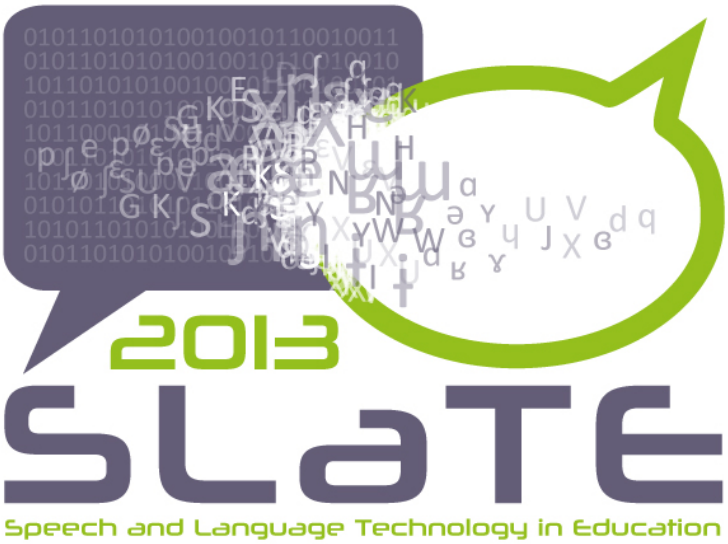


Proceedings of SLaTE 2013

Interspeech 2013 Satellite workshop on

Speech and Language Technology in Education

Grenoble, France - August 30-31 & September 1st, 2013

Organized by

GIPSA-lab and LIDILEM with the ISCA-SLaTE group

Editors:

Pierre Badin – Thomas Hueber – Gérard Bailly

Didier Demolin – Françoise Raby

ISSN 2311-4975

Organization

SLaTE-2013 is being organized by GIPSA-lab and LIDILEM (Grenoble, France), within the framework of the ISCA-SIG SLaTE. The organizers would like to thank Martin Russell, Helmer Strik, and Maxine Eskenazi for giving us instructive advice to organize this workshop.

Local Organizing Committee

Pierre Badin, GIPSA-lab
Thomas Hueber, GIPSA-lab
Gérard Bailly, GIPSA-lab
Didier Demolin, GIPSA-lab
Françoise Raby, LIDILEM

Local logistics

Nadine Bioud, GIPSA-lab
Martine Giglio, CNRS
Emilie Magnat, LIDILEM
Cécilia Mendès, GIPSA-lab
Akila Mokhtari, GIPSA-lab
Jessica Réolon, GIPSA-lab

ISCA – SLaTE International scientific committee

Abeer Alwan, UCLA
Jared Bernstein, Ordinate Corp.
Rodolfo Delmonte, University of Venice
Maxine Eskenazi, Carnegie Mellon
Björn Granström, KTH
Valerie Hazan, UCL
Diane Litman, U. of Pittsburgh
Dominic Massaro, USCS
Nobuaki Minematsu, U. of Tokyo
Patti Price, PPrice.com
Martin Russell, Univ. of Birmingham
Stephanie Seneff, MIT
Helmer Strik, Radboud U., Nijmegen
Catia Cucchiarini, Radboud U., Nijmegen

Scientific Review Committee

The organizers would like to thank the following individuals who took part in the review of papers submitted to SLaTE 2013.

Pierre Badin
Gérard Bailly
Anton Batliner
Kay Berkling
Jared Bernstein
Lei Chen
Jean-Pierre Chevrot
Catia Cuchiarini
Febe de Wet
Rodolfo Delmonte
Didier Demolin
Ryan Downey
Olov Engwall
Donna Erickson
Maxine Eskenazi
Keelan Evanini
Horacio Franco
Björn Granström
Mark Hasegawa-Johnson
Nathalie Henrich
Thomas Hueber
Lewis Johnson
Hiroaki Kato
Mariko Kondo
Jacques Koreman
Diane Litman
Joaquim Llisterri
Nuno Mamede
Nobuaki Minematsu
Hansjörg Mixdorff
Jean-Paul Narcy-Combes
Thomas Pellegrini
Patti Price
Françoise Raby
Martin Russell
Stephanie Seneff
Theban Stanley
Helmer Strik
Joseph Tepperman
Oscar Saz Torralba
Isabel Trancoso
Chiu-Yu Tseng
Hugo Van Hamme
Karl Weilhammer

Message from the Local Organizers

We are very happy to welcome you to SLaTE 2013 in Grenoble!

We are about 60 from all over the world to be gathered here to discuss the last advances in the domain of Speech and Language Technology in Education.

Following SLaTE'2007 in Farmington, USA, SLaTE'2009 in Birmingham, UK, L2WS in 2010 in Tokyo, Japan, and SLaTE'2011 in Venice, Italy, this workshop is the fifth ISCA-supported SLaTE workshop.

Fourty-six contributions have been submitted in all areas of SLaTE. All full or two-page contributions have been reviewed by three reviewers, and given a score. Contributions with an average score greater than +1 have been directly accepted. Those with a score lower than –1 have been rejected. The remaining ones have been carefully discussed by the selection committee. Finally thirty-nine contributions have been selected for presentation at the workshop in either oral or demonstration and poster sessions.

These contributions attest the wealth and dynamism of this research area: children oriented research, corpus based studies of speech diversity, mobile or web-based applications, elaborate dialogue systems and serious games, phonetics / phonology studies on L2.
We have organised the workshop in the following way, with the aim to provide lively discussions.
We have planed a short round table at the end of each of the five oral sessions. We rely on chair persons to raise general or specific questions about the field to the speakers or to the audience.
We will have a session for demonstration and posters with a dozen contributions. This session will start with a series of three minutes presentations of the contributions allowing authors to attract a larger audience.
Long lunches à la Française at the Canberra restaurant will provide time to discuss important matters … or to enjoy the university swimming pool.

In order to promote opening towards fields that are clearly related to SLaTE but finally not so present in this community, we have invited three keynote speakers known for their broad ideas and their vision of their domains.
Diane Litman, from the University of Pittsburgh, will tell us in a few moments her vision of how speech and language processing can be used to enhance dialogue applications for teaching STEM, *i.e.* Science, Technology, Engineering, and Mathematics.
Tomorrow morning, ***Jozef Colpaert***, from the University of Antwerp, will share his experience about what is important in CALL systems, using the design concept behind a project on pronunciation training, and more generally about the new concept of Educational Engineering.
Finally, on Sunday morning, ***Mary Beckman***, from the Ohio State University, will discuss how speech technology can be used to learn more about child speech and in particular to evaluate cross-language differences in the early emergence of contrasting vowel categories.

We are very lucky to be financially supported by a number of institutions.
GIPSA-lab, the Engineering education Grenoble INP group, the Stendhal University that deals with Language and Literature, the Joseph Fourier University involved in Scientific and Medical research, and also from the Grenoble Cognition Pole which aims at initiating and coordinating all Grenoble research teams involved in cognition.
We have also received a strong support from the Rhône-Alpes Region through its Academic Research Community on Information and Communication Technologies and Innovative Computer Usage. Finally, we could also mention the supports of the Grenoble – Alpes metropole consortium of cities and from the Grenoble city itself.
We are grateful to all of them for supporting the workshop.

We would also like to thank all reviewers who accepted to read and assess the papers and provided relevant comments.

We hope that all of you will enjoy the workshop.

Pierre Badin, Thomas Hueber, Gérard Bailly, Didier Demolin & Françoise Raby

Authors list

Sponsors

The SLaTE-2013 local organizing committee is happy to acknowledge the financial support of the following sponsors.

Enhancing the Effectiveness of Spoken Dialogue for STEM Education

Diane J. Litman[1]

[1]Department of Computer Science &
Learning Research and Development Center
University of Pittsburgh
Pittsburgh, PA 15260 USA
dlitman@pitt.edu

Abstract

This talk will discuss the application of speech and language processing to two types of STEM (Science, Technology, Engineering, and Mathematics) dialogue applications: 1) one-on-one physics tutoring, where students engage in dialogues with either a computer or human tutor, and 2) engineering design, where students engage in multi-party dialogue to complete a group project. I will first present results illustrating that relationships exist between student learning and both student affect, as well as lexical/prosodic entrainment between conversational partners. I will then illustrate our use of such findings to build better educational dialogue systems.

Index Terms: spoken dialogue systems, intelligent tutoring, multi-party dialogue, affective systems, lexical and prosodic entrainment

1. Introduction

Students working one-on-one with expert human tutors have scored up to 2.0 standard deviations higher than students working on the same topic in classrooms [1]. In contrast, the best intelligent tutoring systems have yielded much smaller performance gains. One major difference between human tutors and current computer tutors is that only human tutors participate in unrestricted natural language dialogue with students, which has led to the conjecture that human tutoring might be so effective because of its use of dialogue [2].

Computational dialogue systems such as Siri are already providing spoken language access to many types of information services, with potential benefits of remote or hands-free access, ease of use, and naturalness. The use of dialogue technology to build computer tutors similarly has the potential to provide many benefits as a learning environment. For example, a dialogue system can use speech and language technology to infer information about a student's knowledge and/or affective state, which in turn can help a tutoring system better tailor instruction to the needs of a student. In STEM (Science, Technology, Engineering and Mathematics) domains, spoken dialogue can allow students to receive tutoring while simultaneously engaging in learning activities requiring their hands, e.g. scientific lab work. In addition, there has been momentum in the science education literature recognizing the importance of talking, reflecting and explaining as ways to learn; dialogue allows students to participate more actively in the learning process via behaviors such as self-explanation [3]. Finally, it is often the case that students can solve numerical scientific problems while retaining a poor overall knowledge of underlying concepts and principles. Tutorial dialogue can be used to address poor conceptual learning, by adding natural language instruction to quantitative problem solving tutors, by using dialogue to teach conceptual knowledge directly, or by using dialogue in post problem-solving reflective activities.

For all of these reasons, the development of automated tutorial dialogue systems has emerged as a promising method for attempting to close the current performance gap between human and computer tutors. STEM tutorial dialogue systems have been developed for teaching biology [4], circuit design [5], computer science [6, 7], electricity and electronics [8], physics [9, 10, 11, 12], thermodynamics [13], elementary school science [14], and shipboard damage control [15]. Tutorial applications differ in many ways, however, from the types of applications for which dialogue systems are typically developed. The relative educational benefits of using different types of speech and language technology (e.g. the use of spoken versus typed student input [16, 17]) to build tutors and other learning environments is thus an active area of investigation. A related area of research is the use of speech and language technology to annotate and analyze conversational educational data.

2. Adapting to Student States in a Spoken Tutorial Dialogue System for Physics

While most tutorial dialogue systems respond based only on the correctness of a student answer, it has been hypothesized that students could learn even more if the tu-

tor also responded to other pedagogically relevant student states (e.g. student affect and attitudes). While there has been considerable research on user state detection in naturally occurring spoken dialogue, most work has focused on states typically seen during customer care and information-seeking applications (e.g. anger and frustration). Less work has addressed the detection of student states more commonly seen during tutoring interactions (e.g. boredom, confusion, delight, flow, frustration, and surprise [18]). In contrast, while research in intelligent tutoring systems has attempted to detect such pedagogically relevant states, most computer tutors are not spoken tutorial dialogue systems.

We have conducted a series of studies examining the benefits and challenges of building a spoken dialogue system that can detect and adapt to student uncertainty [19, 20, 21] and disengagement [22, 23] during conceptual physics tutoring. Our adaptive dialogue system first detects uncertainty and/or disengagement in each student turn via learned models that use acoustic-prosodic features extracted from the speech signal, lexical features extracted from a noisy speech recognition transcript, as well as contextual features extracted from dialogue system logs to predict student states. The tutor then varies its response content based on the detected student states, using dialogue strategies learned from corpora of human tutoring dialogues. A series of experimental evaluations demonstrate that adapting to student uncertainty over and above answer correctness, as well as further adapting to student disengagement over and above uncertainty, can increase student learning as well as improve other performance measures.

3. Prosodic and Lexical Entrainment in (Multi-Party) STEM Dialogues

Linguistic entrainment refers to the convergence of (para)linguistic features across speakers during the course of a conversation. Research has found that speakers entrain to both human and computer conversational partners, with the amount of entrainment often positively related to conversational success. A variety of measures for automatically computing entrainment have already been developed by speech and language researchers, and have been shown to correlate with task and communicative success in many dialogue contexts [24, 25, 26].

We have been exploring the use of linguistic entrainment as a method for predicting the amount of student learning in STEM educational dialogue, as a first step towards building learning environments that can leverage entrainment. Our particular interests are in predicting educationally-relevant measures of success (e.g. student learning, solution quality), and in moving from two-party to multi-party dialogue, given that much science innovation occurs in teams. To date we have examined entrainment in both one-on-one tutoring dialogues between students and tutors discussing conceptual physics [27], and multi-party conversations between student team members working on a semester engineering project [28]. In tutorial dialogue, we have found that the more a tutor and student entrain prosodically (quantified using both existing and new metrics), the more the student learns. In multi-party dialogues, we proposed a measure of lexical entrainment that extends an existing measure of pair entrainment to groups. We then demonstrated that there is a significant difference between the lexical entrainment of high performing teams, which tended to increase with time, and the entrainment for low performing teams, which tended to decrease with time, but only with respect to task-related words. Our long-term goal is to use our findings for a range of educational purposes, e.g. mining conversational data to support teacher-oriented analytics, or triggering interventions in adaptive conversational agents.

4. Acknowledgments

This research has been supported by the National Science Foundation under Grant Nos. 9720359, 0328431, 0325054, 0631930, and 0914615. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Science Foundation. Some of this research was also supported by ONR (N00014-04-1-0108).

5. References

[1] B. S. Bloom, "The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring," *Educational Researcher*, vol. 13, no. 6, pp. 4–16, 1984.

[2] A. C. Graesser, N. K. Person, and J. P. Magliano, "Collaborative dialogue patterns in naturalistic one-to-one tutoring," *Applied Cognitive Psychology*, vol. 9, pp. 1–28, 1995.

[3] M. Chi, N. D. Leeuw, M.-H. Chiu, and C. Lavancher, "Eliciting self-explanations improves understanding," *Cognitive Science*, vol. 18, pp. 439–477, 1994.

[4] M. W. Evens and J. Michael, *One-on-one tutoring by humans and computers*. Psychology Press, 2006.

[5] R. W. Smith and S. A. Gordon, "Effects of variable initiative on linguistic behavior in human-computer spoken natural language dialogue," *Comput. Linguist.*, vol. 23, no. 1, pp. 141–168, 1997.

[6] C. Kersey, B. Di Eugenio, P. Jordan, and S. Katz, "KSC-PaL: A peer learning agent that encourages students to take the initiative," in *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, 2009, pp. 55–63.

[7] K. E. Boyer, R. Phillips, A. Ingram, E. Y. Ha, M. Wallis, M. Vouk, and J. Lester, "Investigating the relationship between dialogue structure and tutoring effectiveness: A hidden markov modeling approach," *International Journal of Artificial Intelligence in Education*, vol. 21, no. 1, pp. 65–81, 2011.

[8] M. O. Dzikovska, J. D. Moore, N. Steinhauser, G. Campbell, E. Farrow, and C. B. Callaway, "Beetle II: A system for tutoring and computational linguistics experimentation," in *Proceedings of the ACL 2010 System Demonstrations*, 2010, pp. 13–18.

[9] K. VanLehn, A. C. Graesser, G. T. Jackson, P. Jordan, A. Olney, and C. P. Rosé, "When are tutorial dialogues more effective than reading?" *Cognitive Science*, vol. 31, no. 1, pp. 3–62, 2007.

[10] D. J. Litman and S. Silliman, "Itspoke: An intelligent tutoring spoken dialogue system," in *Demonstration Papers at HLT-NAACL 2004*, 2004, pp. 5–8.

[11] M. Chi, K. VanLehn, D. Litman, and P. Jordan, "Empirically evaluating the application of reinforcement learning to the induction of effective and adaptive pedagogical strategies," *User Modeling and User-Adapted Interaction*, vol. 21, pp. 137–180, 2011.

[12] S. Katz, P. Albacete, M. Ford, P. Jordan, M. Lipschultz, D. Litman, S. Silliman, and C. Wilson, "Pilot test of a natural-language tutoring system for physics that simulates the highly interactive nature of human tutoring," in *Proceedings 16th International Conference on Artificial Intelligence in Education*, Memphis, TN, 2013.

[13] C. P. Rosé, R. Kumar, V. Aleven, A. Robinson, and C. Wu, "Cycletalk: Data driven design of support for simulation based learning," *International Journal of Artificial Intelligence in Education*, vol. 16, no. 2, pp. 195–223, 2006.

[14] W. Ward, R. Cole, D. Bolaños, C. Buchenroth-Martin, E. Svirsky, S. V. Vuuren, T. Weston, J. Zheng, and L. Becker, "My science tutor: A conversational multimedia virtual tutor for elementary school science," *ACM Trans. Speech Lang. Process.*, vol. 7, no. 4, pp. 18:1–18:29, 2011.

[15] H. Pon-Barry, K. Schultz, E. O. Bratt, B. Clark, S. Peters, H. Ponbarry, K. Schultz, E. O. Bratt, B. Clark, and S. Peters, "Responding to student uncertainty in spoken tutorial dialogue systems," *International Journal of Artificial Intelligence in Education*, vol. 16, pp. 171–194, 2006.

[16] D. J. Litman, C. P. Rosé, K. Forbes-Riley, K. VanLehn, D. Bhembe, and S. Silliman, "Spoken versus typed human and computer dialogue tutoring," *International Journal of Artificial Intelligence in Education*, vol. 16, no. 2, pp. 145–170, 2006.

[17] S. K. D'Mello, N. Dowell, and A. Graesser, "Does it really matter whether students' contributions are spoken versus typed in an intelligent tutoring system with natural language?." *Journal of Experimental Psychology: Applied*, vol. 17, no. 1, pp. 1–17, 2011.

[18] S. D'Mello, B. Lehman, J. Sullins, R. Daigle, R. Combs, K. Vogt, L. Perkins, and A. Graesser, "A time for emoting: When affect-sensitivity is and isn't effective at promoting deep learning," in *Proc. Intelligent Tutoring Systems Conference*, Pittsburgh, 2010, pp. 245–254.

[19] K. Forbes-Riley and D. Litman, "Designing and evaluating a wizarded uncertainty-adaptive spoken dialogue tutoring system," *Computer Speech & Language*, vol. 25, no. 1, pp. 105–126, 2011.

[20] ——, "Benefits and challenges of real-time uncertainty detection and adaptation in a spoken dialogue computer tutor," *Speech Communication*, vol. 53, no. 9–10, pp. 1115–1136, 2011.

[21] D. Litman and K. Forbes-Riley, "Towards improving (meta) cognition by adapting to student uncertainty in tutorial dialogue," in *International Handbook of Metacognition and Learning Technologies*. Springer, 2013, pp. 385–396.

[22] K. Forbes-Riley, D. Litman, H. Friedberg, and J. Drummond, "Intrinsic and extrinsic evaluation of an automatic user disengagement detector for an uncertainty-adaptive spoken dialogue system," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2012, pp. 91–102.

[23] K. Forbes-Riley and D. Litman, "Adapting to multiple affective states in spoken dialogue," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2012, pp. 217–226.

[24] A. Nenkova, A. Gravano, and J. Hirschberg, "High frequency word entrainment in spoken dialogue," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Short Papers*, 2008, pp. 169–172.

[25] G. Parent and M. Eskenazi, "Lexical entrainment of real users in the Let's Go spoken dialog system," in *Proceedings of Interspeech*, 2010.

[26] C. M. Mitchell, K. E. Boyer, and J. C. Lester, "From strangers to partners: Examining convergence within a longitudinal study of task-oriented dialogue," in *Proceedings of the 13th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2012, pp. 94–98.

[27] J. Thomason, H. Nguyen, and D. Litman, "Prosodic entrainment and tutoring dialogue success," in *Proceedings 16th International Conference on Artificial Intelligence in Education (AIED)*, Memphis, TN, 2013.

[28] H. Friedberg, D. Litman, and S. B. Paletz, "Lexical entrainment and success in student engineering groups," in *Spoken Language Technology Workshop (SLT)*. IEEE, 2012, pp. 404–409.

The role and shape of speech technologies in well-designed language learning environments.

Jozef Colpaert

IOIW, Universiteit Antwerpen, Belgium

jozef.colpaert@ua.ac.be

Abstract

No technology carries an inherent, direct, measurable and generalizable effect on learning. Nor does speech technology. Its assumed added value is not a starting point for design, but a hypothesis resulting from design. Role and shape of speech technologies are simply logical and natural consequences of the specification of an optimal language learning environment for a specific context and for specific subconscious personal goals. These *ecological* and *psychological* paradigm shifts will be illustrated by a discussion of the design concept behind a project on pronunciation training. Challenges for CALL research on this topic will be briefly discussed.

Index Terms: pronunciation training, courseware design, learning environments

1. Introduction

When we look back on thirty years or so of speech technology, we observe an impressive technological evolution, in hardware and software, in concepts, routines and models. On the other hand, we have to admit that this speech technology has not yet permeated our daily life and certainly not language education. There was indeed no pot of gold at the end of the voice rainbow. The reason for this, in our view, had less to do with technological limitations than with underestimated psychological aspects [1].

CALL is the field *par excellence* where speech technology could play an obvious role: on the level of the interface (navigation), for practicing oral communication (simulation of natural interaction) or for specific pronunciation training (remediation of specific topics). Despite their technological complexity and power, despite their pedagogical usefulness, despite their theoretical soundness, speech systems have not yet been able to show their effectiveness on the basis of evidence in terms of usage and satisfaction. The problem is a design problem.

2. Educational Engineering

Educational Engineering is our Instructional Design model for guiding the design, development, implementation and evaluation of educational artefacts for learning, testing and teaching. These educational artefacts can be documents, tools, content, concepts, models and solutions such as textbooks, syllabi, lesson plans, curricula, graded readers, exercises, tests, applications or electronic learning platforms.

The term engineering does not necessarily refer to technology, but it primarily denotes the typical actions we have to undertake when not enough knowledge is available for attaining our goal. Engineering is not only about solving practical problems by applying scientific knowledge, it is also about building knowledge through real-world implementations, in a systematic and verifiable way, using working hypotheses that should be empirically and theoretically validated.

Educational engineering is needed because there is not enough knowledge available for creating perfect artefacts. By its very nature, education can and will never be perfect. It will always be *l'art du possible*. Educational Engineering is geared towards obtaining the best possible results, applying the best possible methodologies, taking into account as many actors and factors as possible.

Educational Engineering can be considered a design-process oriented model. Generally speaking it distinguishes itself from other design-process oriented approaches on several points. First, Educational Engineering focuses on a larger process than on design alone. It also embraces and clearly specifies other stages such as Analysis, Development, Implementation and Evaluation (ADDIE). The Analysis stage focuses on the identification of elements of the learning context which are amenable to improvement, or which should be taken into account during design, but does not state anything about the eventual design. Design focuses on the conceptualization, specification and possible prototyping of educational artefacts. Result of the design stage is a mental representation, a virtual construct, a blueprint or even a metaphor. Actual development is clearly left for its proper stage, followed by Implementation and Evaluation. Educational Engineering states that more time, energy and effort should be put in Analysis and Design as these stages are crucial for the eventual quality and effect of the targeted product.

Educational Engineering is based on real-world iteration: its starting point is a concrete problem, but it does not focus on the problem alone. It focuses on how to reach an optimal solution which in most cases cannot be realized in one step, due to resistance, financial limitations, technological challenges or practical constraints. The long-term engineering process goes through a series of ADDIE lifecycles (Fig.1), and each of these cycles formulates a very precise and justifiable intermediate change on the pathway to the optimal solution. The evaluation stage of every cycle not only validates the suggested changes, but also confirms or readjusts the concept of the optimal solution along the way. This concept serves as a lighthouse for all actors involved. The lighthouse metaphor is not accidental here: a lighthouse shows direction to boats, but it never is their final destination. So will the eventual solution be different than the initially conceived optimal solution.

Contrary to Rapid Application Design (RAD) or Rapid Instructional Design (RID), Educational Engineering does not insist on much iteration during design itself, as it considers its typical real-world iteration as the most important source of information. While RAD and RID consider real-world iteration too slow and time-consuming – which in a large number of cases can be a valid argument – the goal of Educational Engineering is more than creating optimal solutions: it also intends to validate research hypotheses, build knowledge and share expertise in an academic context.

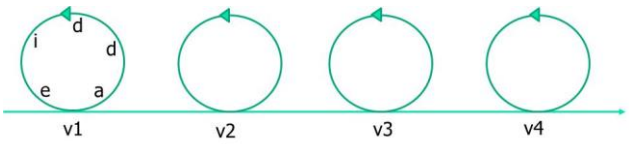


Figure 1. *Engineering cycles*

The optimal solution is a hypothesis, but so are the consecutive intermediate loops. Every hypothesis is based on theoretical findings, practical experience, and the outcome of previous loops. The role of theory is to feed the process with as many useful, relevant and substantiated findings and concepts as possible, in order to increase the efficiency of the process, to guarantee the effectiveness of its product and to reduce the risk of failures. In the case of language learning and teaching for example, theories to be taken into account pertain roughly speaking to the following fields: pedagogy, psychology, technology, linguistics and specific sub-disciplines such as Human-Computer Interaction (HCI), Second Language Acquisition (SLA), Computer Mediated Communication (CMC), Motivation Theory, Activity Theory and Cognitive Multimedia Theory. The integration of this theory typically happens during two stages. During the Analysis phase, the educational engineer checks whether or not enough theoretical knowledge and findings are available for carrying out the project, and (s)he has to make the inventory of all required knowledge for designing the optimal solution. Secondly, during Design, the final shape will to a great extent, but not exclusively, be determined by theory. So theory is not directly applied to nor translated into the solution (like applying 4C/ID as such in a concrete learning situation), but it serves as one of the premises of a logical reasoning that forms the hypothesis.

Educational Engineering intends to be a universally applicable model (statement to be validated), but it does not state anything about the eventual shape of the solution (as this mainly depends on the context), nor about which theories are relevant, useful and/or applicable. The product should not be evaluated as such on its features, and this for two reasons: a/ a product is by definition always an intermediate solution and b/ as the product will always depend on the local context. Applying the same model leads to polymorphous results. Design should not be confused with shape. While we can easily observe shape, good design often remains invisible. Because design refers to the work behind the shape.

3. Distributed Design

This optimal solution, the virtual product of the design process, will always be a hypothesis, but the proposed process itself will also remain a hypothesis to be continuously subject to theoretical and empirical validation. Seven hypotheses, paradigm shifts really, grown out of the confrontation of practical experience with theoretical findings, have led to our own specific design model:

- The ecological paradigm shift
- The psychological paradigm shift
- Focus on the process
- Distributed design
- Ontological specification
- Generic Content
- Teacher Support

We will now briefly discuss the first two paradigm shits, leaving the others for upcoming publications.

3.1. The ecological paradigm shift

Research into the learning effect created by a single educational artefact, such as a new technology (tablet, serious game, IWB …) often leads to the *No Significant Difference Syndrome* (the more difficult to prove the effect, the more complex the statistics), or at least to non-generalizable results. The main tenet of Distributed Design is that the added value of a particular educational artefact is proportional to the extent to which it contributes to the creation of an optimal learning environment (OLE).

The term learning environment in its traditional acceptance refers to a collection of components such as actors (learner, teacher, parent, policy maker, content provider …), content, infrastructure, technology and models (for teaching, learning and evaluation). The Distributed Design approach defines the learning environment more as a self-regulating system, a learning ecology, where more attention goes to the interplay between the components of the environment, the context and the rationale behind its design. It focuses on the possible effect on learning of this entire ecology, and tries to research to what extent this ecology can be optimized, in other words leading to better results for all actors involved, both in quantitative and qualitative terms.

An OLE is a blueprint of an ideal learning environment which by definition will (perhaps) never exist. As already stated, its function is that of a lighthouse: it shows direction. In the same vain, an OLE should perhaps never be realized as such, but its main purpose will be to guide the decision process along the way. An OLE also has its specific scope. This scope is determined by the users of our educational artefacts: it can be a class, a grade or degree, an institution, a country or even the entire world (e.g. Open University).

An OLE cannot be realized in one step, but it should inspire small changes to be undertaken in the existing learning environment, typically every year. Every redesigned learning environment should always be seen as an instantiation of the OLE. This instantiated learning environment or ILE should be

specified in detail. The purpose of an ILE is to test a hypothesis, and after evaluation and validation, formulate a new hypothesis leading to a new ILE along the pathway to the OLE. The number of changes in the design of a new ILE, compared to the previous one, depends on available resources, on resistance to be expected, on the research-oriented nature of the activity etc. Hypotheses are based on previous experience (evaluation of previous ILEs, exchanges with colleagues worldwide …) and on theory. The reasoning leading to a new hypothesis should be based on a sound construct based on substantiated evidence. It is obvious in this respect that also the design of the OLE can and should be adjusted along the way on the basis of these intermittent evaluations.

3.2. The psychological paradigm shift

An OLE should be designed with a clear focus on a particular pedagogical goal. Pedagogical goals are mostly well documented, easy to find, explicit and detailed. Their formulation largely depends on the scope of the OLE, and they range from lesson plans over course goals ("At the end of this course you will be able to …") and grade descriptors (French 101 or Common European Framework for Languages), to country level (official learning programmes). The term optimal learning environment refers to its very *reason d'être*, i.e. to offer the best possible guarantee that the set pedagogical goals can be realized as efficiently and effectively as possible.

However, especially in cases of lesser motivation, it is counterproductive to focus exclusively or too directly on the realization of these pedagogical goals. It is far more efficient to focus on personal goals first. Personal goals can be considered subconscious volitions that hinder or stimulate the learning process. The problem is that these goals are quite difficult to elicit [2].

The starting point – or angle of attack – of Distributed Design is the point where personal goals and pedagogical goals conflict such as in cases where students have to learn French but may not be motivated to do so, where students have to learn to be autonomous but they may prefer strong guidance, or where students should learn how to cope with chaos but they may prefer a strong structure.

Most of the effort in Distributed Design goes into trying to reconcile these conflicting goals into a strong concept. The concept underlying an eventual OLE can be expressed as a metaphor (such as a city, a space station, a forest or a power plant). This concise representation makes sure that all actors involved (designers, developers, users and stakeholders) carry more or less the same mental image. This is important for the design team, but also for the teachers and learners.

4. Designing for pronunciation training

4.1. Computer Assisted Pronunciation Training

Computer Assisted Pronunciation Training (CAPT) is a fairly recent discipline. As its history is well known by the readers of this volume, we will not try to provide another overview here, but we just wish to refer to the many publications on the topic by a.o. our eminent colleagues Catia Cucchiarini and Helmer Strik, such as [3]. There are, however, a couple of aspects that are worth mentioning at this stage. Epistemologically speaking, we see four challenges. The first one is a *technological* one: the use of spectral analysis versus speech recognition routines. Each approach has its limitations and affordances. The second one is related to the *feedback* problem: how to make sure that visual or auditory feedback is meaningful for the learner. Thirdly, the effect of *auditory discrimination training* on pronunciation: it makes little sense to train pronunciation if the learner does not hear the difference between two sounds. And finally, the issue of *washback*: how can the analysis and evaluation of CAPT system usage impact on the way pronunciation is taught in language education?

4.2. The DISCO project

The DISCO project originated within the framework of the Dutch-Flemish STEVIN programme of the *Nederlandse Taalunie* and aimed at developing and testing a prototype of an ASR-based CALL application for Dutch as a second language (DL2). Radboud University Nijmegen focused on the linguistic-didactic aspects of pronunciation, while the University of Antwerp focused on design, usage and usability issues.

While the initial design [4] reflected a more traditional user-driven and lexis/grammar-oriented approach (with the three main program components phonology – morphology – syntax), we very soon became aware that a different interface was needed in order to guarantee acceptance and willingness in the users' mind. With respect to pronunciation, we aimed at the achievement of intelligibility, rather than accent-free pronunciation. The target user group consisted of highly educated immigrants in Flanders and the Netherlands who wanted to be able to communicate with locals in the most natural way (Fig. 2).

Figure 2: *Interaction example*

Our analysis in terms of personal goals ([2]) indeed pointed out that the targeted learners wanted to feel integrated and accepted as quickly as possible, without losing or betraying their own identity and culture. They also preferred a learning process that would give them a feeling of freedom, respect and

autonomy, without a school-like approach with strong guidance nor exaggerated focus on metalanguage or –again- civic duties. The system had to be usable as a part of the learning environments at various DL2 centers in Flanders and the Netherlands [5].

The design we eventually came up with was based on the following premises: a/ not the learner, but the system has to make the choice between phonology, morphology or syntax as exercise items, based on an intelligent analysis of user behavior and performance; b/ the user has to interact with an agent or a persona in the program who is interested in the user and who wants to help him/her; c/ the user has to interact in a natural way with this agent, in a simulated branched conversation through oral selection without being aware that his/her pronunciation is constantly being analyzed; d/ only after detecting systematic pronunciation errors, the system will interrupt the interaction with the appearance of doctor Spraak (Fig. 3), who will invite the user to work on some specific listening and speaking tasks.

Figure 3: *Doctor Spraak*

5. Discussion

The current system affords the most advanced and up-to-date functionalities in pronunciation training. The design stage considerably simplified the interface and the system architecture in general. The goal of the project was not to deliver a market-ready product, but we would have reached this stage, and we would have been able to test it extensively, if one of the project partners had been able to deliver the system on time.

Challenges for further CAPT research include a deeper psychological study of feedback scenarios, the role of recast and auditory discrimination and a thorough analysis of user behavior and performance before formulating our next hypothesis.

6. Conclusions

Speech technology in general, and CAPT in particular, seem to have the advantage that they can be used in many language learning and teaching situations, without having to be adapted significantly to the local users and their context. This feature is quite deceptive as we state that design, meaning adapting a system construct to the ecology and the psychology of the user, not only leads to more and better use, to more learning results, but also to a less complex system architecture and a simple interface.

Having made an analysis of a target group (DL2) and of a number of similar learning contexts (university context) we were able to design a specific learning environment. The role and shape of the system we needed for pronunciation training are completely determined by that learning environment and are not transportable as such to other environments.

The role and shape of speech technologies in general should not be determined by affordances, expectations, hypes nor perceptions. They should be the result of a methodological and justifiable design process. Their eventual role and shape depend on the local context and will always be polymorphous.

In order to end on a less positive tone: what we have learned on the level of the design process is, again similar to our experience with other projects, that most of the resistance against ecological and psychological design seems to come from the actors involved themselves: project partners, learners, teachers, policy makers, publishers etc. We will come back to this issue in a forthcoming publication.

7. References

[1] Venkatesh, V., Thong, J. Y.L. and Xu, X., “Consumer Acceptance and Use of Information Technology: Extending the Unified Theory of Acceptance and Use of Technology”, MIS Quarterly, 36(1):157-178, 2012.

[2] Colpaert, J., “Elicitation of language learners’ personal goals as design concepts”, Innovation in Language Learning and Teaching. 4(3):259-274, 2010.

[3] Neri, A. Cucchiarini, C., Strik, H. and Boves, L., "The pedagogy-technology interface in Computer-Assisted Pronunciation Training", Computer-Assisted Language Learning, 15(5): 441-467, 2002.

[4] Strik, H., Cornillie F., Colpaert, J., Van Doremalen, J, and Cucchiarini, C., “Developing a CALL System for Practicing Oral Proficiency: How to Design for Speech Technology, Pedagogy and Learners”, SLaTE 2009 Online proceedings (http://www.eee.bham.ac.uk/SLaTE2009/index.html), University of Birmingham, 2009.

[5] Strik, H., Van Doremalen, J., Colpaert, J. and Cucchiarini, C.. “Development and Integration of speech technology into courseware for language learning: the DISCO project”, in P. Spyns and J. Odijk (eds.), Essential Speech and Language Technology for Dutch, Theory and Applications of Natural Language Processing, STEVIN Programme, Nederlandse Taalunie. Springer, 2013.

Enriched technology-enabled annotation and analyses of child speech

Mary E. Beckman

Department of Linguistics, Ohio State University, Columbus, OH, USA
mbeckman@ling.osu.edu

Abstract

This paper reviews a range of studies illustrating the ways in which speech technology has enabled richer analyses of corpora of young children's speech and infants' speech-like vocalizations. One set of studies illustrates the use of speech synthesis technology, such as VLAM, an articulatory synthesis system that models the transfer functions of child- or infant-proportioned vocal tracts. VLAM has been used to evaluate cross-language differences in the emergence of contrasting vowel categories. Another set of studies illustrates the use of modern corpus development and annotation tools to create and analyze the paidologos corpus, a databse of utterances elicited in a picture-prompted word-repetition task from 2- through 5-year-old child speakers of a variety of languages. Flexible, incremental annotation on multiple tiers allows researchers to extract target sounds for spectral analysis as well as for calculating transcribed accuracy rates. Tag sets also can be used to extract stimuli for perception experiments, yielding naive-listener responses that can become another layer of tags.

Index Terms: child speech corpora, articulatory synthesis, perceptual evaluation of child productions

1. Introduction

Corpus studies have long played an important role in advancing our understanding of first language acquisition. For example, in the early part of the 20th century, Otto Jespersen and Roman Jakobson, among others, portrayed children's phonological development as an orderly sequence of stages, which begins when the rich inventory of sounds randomly produced in babbling abruptly gives way to a very reduced phonetic inventory in first words, followed by a re-acquisition of consonant contrasts that proceeds in "a strict and invariable temporal sequence" [1]. This characterization of phonological development as invariable across children and fundamentally different from the richly variable productions of babbling was an extrapolation from the few published diary studies, in which the researcher (typically the parent) recorded typical pronunciations of words and word-like forms observed in the course of daily interaction with a child. It was the dominant view in linguistics until the commercialization of good-quality magnetic analog tape recorders enabled research using more controlled phonetic transcription and even acoustic analysis of babbling and early word productions in corpora that included more children (e.g., [2], [3]).

This paper reviews a range of studies illustrating the ways in which more recent developments in speech technology have enabled even richer analyses of corpora of children's speech and younger infants' non-speech vocalizations. One set of studies uses speech synthesis programs such as VLAM [4], an articulatory synthesis program based on Maeda's Vtcalcs [5] that can model the transfer functions of infant- and child-proportioned vocal tracts. Another set of studies illustrates the application of resources that were developed in building speech databases for training speech synthesis and automatic speech recognition systems in designing and annotating the paidologos cross-language corpus of words and nonwords elicited from 2- through 5-year-old children [6]. Sections 2 and 3 review these two ways of applying speech technology to the analysis of child speech. Each section also cites several studies that use speech synthesis or the tagged corpus to create controlled stimulus sets for perception experiments.

2. Analysis by speech synthesis

The general method of analysis by speech synthesis has been used to investigate phonological representations and processes in adults for many decades. Classic examples include Cohen and 't Hart's use of close-copy stylization of the fundamental frequency (F0) contour to develop their model of the Dutch intonation system [7], and Stevens and House's application of a simple tongue-arc model [8] to explore the features of the American English vowel space as estimated for adult male talkers by the formant values measured in the Peterson and Barney corpus [9].

In the decades since these classic studies, the revolution in computer hardware that gave the world small, fast, powerful computers also led to the eventual incorporation of more recent techniques such as LPC analysis-resynthesis [10] and PSOLA [11] into free signal-analysis programs, such as Praat [12], that would have been unimaginable even as recently as the 1990s. This development in turn has made the analysis-by-synthesis method of close-copy stylization available to anyone with access to a personal computer, so that any researcher studying the acquisition of intonation (e.g., [13]) or of lexical tone (e.g., [14]) can apply the method to estimate pitch values in young children's productions even for intervals where the interaction of glottal flow and vocal tract resonances is highly nonlinear (cf. [15]), making F0 estimation as well as formant value estimation difficult.

At an even earlier period of this revolution in computer hardware, most phonetics laboratories and speech laboratories acquired enough computational power to be able to apply statistical techniques such as factor analysis to articulatory data such as cross-sectional distances measured at a vector of points along the mid-sagittal tongue surface, making it possible to build more sophisticated articulatory synthesis models than the simple tongue-arc model of Stevens and House – see, e.g., Harshman et al.'s model [16], as well as Maeda's model [5] mentioned above (5, [17]-[19]). Among these, Maeda's Vtcalcs model is especially noteworthy, both for the size of the database on which it was based (more than 1000 frames of simultaneous cineradiographic and labiofilm recordings from two adult female speakers of French) and for the way in which the code has been distributed freely [20] to allow its use as a research tool by other researchers (e.g., [21]) and also as a tool for teaching phonetics and speech science (e.g., [22]).

Maeda and his colleagues have incorporated Vtcalcs into several different programs for estimating transfer functions of infant- or child-proportioned vocal tracts, the earliest being the

Boë and Maeda's Variable Linear Articulatory growth Model (VLAM) mentioned above [4]. VLAM squeezes (or stretches) the physical scale corresponding to the model parameters that specify the adult female mid-sagittal tongue surface and jaw position, so that the tongue can fit into the dimensions of a scaled-down (or scaled-up) model of the various fixed vocal tract structures that determine the lengths of the pharyngeal and oral cavities, with the sizes of the relevant fixed structures estimated using Goldstein's [23] model of vocal tract growth from birth to 21 years of age, which was based on her survey of extant measurements reported in the literature from the fields of dentistry, anatomy, and physical anthropology. (A more recent model developed by Callan et al. [24] estimates the sizes of the relevant fixed vocal tract structures instead from measurements of several of the first images of infants and children to be included in the Vorperian et al. database of medical MRIs [25].)

Boë, Ménard, and their colleagues have used VLAM to synthesize both maximal vowel spaces (MVS) and a set of "vowel prototype" stimuli for perception experiments ([26]-[32]), in studies designed to explore how vocal tract anatomy constrains the vowel qualities that infants and children can produce at different ages, as the vocal tract is restructured from the shape that enables safe suckling in early infancy to the adult female or male shape. Together these simulations and perceptual studies provide strong support for a picture of vowel acquisition in which vowel-like vocalizations are at first focused in the mid front or the low central region of the vowel space, with gradual expansion along both the first and second formant dimensions until the child has achieved control of the full range of qualities needed to contrast the vowel phonemes of the ambient language.

One of the simulation studies ([28]) also is an explicit cross-language study that uses two different synthesized MVS to interpret formant patterns in a cross-sectional corpus of recordings of 10- to 19-month-old toddlers who were acquiring either Canadian French or Canadian English as their first language. The two groups of toddlers differed both in the typical formant values for the youngest participants and in the dimension that was observed to expand more as a function of the toddler's age. Specifically, the youngest English-learning toddlers had a higher (more front) mean F2 value than did the youngest French-learning toddlers, in keeping with the values that de Boysson-Bardies and colleagues reported in their earlier cross-linguistic study of vowel formants in babbling of 10-month-old infants [3]. The primary developmental trends were that the MVS expanded in the F2 dimension for the older English-learning toddlers, as they mastered the less frequent high and mid back vowels of their language. By contrast, the MVS expanded (lowered) in the F1 dimension for the older French-learning toddlers, as they gained control of the more densely populated high vowel region of French, and the contrast among [y] and a more peripheral target for both [i] and [u] by comparison to the analogous high vowel of English.

Two of the most recent perception studies ([31]-[32]) also are explicit cross-language comparisons, with results showing that adult speakers of different languages interpret the synthesized vowel qualities differently, in ways that suggest the influence of culture-specific processes of talker-size normalization as well as the influence of language-specific representations for even "prototypical" focal vowel categories.

All together, these studies illustrate how speech synthesis technology can be applied effectively to enable a much richer picture of the role of ambient language input even at the very earliest stages of phonological acquisition.

3. Developing the paidologos corpus

Another way in which speech technology has enabled a richer picture of ambient language input is by providing ancillary resources for developing controlled cross-language databases of children's speech. This section illustrates this point by describing the development of the paidologos corpus [6].

The corpus is a set of audio recordings of words and nonwords that are elicited in a picture-prompted word-repetition task from 2- through 5-year-old child speakers. We designed this task so as to be able to elicit productions both of highly familiar early-acquired words and of less familiar words and even nonsense words, using prompts that are consistent across all types of words and across all ages. We wrote a tcl/tk script that loads a suitably randomized stimulus list and then presents, for each stimulus, a picture prompt (shown on the monitor of a laptop computer) and an associated audio prompt (an audio file recording of a woman's voice saying the target word in a child-directed voice, that is played over loudspeakers plugged into the laptop's audio output port). To the left of the picture prompt, there is a picture of a duck (or frog or koala), which moves up a ladder when the program is advanced to the next stimulus, to give the child feedback about the progress of the experiment.

Our primary original aim in developing this picture-prompted word-repetition method was to be able to compare transcribed production accuracy for word-initial pre-vocalic lingual consonants that occur in several languages, but that differ either in phoneme frequency or phonotactic probability between pairs of target languages (see [33][34]). For example, the consonant [ts] occurs in both Cantonese and Greek, but where it occurs in many words of Cantonese, it has a much lower phoneme frequency in Greek. The accuracy rates for [ts] in young Cantonese and Greek children's productions differ in a way that reflects the impact of the cross-language difference in phoneme frequency. Also, the consonant-vowel sequence [ti] occurs in both English and Japanese, but where this sequence begins many words of English, it occurs in only a handful of words of Japanese. And the accuracy rates for [t] before [i] in young English- and Japanese-speaking children's productions differ in a way that reflects the cross-language difference in phonotactic probability.

The first languages we recorded using this paradigm were Hong Kong Cantonese, U. S. English, a northern variety of Greek, and standard (Tokyo) Japanese (see [6], [33], [34]). Not long after, graduate students working with us expanded the corpus to include recordings of a northeastern variety of Mandarin Chinese [35] and of standard (Seoul) Korean [36]. More recently, other graduate student collaborators have adapted the paidologos paradigm to record bilingual speakers of Drehu and French [37] and of the Taiwan varieties of Southern Min and Mandarin Chinese [38].

For many of the languages that we have recorded we used pronunciation dictionaries developed for speech and language technology in order to determine the frequencies of different target consonants and consonant-vowel sequences. We also used these resources to be able to devise nonwords that controlled for phoneme frequency and phonotactic probability in the "frame" portion – i.e., the portion after the form-initial

consonant-vowel sequence that was the target of our investigation.

To be able to calculate transcribed phoneme accuracy, we developed an annotation protocol and annotation scripts that used the tool and scripting language of Praat to automate as much of the process as possible. In developing the protocol, we benefited from from the by-now long history of annotating corpora for speech synthesis and automatic speech recognition (e.g., [39]), to design the annotation to proceed on as many different annotation tiers as we needed for different analysis purposes, with the annotation scripts linking time points that should be linked. The tags we developed allow researchers to easily extract target sounds or shorter intervals for spectral analysis (e.g., [40]-[42]) as well as for calculating and comparing transcribed accuracy rates.

The technology for tagging child productions and for extracting target sounds for spectral analysis also can be applied to develop more sensitive measures of phonological contrast (e.g., [40],[43]). And the tag sets also can be used to extract stimuli for perception experiments, yielding naive-listener responses of various types, including continuous category-goodness ratings, which can act as an added layer of tags to augment the much more time-consuming symbolic tags that are provided by phonetically-trained transcribers (e.g., [41], [42], [44], [45]). Moreover, speech language pathologists also can use these more continuous response types, to evaluate the role of clinical experience in training sensitivity to relevant sub-phonemic variation [47] and to develop more sensitive diagnostics of phonological delay or of incremental progress in response to therapy for speech sound disorders [47].

In short, as with the application of speech synthesis technology, the application of auxiliary tools and resources from speech technology has enabled much richer annotations and analyses of child speech corpora. In both cases, too, the applications yield to enhanced tools for educating students of speech, including speech pathologists as well as phoneticians and engineers.

4. Acknowledgements

Work on this paper was supported by NIDCD grant 02932.

5. References

[1] Jakobson, R., "The sound laws of child language and their place in general phonology", Fifth International Congress of Linguists, Brussels, 1939. [translation by R. Sangster in R. Jakobson [Ed], Studies on Child Language and Aphasia, 7-20, 1971.]

[2] Vihman, M. M., Macken, M. A., Miller, R., Simmons, H., and Miller, J., "A re-assessment of the continuity issue", Language 61(2):397-445, 1985.

[3] de Boysson-Bardies, B., Hallé, P., Sagart, L., and Durand, C., "A cross-linguistic investigation of vowel formants in babbling", Journal of Child Language, 16:1-17, 1989.

[4] Boë, L.-J., and Maeda, S., "Modélisation de la croissance du conduit vocal. Espace vocalique des nouveaux-nés et des adultes. Conséquences pour l'ontogenèse et la phylogenèse", Journées d'Études Linguistiques: La Voyelle dans Tous ces États, Nantes, 98-105, 1997.

[5] Maeda, S., "Une analyse statistique sur les positions de la langue: Étude préliminaire sur les voyelles françaises", 9èmes Journées d'Étude sur la Parole, Lanion, 191-199, 1978.

[6] Edwards, J., and Beckman, M. E., "Methodological questions in studying consonant acquisition", Clinical Linguistics and Phonetics, 22(12):939-958, 2008.

[7] Cohen, A., and Hart, J. 't, "On the anatomy of intonation", Lingua, 19:177-192, 1967.

[8] Stevens, K. N., and House, A. S., "Development of a quantitative description of vowel articulation", Journal of the Acoustical Society of America, 27(3):484-493, 1955.

[9] Peterson, G. E., and Barney, H. L., "Control methods used in a study of the vowels", Journal of the Acoustical Society of America, 24(2):175-184, 1952.

[10] Atal, B. S., and Hanauer, S. L., "Speech analysis and synthesis by linear prediction of the speech wave", Journal of the Acoustical Society of America, 50:637-655, 1971.

[11] Moulines, E., and Charpentier, F., "Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones", Speech Communication, 9, 453-467, 1990.

[12] Boersma, P., "Praat, a system for doing phonetics by computer", Glot International, 5(9-10):341-345, 2001.

[13] Frota, S., Butler, J., Correia, S., Severino, C., and Vigário, M., "Pitch first, stress next? Prosodic effects on word learning in an intonation language", Proceedings of the 36th annual Boston University Conference on Language Development, 120-201, 2012.

[14] Wong, P., "Perceptual evidence for protracted development in monosyllabic Mandarin lexical tone production in preschool children in Taiwan", Journal of the Acoustical Society of America, 133(1):434-443, 2013.

[15] Story, B. H., and Bunton, K., "Production of child-like vowels with nonlinear interaction of glottal flow and vocal tract resonances", Proceedings of Meeting on Acoustics, 19:5pSC2, 2013.

[16] Harshman, R., Ladefoged, P., and Goldstein, L., "Factor analysis of tongue shapes", Journal of the Acoustical Society of America, 62(3):693-707, 1977.

[17] Maeda, S., "Un modèle articulatoire de la langue avec des composantes lineaire", 10èmes Journées d'Étude sur la Parole, Grenoble, 152-162, 1979.

[18] Maeda, S., "Compensatory articulation in speech: Analysis of x-ray data with an articulatory model", First European Conference on Speech Communication and Technology (EUROSPEECH), 2441-2444, 1989.

[19] Maeda, S., "On articulatory and acoustic variabilities", Journal of Phonetics, 19:321-331, 1991.

[20] http://www.cns.bu.edu/~speech/VTCalcs.php

[21] Guenther, F. H., "Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production", Psychological Review, 102(3):594-621, 1995.

[22] http://www.phon.ucl.ac.uk/resource/vtdemo/

[23] Goldstein, U. G., "An articulatory model of the vocal tracts of growing children", Doctoral thesis, Massachusetts Institute of Technology, 1983.

[24] Callan, D. E., Kent, R. D., Guenther, F. H., and Vorperian, H. K., "An auditory-feedback-based neural network model of speech production that is robust to developmental changes in the size and shape of the articulatory system", Journal of Speech Language and Hearing Research, 43:721-736, 2000.

[25] Vorperian, H. K., Wang, S., Chung, M. K., Schimek, E. M., Durtschi, R. B., Kent, R. D., Ziegert, A. J., and Gentry, L. R., "Anatomic development of the oral and pharyngeal portions of the vocal tract: An imaging study", Journal of the Acoustical Society of America, 125(3): 1666-1678, 2009.

[26] Ménard, L., Schwartz, J.-L., and Boë, L.-J., "Auditory normalization of French vowels synthesized by an articulatory model simulating growth from birth to adulthood", Journal of the Acoustical Society of America, 111(4):1892-1905, 2002.

[27] Ménard, L., Schwartz, J. L., and Boë, L. J., "Role of vocal tract morphology in speech development: Perceptual targets and sensorimotor maps for synthesized French vowels from birth to adulthood", Journal of Speech Language and Hearing Research, 47:1059-1080, 2004.

[28] Rvachew, S., Mattock, K., Polka, L., and Ménard, L., "Developmental and cross-linguistic variation in the infant vowel space: The case of Canadian English and Canadian

French", Journal of the Acoustical Society of America, 120(4):2250-2259, 2006.
[29] Serkhane, J. E., Schwartz, J. L., Boë, L. J., Davis, B. L., and Matyear, C. L., "Infants' vocalizations analyzed with an articulatory model: A preliminary report", Journal of Phonetics, 35(3):321-340, 2007.
[30] Ménard, L., Davis, B. L., Boë, L.-J., and Roy J.-P., "Producing American English vowels during vocal tract growth: A perceptual categorization study of synthesized vowels", Journal of Speech Language and Hearing Research, 52:1268-1285, 2009.
[31] Plummer, A. R., Ménard, L., Munson, B., and Beckman, M. E., "Comparing vowel category response surfaces over age-varying maximal vowel spaces within and across language communities", Interspeech, 2013.
[32] Plummer, A. R., Munson, B., Ménard, L., and Beckman, M. E., "Examining the relationship between the interpretation of age and gender across languages", Proc. 21st International Congress on Acoustics, POMA, 19:2aSC36, 2013.
[33] Edwards, J., and Beckman, M. E., "Some cross-linguistic evidence for modulation of implicational universals by language-specific frequency effects in phonological development". Language Learning and Development, 4(2):122-156, 2008.
[34] Beckman, M. E. and Edwards, J., "Generalizing over lexicons to predict consonant mastery", Laboratory Phonology, 1(2):319-343, 2010.
[35] Li, F., "The phonetic development of voiceless sibilant fricatives in English, Japanese and Mandarin Chinese", Doctoral thesis, Ohio State University, 2008.
[36] Kong, E. J., "The development of phonation-type contrasts in plosives: Cross-linguistic perspectives", Doctoral thesis, Ohio State University, 2009.
[37] Monnin, J., "Influence de la langue ambiante sur l'acquisition phonologique: une comparaison du français et du drehu", Doctoral thesis, Université Joseph Fourier, Grenoble I, and Université de Nouvelle-Calédonie, 2010.
[38] Shih, Y., "Taiwanese-Guoyu bilingual children and adults' sibilant fricative production patterns", Doctoral thesis, Ohio State University, 2012.
[39] Bird, S., and Liberman, M., "A formal framework for linguistic annotation", Speech Communication, 33:23-60, 2001.
[40] Li, F., Edwards, J., and Beckman, M. E., "Contrast and covert contrast: The phonetic development of voiceless sibilant fricatives in English and Japanese toddlers", Journal of Phonetics, 37(1):111-124, 2009.
[41] Kong, E. J., Beckman, M. E., and Edwards, J., "Why are Korean tense stops acquired so early? The role of acoustic properties", Journal of Phonetics, 39(2):196-211, 2011.
[42] Kong, E. J., Beckman, M. E., and Edwards, J., "Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese", Journal of Phonetics, 40: 725–744, 2012.
[43] Holliday, J. J., Beckman, M. E., Mays, and C. E., "Did you say susi or shushi? Measuring the emergence of robust fricative contrasts in English- and Japanese-acquiring children", InterSpeech, 2010.
[44] Munson, B., Edwards, J., Schellinger, S. K., Beckman, M. E., and Meyer, M. K., "Deconstructing phonetic transcription: Covert contrast, perceptual bias, and an extraterrestrial view of *Vox Humana*", Clinical Linguistics and Phonetics, 24(4-5):245-260, 2010.
[45] Li, F., Munson, B., Edwards, J., Yoneyama, K., and Hall, K., Language specificity in the perception of voiceless sibilant fricatives in Japanese and English: Implications for cross-language differences in speech-sound development", Journal of the Acoustical Society of America, 129(2): 999-1011, 2011.
[46] Meyer, M. K., "Do attitudes and practice patterns predict the perception of children's speech? Evidence from a web-based audio survey of Speech Language Pathologists", Masters thesis, University of Minnesota, 2012.
[47] Munson, B., Schellinger, S. K., and Urberg Carlson, K., "Measuring speech-sound learning using Visual Analog Scaling", Perspectives on Language Learning and Education, 19(1):19-30, 2012.

Young Children's Performance on Self-administered iPad Language Activities

Jared Bernstein[1], *Ognjen Todic*[2], *Kayla Neumeyer*[1], *Katharyn Schultz*[1], *Liang Zhao*[1]

[1]Knowledge Technologies, Pearson, Menlo Park, California, USA
[2]Keen Research LLC, Mill Valley, California, USA
jared.c.bernstein<>pearson.com, Schultz.katharyn<>gmail.com

Abstract

Twenty-nine English language activities were implemented on a touch-tablet computer. Some activities were focused on a single skill (e.g. reading or speaking), while others involved several integrated skills (e.g. listening and writing). Materials were presented in several modalities, including speech only, speech with figure, silent video, text, speech with text, and speech with text and figure. Response modalities included speech, typing, touch, dragging screen objects, selecting and/or arranging words, and drawing figures. Various test-like sequences of 24 to 45 items were presented to 784 children, 53% from non-English speaking homes. Analysis of over 28,000 responses to these self-administered activities indicates that most activities can be successfully modeled by a single short instructional video example. By age 8 years, nearly all children will respond meaningfully to about 95% of these specific activities. Examples of these activities and child responses are presented.
Index Terms: Language, assessment, touch tablet, four skills, ELL, children, modeling, modality.

1. Introduction

The design and development of instructional systems often proceeds by trial and error. For example, in the United States, various committees formulate standards that describe activities students should be able to perform (e.g. "*Students at all levels of English language proficiency will evaluate [an] author's bias*."), giving finer definitions for different levels, ages, or grades (http://www.wida.us/standards/eld.aspx). Publishers and system developers then design instructional material and tests that align with these standards, even when standards have not been developed with a view to making the defined skills easily measurable. This situation can leave test developers searching for methods that will elicit scoreable samples of student performance within feasible bounds of time and cost. If we understood which kinds of computer-based activities are intuitive for children, we could more accurately plan the development of computer-based assessments and instructional systems.

At the same time, a remarkable proliferation of smart phones and touch-tablet computers, like the Apple iPhone and iPad, and similar devices, has prompted interest in using these devices in education. Touch-tablet computers have touch-sensitive screens, virtual keyboards, accelerometers, and good quality audio I/O. Children are excited to use touch tablets, and their availability opens new possibilities in elicitation and capture of student performance with first and second language, as well as other areas including mathematics. We report an experiment to identify the touch-tablet-enabled presentation and response modalities that children can handle successfully (in 2012 in North America) without any adult guidance or instruction.

2. Approach

Streeter et al. (2011) organize most current applications of technology in language assessment within the grid of Table 1, however touch tablets and smart phones allow several extensions not foreseen in that table of possibilities.

Table 1. *Traditional presentation and response modalities.*

Scoring Focus	Presentation		Response	
	Spoken	Written	Spoken	Written
Declarative Knowledge	+	+	+	+
Language Skills	+	+	+	+

In the present study, a set of 29 different language activities (or test item types) were implemented on an iPad-2 touch-tablet computer. These items were designed to cover many combinations of input and output modalities that are available on a touch-tablet, such that the performance of a young child (age 4-11 years) can, in principle, be measured automatically from the responses. Some activities were designed to elicit information about a single skill (e.g. just reading or speaking), while others elicited performances that reflected several integrated skills (e.g. both listening and writing). Materials were presented in several modalities, including speech only, speech with figure, silent video, text, speech with text, and speech with text and figure. Response modalities included speech, typing, touch, dragging screen objects, selecting and/or arranging words, and drawing figures. Within each activity type, among 15-20 items, certain items were designed for either first or fifth graders (aged 5-6 or 9-10), while other items were designed for use with both groups. Our first analysis of the children's responses was designed to answer three questions:

(1) Which activities can children understand well enough (without adult help) to perform meaningfully on?
(2) Which specific activities yield the most information about a child's relative language skills, and which materials are most appropriate for ages 4-7 and 8-11?
(3) Which activities best discriminate English language learners (**ELL**s) from 'mainstream' students?

2.1. Materials

Apple's iPad was used as the experimental delivery platform because it had a rich Software Development Kit that easily supports multimedia presentation and accepts a number of different gesture controls. A test-like presentation flow was implemented on the iPad, as shown in Figure 1.

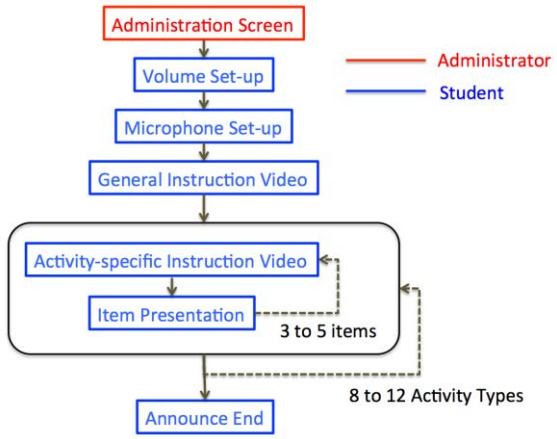


Figure 1: P*resentation-flow of items within Activity Types.*

Twenty-five of the Activity Types studied are listed in Table 1. Some isolate a receptive skill (listening or reading) by using a non-linguistic response mode. The Silent RT Description isolates speaking by presenting a silent video clip (8-15 seconds) for the student to narrate or describe in real time (RT).

Table 2. *Activities, with modalities and language skills (**L**istening, **S**peaking, **R**eading, **W**riting, **U**sage).*

Activity Types (Information Rank)	Elicit	Respond	Skills
Repeat	Voice	Speak	L,S
Silent RT Description **(3)**	Silent Video	Speak	S
Oral Read & Answer **(1)**	Text	Speak	R,S,L
Situated Polite Request **(4)**	Voice, Figure	Speak	L,S,U
Seeded Sentence	Voice, Text	Speak	R,S
Passage retell	Text, Voice	Speak	R,S
Teacher Teach	Video, Voice	Speak	L,S
Hear-Touch	Voice, Figure	Touch	L
Hear-Point Conversation	Voice, Figure	Touch	L
Hear-Move	Voice, Figure	Gesture	L
Hear-Path	Voice, Figure	Gesture	L
Anaphora	Text	Touch	L,R
Read-Move	Text	Gesture	R
Place Words	Text	Gesture	R
Passage Read & Answer	Text	Type	R,W
Non-word Read Aloud	Text	Speak	R,S
Recognize Word	Text	Touch	R
Recognize Non-word	Text	Touch	R
Insert Adverbial	Text	Gesture	R,W,U
Word Choice	Text	Gesture	R,W,U
Sentence Cloze	Text	Type	R,W,U
Find Error	Text	Touch	R,U
Spell Words **(5)**	Voice, Figure	Type	W
Gr1 Passage, write-word	Text	Type	R,W
Gr5 Passage re-write **(2)**	Text	Type	R,W

For example, in a **Hear-Point Conversation**, the student hears a conversation between two voices, during which two or more concrete objects in the figure are mentioned in passing. The student has been instructed to touch any object as it is mentioned in the running conversation. See Figure 2.

2.2. Procedures

Each child sat before an iPad on a table-top stand, in a small room. Typically 3 to 8 children were doing the activities individually in the same room, but not in synchrony with each other. The audio was presented via the iPad's built-in speaker and spoken responses were recorded through the built-in microphone. After an administrator entered the child's ID and selected an age-appropriate version to run, students were left alone to figure out what the task demands were. They had been told that the iPad gave a test, but most referred to it as a game. Each student encountered a sequence of activities, with 3 to 5 items of each activity type presented, following a single 10-20 second video showing a child doing a sample item (see Figure 1). Most children aged 4-7 encountered 36 items, while those aged 8-11 most often had 43 items to do.

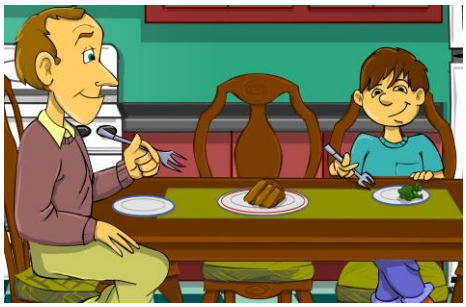


Figure 2: *Example Hear-Point Conversation figure. Several objects are touchable (e.g. dad, chair, cake, boy, fork), although only two are mentioned in the audio conversation.*

3. Results

In July-August 2012, we ran 326 students; then we revised some items, introduced several new activities, and in October 2012, ran another 458 students (N=784). Almost all students who participated were from low-income homes, with 40% currently in official ELL status, but only 47% listed English as their home language. Specific results:

(1) Which activities work? 27 of 29 activities elicited useful responses from > 85% of students aged 8-11. Among students 4-7, 24 of 29 activities elicited useful responses from at least 75% of the students.

(2) Which yield the most skill information? The best are Oral Read Passage, Situated Polite Request, and Silent Video RT Description, with traditional Non-word Read Aloud and Spell Word nearly as good.

(3) Which discriminate ELLs from other students? Items types yielding the most information are noted in Table 2. Across all activities and ages, the ELL-nonELL score difference averaged about 3% of the observed score range. The items that best discriminated ELLs were Repeats, Read-Moves, Hear-Moves, Polite Requests, and Hear-Paths.

In conclusion, most activities can be successfully modeled by a single short video example. In this U.S. sample, by age 8 years, children respond meaningfully to almost all these items about **95%** of the time, regardless of first language.

4. References

[1] L.Streeter, J.Bernstein, P.Foltz, & D.DeLand, D. (2011). *Pearson's Automated Scoring of Writing, Speaking and Mathematics*. (pdf) at http://kt.pearsonassessments.com

A Survey about ASR for Children

Felix Claus[1], *Hamurabi Gamboa Rosales*[2], *Rico Petrick*[3], *Horst-Udo Hain*[3], *Rüdiger Hoffmann*[1]

[1]Dresden University of Technology, Chair for System Theory and Speech Technology, 01062 Dresden, Germany
[2]Autonomous University of Zacatecas, 98000, Zacatecas, Mexico
[3]Linguwerk GmbH, Research & Development, 01069 Dresden, Germany
felix_claus@gmx.de, hamurabigr@uaz.edu.mx, Ruediger.Hoffmann@tu-dresden.de
[rico.petrick,udo.hain]@linguwerk.de

Abstract

This paper is intended to survey the state of the art of automatic speech recognition (ASR) for children's speech. Investigating ASR for children is a current trend in research. Therefore databases of children's speech are needed for training and testing of ASR systems. In the first part of this paper the most relevant databases of children's speech are described. There are less speech data of children available than of adults and speech of preschool children is even more rarely available.

In the second part of this paper the common techniques for recognizing children's speech are summarized. Most investigations about children's ASR focus on the acoustic model. The common methods are described and approaches regarding the lexical and speech model are mentioned subsequently.

In an extensive literature research we collected papers investigating ASR for children. Several studies have been carried out investigating children's ASR. Due to the lack of data from preschool children only a few investigations for this age group have been accomplished. This is illustrated by presenting a statistic on the age of the children in past studies.

Index Terms: children's speech, preschool children's speech, ASR for children, child computer interaction, statistics on children's speech, children's speech corpora

1. Introduction

Most applications using speech interfaces are mainly designed for adults and therefore use automatic speech recognition (ASR) systems for adults' speech. Examples are: speech interfaces of navigation systems, mobile phones or dictation systems for the computer. In recent years systems for children, like reading tutors, tools for foreign language learning or computer games came up. Therefore children's speech recognition became more relevant.

Recognition accuracy of children's speech is usually lower than for adults [1, 2, 3], which is caused by the differences between children's and adults' speech [2]. There exist differences due to anatomical differences and differences in linguistic skills. The shorter vocal tracts of children cause higher formant frequencies and the smaller and lighter vocal folds lead to a higher fundamental frequency. Furthermore, linguistic skills of children are poorer than those of adults. Especially young children are not able to articulate all phonemes correctly. Even if they are able to pronounce single phonemes right they pronounce some words wrong. In [4] D'Arcy and Russel investigated the human perception. They compared the recognition accuracy of ASR systems and human listeners for recognizing children's and adults' speech. The results show that for both listeners, ASR systems and humans, recognition accuracy is worse for recognizing children's speech than for recognizing adults'.

Regardless of the difficulties a huge market for potential applications using speech recognition for children can be assumed. In 2002, Narayanan published a study about creating conversational interfaces for children [5]. For the motivation of his study he pointed out how firm children are in using the computer and that they would like to operate with the computer by a speech interface. 60 % of the U. S. children, aged between four and eleven years, use a computer at home, where only 40 % of the adults does. Recent German studies [6] corroborate these results and show the habits of German children in the usage of computers and mobile phones. 80 % of the children, aged six years and older, have access to a computer and 23 % of the four to five years old children have experiences with computers, too. Children are curious to deal with new technologies and consequently they are potential users for applications using speech recognition.

In order to improve ASR for children databases of children's speech and further research are required. Currently there are less databases of children's speech available than of adults'. One reason is that recording children's speech is more difficult than recording adults' and it becomes even more challenging with decreasing age of the children [7]. Several studies were made investigating ASR for children. Due to available databases most of these studies focus on school children and only a few studies were made for recognizing preschool children's speech.

The paper is structured as following. In the first part we survey the most relevant databases consisting of speech from school children as well as from preschool children. Thereafter current methods for recognizing children's speech are described for the acoustic, the lexical and the speech model. Furthermore a statistic about the age of children in past studies about children's ASR is presented and the trends in research are shown.

2. Databases

Databases are needed for training and testing of ASR systems. There exists less data of children's speech than of adults'. One reason is that recording children's speech is more difficult than recording adults' and it gets even more difficult with decreasing age of the children [7]. An extensive overview about existing databases of children's speech can be found in [8].

2.1. Databases of school children

Most databases consist of speech from children aged between 6 and 18 years. The language of the databases is mainly English, German, Italian, Swedish and Dutch. Databases in other languages are more rarely available. The most relevant corpora are:

- Tball corpus (non-native English from native Spanish, 256 children, aged between 5 and 8 years) [7],
- CID children's speech corpus (American English, read speech, 436 children aged between 5 and 17 years) [9],
- CU Kid's Prompted and Read Speech corpus (American English, read speech, 663 children, aged between 4 and 11 years) [10],
- CU Kid's Read and Summarized Story corpus (American English, spontaneous speech, 326 children, aged between 6 and 11 years) [11],
- CMU Kid's speech corpus (American English, read speech, 76 children, aged between 6 and 11 years) [12],
- OGI Kid's speech corpus (English, read speech, 1100 children, aged between 5 and 15 years) [13],
- PF-STAR corpus (multilingual, including English, German, Swedish and Italian, 491 children, aged between 4 and 15 years, including spontaneous and emotional speech corpus FAU-AIBO) [14],
- ChildIt corpus (Italian, 171 children, aged between 7 and 13 years) [15],
- CHOREC corpus (Dutch, read speech, 400 children, aged between 6 and 12 years) [16] and
- JASMIN-CGN corpus (Dutch, read and spontaneous speech of native and non-native speakers, more than 60 hours of speech from children aged between 7 and 16 years) [17].

2.2. Databases of preschool children

As mentioned above recording children's speech is more difficult than recording adults'. Especially recording preschool children is time-consuming since the children are not able to read. Therefore alternative methods have to be applied to obtain the recordings. Furthermore young children can concentrate for only a short period (5...10 min) [3]. Hence less speech than of older children or adults can be recorded within one recording session. Accordingly, there exists significantly less data of young children than of older ones and in most cases the quality of the data is inferior to those of older children or adults.

Most preschool children's speech data exists in the context of the project CHILDES (Child Language Data Exchange System) [18] from the Carnegie Mellon University. CHILDES is part of the TalkBank system for sharing and studying conversational interactions and consists of data from more than 100 corpora of different languages. The multilingual PHON corpus is one of these corpora. It is meant to be used in order to study the phonological development of children. German data attached to PHON is published in [19]. Data from ten children are included and analyzed in detail. Six of these children are recorded from the 5th to the 36th month and four children are recorded from the 36th month to eight years. More details can be found in [19]. Regrettably, for most of the data from CHILDES only the transcript is publicly available, but without media. More data of young children's speech is recorded with the LENA device [20] or in the context of the SpeechHome project [21]. Unfortunately, these databases are not publicly available, too.

3. Acoustic model

Most work dealing with ASR for children focuses on the acoustic model. In this section existing appendages are described. More details can be found in [15, 22, 23].

3.1. Training with children's speech

In order to obtain the best recognition performance the data used for the training of the speech recognizer should be akin to the data which has to be recognized. Therefore recognizers should be trained with children's speech in order to recognize it. But it depends on the language and the considered age group whether enough data is available to train a hidden markov model (HMM) recognizer.

In 1996, there was a key study made by Wilpon and Jacobsen [1]. They wanted to recognize speech of different age groups. Therefore they created different acoustic models, one per each age group. The recognition performance was always the best, when the acoustic model was trained with speech from the same age group. Additionally they noticed that the recognition performance for children's speech is not as good as for adults, even when the acoustic model was trained with children's speech. The word error rates they achieved were 1.9 % for adults from 35 to 59 years and 4.7 % for children from 8 to 12 years, which is remarkable low compared to other studies.

Work of Hagen et al. [24] as well as work of D'Arcy et al. [25] confirmed the results of Wilpon and Jacobsen. They created different acoustic models for different age groups of children and received the best results when training and testing data were from the same age group. The recognition performance also decreased with decreasing age of the children. The disadvantage of this approach is that in most cases there is not enough children's speech data available to train every age group separately. So often it is used to create one acoustic model for children in general [26, 27].

In [28], an automatic reading tutor system is presented, which is trained with children's speech. Later the system is ported on two hand-held devices [29]. The recognizer used for the system is trained with four children's speech databases: CU Kid's Prompted and Read Speech corpus, CU Kid's Read and Summarized Story corpus, OGI Kid's speech corpus and CMU Kid's speech corpus (see section 2). The test data is a subset of Kid's Read and Summarized Story corpus, which was excluded from the training data. With this approach a WER of 11.45 % was achieved.

3.2. Training with adults' speech and VTLN

If there is not enough children's speech data available to train the recognizer, adults' speech is used and the differences in the positions of the formant frequencies are compensated by vocal tract length normalization (VTLN).

This was already done in early studies in 1977 [30]. Many other studies approve that recognition performance of children's speech after training with adults' speech can be increased by applying VTLN methods [26, 27, 31, 32, 33, 34]. For example in [26] the word error rate can be decreased from 15.9 % to 8.7 %. Often the implementation is simple and a general linear, piecewise linear or bilinear distortion function is used. But also more extensive techniques like a phoneme depended distortion function [35] or the distortion with a distortion matrix instead of a simple scale factor [36] are utilized. Further appendages can be found in [37].

3.3. Training with adults' speech and adaptation to children's speech

Due to the fact that there are more differences between children's and adults' speech than only the positions of the formant frequencies [38] adaptation techniques like maximum likelihood linear regression (MLLR), speaker adaptive training (SAT) or maximum a posteriori adaptation (MAP) are used in order to create age depended acoustic models. Several works [31, 33, 39] show that the recognition accuracy could be increased by applying these methods. For example in [39] the word error rate after using VTLN was 10.9 % and could be further decreased to 8.0 % by using adaptation techniques.

4. Lexical and speech model

4.1. Lexical model

A further appendage to improve the recognition performance of children's speech is pronunciation modeling. Young children and children with a poor pronunciation do not use a canonical pronunciation. In these cases it is subsidiary to adapt the lexical model. In [2] user-depended pronunciation dictionaries are employed. Depending on the pronunciation of the child, the recognition performance could be raised in a small range. For a child estimated to have a good pronunciation the word accuracy could be raised from 75.83 % to 76.89 % and for a child estimated to have a poor pronunciation the word accuracy could be raised from 35.47 % to 43.92 %.

In order to improve the recognition performance for recognizing Japanese preschool children, age-depended pronunciation dictionaries are applied in the Takemaru-kun system [40]. They are created by manually adding pronunciation variants of the children from the training data. For example, the word 'Takemaru' is often pronounced as 'Tachimaru', 'Takebaru' or 'Takemau'. These pronunciation variants are added to the pronunciation dictionary and the recognition accuracy could be increased from 51.2 % to 54.7 %. This increase disappears when using acoustic models of children. The authors constitute that in this case, the specifics of children's speech are modeled within the acoustic model already, so there is no room for further improvement through pronunciation modeling.

4.2. Speech model

The use of specific speech models has also been investigated in several studies [5, 27, 40]. Either the speech model is extracted from domain-specific texts or children are recorded during the use of respective systems, wherefrom the speech model is extracted. In [5] the word error rate could be decreased upon 5 to 20 % relatively, according to a word error rate of 22 %.

5. Literature research

5.1. Statistics on the age

In order to obtain the state of the art in children's ASR we did an extensive literature research. We collected over 1000 papers about children's speech recognition and related topics like VTLN and adaption techniques. For this purpose we collected eligible papers from conference proceedings of Interspeech, ICASSP and further conferences with regard to speech processing. Studies, published in the diverse fields of science like medicine or pedagogics, are not considered. Additionally IEEE Xplore and Google Scholar were browsed for papers. In 100 of these 1000 papers aspects of Children's ASR are investigated. These 100 papers are analyzed and used for the statistic presented in this section.

In figure 1 the ages of the children used in the investigations are shown. Every paper is represented by a vertical line. The length of each line specifies the range of the data. Lines next to each other with the same length and the same covered range indicate that the same data is used for the investigations.

Publications are available for children from three years forward with the exception of [41]. In this study analysis and automatic estimation of children's subglottal resonances, which may benefit children's ASR, are investigated for children from birth on. Most studies deal with speech from children in school age. This is a matter of fact of available databases. Some of these studies use data including also speech from preschool children [41, 42, 43, 44]. However most of the data is from older children, and therefore the studies do not focus on preschool children. To the best of our knowledge only a few studies were made for preschool children only (highlighted in figure 1). These studies are [3, 40, 45, 46, 47, 48]. Already in 1993, Strommen published an investigation about simple ASR experiments on preschool children users aged three years [45]. Unfortunately, further investigations of Strommen do not focus on ASR for preschool children. In [3, 46] children's ASR is investigated using a little German database of children aged between three and six years. In [40] Cincarek et al. describe the development of a module for speech recognition and answer generation for preschool children for a speech-oriented guidance system. For their investigations they use data from the Takemaru database consisting of spontaneous speech from Japanese children. Unfortunately, the ages of the children are not documented exactly. Further research on preschool children's speech is published in [47], where the authors investigated reference marking in children's computer-directed speech using a Wizard of Oz scenario for children from three to six years. In this context data was recorded which was used in [48] for automatic detection of disfluency boundaries in spontaneous young children's speech.

5.2. Trends in research

One trend in research is to analyze speech from preschool children. For example Marklund investigated the phonological complexity and vocabulary size in 30-month-old Swedish children [49]. Further research is done on the recognition of children's emotional speech [50, 51, 52]. Most studies dealing with that use the FAU-AIBO corpus. Additionally research is done according to existing applications like reading tutors, tools for pronunciation training or medical assessment [44, 53, 54, 55]. Since the performance of speech recognizers for recognition of children's speech is still lower than for adults, recognition of children's speech in general is also a current field of research. Recent studies are [23, 56, 57].

6. Conclusions

Due to current applications like reading tutors, tools for foreign language learning or computer games children's ASR became more relevant in recent years. But results for recognizing children's speech are worse than for recognizing adults'. Several studies have been carried out in order to improve children's speech recognition. Most work focuses on the acoustic model. The recognition accuracy for children is higher if the recognizer is trained with children's speech. When not enough child data is available VTLN and adaption methods are applied to increase

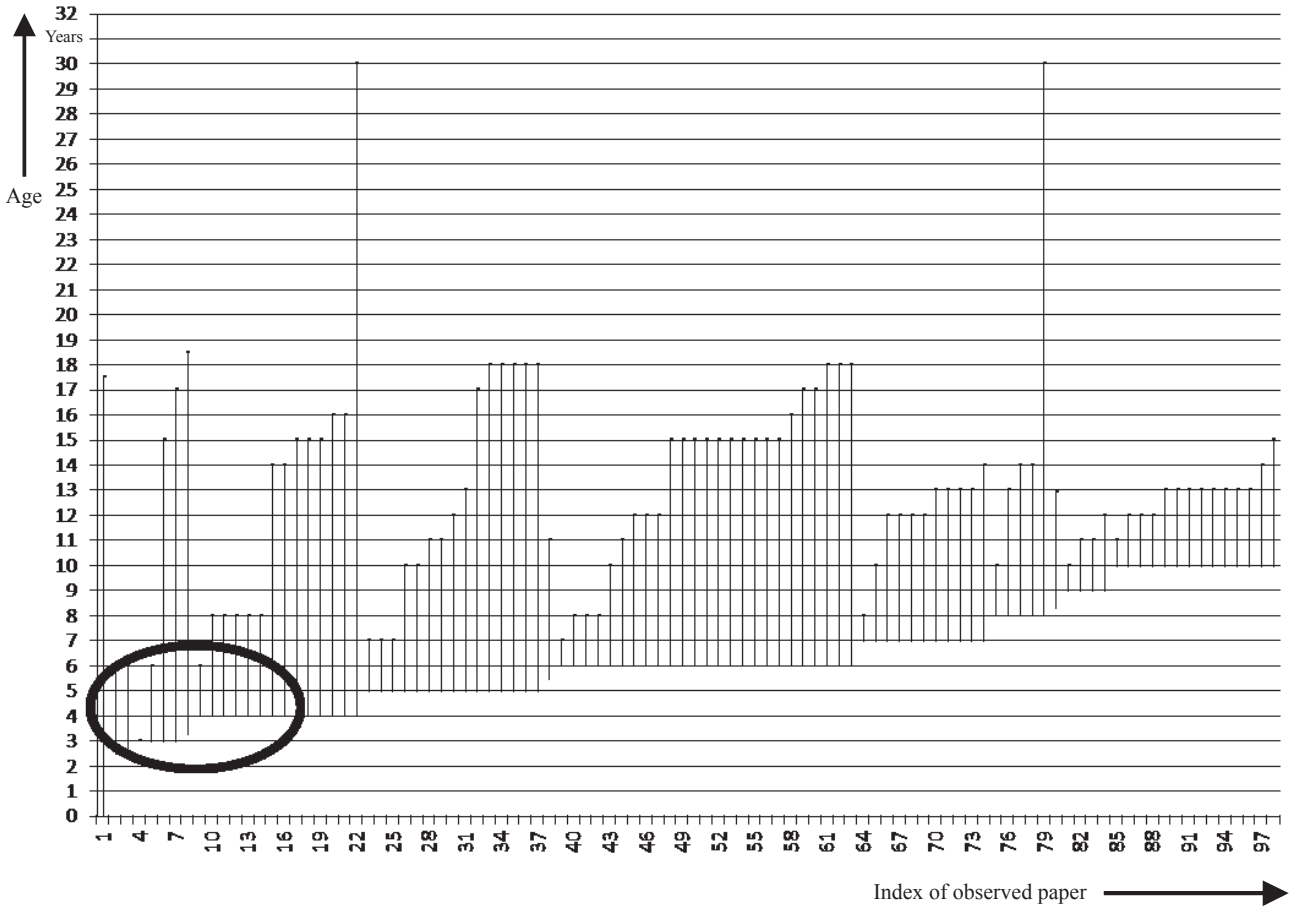


Figure 1: *Considered age group in investigations related to children's ASR.*

the recognition accuracy. Further studies have been carried out on the adaption of the lexical and the speech model. The recognition accuracy could be increased with the described methods but nevertheless results for recognizing children's speech are worse than for recognizing adults'.

Further research is needed in order to improve ASR for children. Therefore databases are required. But databases of children's speech are rare and databases of preschool children are even more rarely available. Our statistic about the age of children in past studies corroborates the lack of young children's speech data (3. . . 6 years) which are eligible for children's ASR. Accordingly, further databases have to be developed in future.

7. References

[1] J.G. Wilpon and C.N. Jacobsen, "A study of speech recognition for children and the elderly," in *Proc. of ICASSP*, 1996.

[2] Q. Li and M.J. Russell, "An analysis of the causes of increased error rates in children's speech recognition," in *Proc. of ICSLP*, 2002.

[3] K. Matthes, F. Claus, H.-U. Hain, and R. Petrick, "Herausforderungen an Sprachinterfaces für Kinder," in *Proc. of ESSV*, 2010.

[4] S.M. D'Arcy and M.J. Russell, "A comparison of human and computer recognition accuracy for children's speech," in *Proc. of Interspeech*, pp. 2197-2200, 2005.

[5] S.S. Narayanan and A. Potamianos, "Creating conversational interfaces for children," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 2, pp. 65-78, February 2002.

[6] Egmont-mediasolutions, "KidsVerbraucherAnalyse 2012," 2012.

[7] A. Kazemzadeh, H. You, M. Iseli, and B. Jones, "Tball data collection: the making of a young children's speech corpus," in *Proc. of Interspeech*, 2005.

[8] F. Claus, H. Gamboa Rosales, R. Petrick, H.-U. Hain, and R. Hoffmann, "A Survey about Databases of Children's Speech," in *Proc. of Interspeech*, 2013.

[9] S. Lee and A. Potamianos, "Acoustics of children's speech: Developmental changes of temporal and spectral parameters," *Journal of the Acoustical Society of America*, vol. 105, no. 3, pp. 1455-1468, March 1999.

[10] R. Cole, J.-P. Hosom, and B. Pellom, "University of Colorado Prompted and Read Children's Speech Corpus," *Technical Report TR-CSLR-2006-02*, University of Colorado, 2006.

[11] R. Cole and B. Pellom, "University of Colorado Read and Summarized Stories Corpus," *Technical Report TR-CSLR- 2006-03*, University of Colorado, 2006.

[12] M. Eskenazi, "Kids: a database of children's speech," *Journal of the Acoustical Society of America*, vol. 100, no. 4, 1996.

[13] K. Shobaki, J.-P. Hosom, and R. Cole, "The OGI kids' speech corpus and recognizers," in *Proc. of ICSLP*, 2000.

[14] A. Batliner, M. Blomberg, and S.M. D'Arcy, "The PF-STAR Children's Speech Corpus," in *Proc. of Interspeech*, pp. 2761-2764, 2005.

[15] M. Gerosa, *Acoustic Modeling for Automatic Recognition of Children's Speech*, Ph.D. thesis, University of Trento, 2006.

[16] L. Cleuren, J. Duchateau, P. Ghesquiere, and H. Van Hamme, "Children's Oral Reading Corpus (CHOREC): Description and Assessment of Annotator Agreement," in *Proc. of LREC*, 2008.

[17] C. Cucchiarini, J. Driesen, H. Van hamme, and E. Sanders, "Recording Speech of Children, Non-Natives and Elderly People for HLT Applications: the JASMIN-CGN Corpus," in *Proc. of LREC*, 2008.

[18] B. MacWhinney, "The CHILDES Project: Tools for Analyzing Talk," *Lawrence Erlbaum Associates*, 2000.

[19] B. Möbius, "Ein exemplartheoretisches Modell zum Erwerb der akustischen Korrelate der Betonung," *DFG-Abschlussbericht*, 2007.

[20] Project LENA, http://www.lenafoundation.org, 2013.

[21] SpeechHome project, http://www.media.mit.edu/cogmac/projects/hsp.html, 2013.

[22] C. Hacker, *Automatic assessment of children speech to support language learning*, Ph.D. thesis, University of Erlangen-Nuremberg, 2009.

[23] D. Elenius, *Accounting for Individual Speaker Properties in Automatic Speech Recognition*, Ph.D. thesis, KTH Stockholm, 2010.

[24] A. Hagen, B. Pellom, and R. Cole, "Children's speech recognition with application to interactive books and tutors," in *Proc. of Workshop on Automatic Speech Recognition and Understanding*, 2003.

[25] S.M D'Arcy, L.P. Wong, and M.J. Russell, "Recognition of read and spontaneous children's speech using two new corpora," in *Proc. of ICSLP*, 2004.

[26] A. Potamianos, S.S. Narayanan, and S. Lee, "Automatic speech recognition for children," in *Proc. of Eurospeech*, 1997.

[27] S. Das, D. Nix, and M. Picheny, "Improvements in children's speech recognition performance," in *Proc. of ICASSP*, 1998.

[28] X. Li, Y.-C. Ju, L. Deng, and A. Acero, "Efficient and robust language modeling in an automatic children's reading tutor system," in *Proc. of ICASSP*, pp. 193-196, 2007.

[29] X. Li, L. Deng, Y.-C. Ju, and A. Acero, "Automatic children's reading tutor on hand-held devices," in *Proc. of Interspeech*, pp. 1733-1736, 2008.

[30] H. Wakita, "Normalization of vowels by vocal-tract length and its application to vowel identification," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 25, no. 2, pp. 183–192, April 1977.

[31] D. Elenius and M. Blomberg, "Adaptation and normalization experiments in speech recognition for 4 to 8 year old children," in *Proc. of Interspeech*, pp. 2749-2752, 2005.

[32] D.C. Burnett and M. Fanty, "Rapid unsupervised adaptation to children's speech on a connected-digit task," in *Proc. of ICSLP*, pp. 1145-1148, 1996.

[33] M. Gerosa, D. Giuliani, and F. Brugnara, "Acoustic variability and automatic recognition of children's speech," *Speech Communication*, 2007.

[34] O. Jokisch, H.-U. Hain, R. Petrick, and Rüdiger Hoffmann, "Robustness optimization of a speech interface for child-directed embedded language tutoring," in *Proc. of Workshop on Child, Computer, and Interaction*, 2009.

[35] A. Potamianos and S.S. Narayanan, "Robust recognition of children's speech," *IEEE Transactions on Speech and Audio Processing*, vol. 11, no. 6, pp. 603-616, November 2003.

[36] D. Saito, R. Matsuura, and S. Asakawa, "Directional dependency of cepstrum on vocal tract length," in *Proc. of ICASSP*, pp. 4485–4488, 2008.

[37] S. Molau, *Normalization in the acoustic feature space for improved speech recognition*, Ph.D. thesis, University of Aachen, 2003.

[38] M. Gerosa, D. Giuliani, S.S. Narayanan, and A. Potamianos, "A review of ASR technologies for children's speech," in *Proc. of Workshop on Child, Computer, and Interaction*, 2009.

[39] A. Hagen, B. Pellom, S. Van Vuuren, and R. Cole, "Advances in children's speech recognition within an interactive literacy tutor," in *Proc. of NAACL HLT*, 2004.

[40] T. Cincarek, I. Shindo, T. Toda, H. Saruwatari, and K. Shikano, "Development of Preschool Children Subsystem for ASR and Q&A in a Real-Environment Speech-Oriented Guidance Task," in *Proc. of Interspeech*, 2007.

[41] S.M. Lulich, H. Arsikere, J.R. Morton, G.K. Leung, A. Alwan, and M.S. Sommers, "Analysis and automatic estimation of children's subglottal resonances," in *Proc. of Interspeech*, pp. 2817–2820, 2011.

[42] J. Gustafson and K. Sjölander, "Voice transformations for improving children's speech recognition in a publicly available dialogue system," in *Proc. of ICSLP*, 2002.

[43] W.R. Rodríguez and E. Lleida, "Formant Estimation in Children's Speech and its application for a Spanish Speech Therapy Tool," in *Proc. of Workshop on Speech and Language Technology in Education*, 2009.

[44] T. Bocklet, A. Maier, U. Eysholdt, and E. Nöth, "Improvement of a speech recognizer for standardized medical assessment of children's speech by integration of prior knowledge", in *Proc. of Spoken Language Technology Workshop*, pp. 259–264, 2010.

[45] E.F. Strommen and F.S. Frome, "Talking back to big bird: Preschool users and a simple speech recognition system," *Educational Technology Research and Development*, vol. 41, no. 1, pp. 5-16, 1993.

[46] F. Claus, *Integrierte Spracherkennung für Kindersprache: Evaluierung phonembasierter Spracherkenner*, diploma thesis, Hochschule für Technik und Wirtschaft Dresden (FH), 2010.

[47] S. Montanari, S. Yildirim, E. Andersen, and S.S. Narayanan, "Reference Marking in Children's Computer-Directed Speech: An Integrated Analysis of Discourse and Gestures," in *Proc. of ICSLP*, 2004.

[48] S. Yildirim and S.S. Narayanan, "Automatic Detection of Disfluency Boundaries in Spontaneous Speech of Children Using Audio Visual Information," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 2–12, January 2009.

[49] U. Marklund, U. Sundberg, I.-C. Schwarz, and F. Lacerda, "Phonological complexity and vocabulary size in 30-month-old Swedish children," in *Proc. of Interspeech*, 2012.

[50] S. Planet and I. Iriondo, "Spontaneous children's emotion recognition by categorical classification of acoustic features," in *Proc. of CISTI*, 2011.

[51] Z. Zhang and B. Schuller, "Active Learning by Sparse Instance Tracking and Classifier Confidence in Acoustic Emotion Recognition," in *Proc. of Interspeech*, 2012.

[52] N. Ding, V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in emotion recognition – an adaption based approach," in *Proc. of ICASSP*, pp. 5101-5104, 2012.

[53] E. Ylmaz, D. Van Compernolle, and H. Van hamme, "Robust tracking for automatic reading tutors," in *Proc. of Interspeech*, 2012.

[54] M.P. Black and S.S. Narayanan, "Improvements in predicting children's overall reading ability by modeling variability in evaluators subjective judgements," in *Proc. of ICASSP*, pp. 5069-5072, 2012.

[55] S.-C. Yin, R. Rose, and Y. Tang, "Verifying session level pronunciation accuracy in a speech therapy application," in *Proc. of Interspeech*, 2012.

[56] S. Ghai and R. Sinha, "Analyzing pitch robustness of PMVDR and MFCC features for children's speech recognition," in *Proc. of SPCOM*, 2010.

[57] S. Ghai and R. Sinha, "A Study on the Effect of Pitch on LPCC and PLPC Features for Children's ASR in Comparison to MFCC," in *Proc. of Interspeech*, pp. 2589-2592, 2011.

A Multimodal Educational Game for 3-10-Year-Old Children: Collecting and Automatically Recognising European Portuguese Children's Speech

Annika Hämäläinen[1,2], *Fernando Miguel Pinto*[1], *Silvia Rodrigues*[1], *Ana Júdice*[1], *Sandra Morgado Silva*[3], *António Calado*[1], *Miguel Sales Dias*[1,2]

[1]Microsoft Language Development Center, Lisbon, Portugal
[2]ADETTI – ISCTE, IUL, Lisbon, Portugal
[3]Diferente Jogo, Caldas da Rainha, Portugal

{t-anhama, a-fpinto, v-antonc, Miguel.Dias}@microsoft.com, sandrasilva@diferencas.net

Abstract

Speech interfaces have tremendous potential in education. In this paper, we present our work in the Contents for Next Generation Networks project, an ongoing Portuguese industry-academia collaboration developing a multimodal educational game aimed at improving the physical coordination and the basic mathematical and musical skills of 3-10-year-old children. We focus on our work in the area of children's speech recognition: designing, collecting, transcribing and annotating a 21-hour corpus of prompted European Portuguese children's speech, as well as our first experiments with different acoustic modelling approaches. Our speech recognition results suggest that training children's speech models from scratch is a more promising approach than retraining adult speech models using children's speech when a sufficient amount of training data is available from the targeted age group. This finding also holds for adult female speech models retrained using children's speech. As compared with a baseline recogniser comprising gender-dependent adult speech models, the best-performing children's speech models that we have trained so far – gender-independent cross-word triphones trained with 17.5 hours of speech from 3-10-year-old children – resulted in a 45-percent (relative) decrease in word error rate in a task expecting isolated cardinal numbers, sequences of cardinal numbers or musical notes as speech input.

Index Terms: acoustic modelling, ASR, child-computer interaction, corpus, educational game, European Portuguese

1. Introduction

Speech interfaces have tremendous potential in the education of children. Speech provides a natural modality for child-computer interaction and can, at its best, contribute to a fun, motivating and engaging way of learning [1]. However, it is well known that automatically recognising children's speech is a very challenging task. Recognisers trained on adults' speech tend to suffer from a substantial deterioration in recognition performance when used by children [1-6]. Moreover, word error rates (WERs) on children's speech are usually much higher than those on adults' speech even when using a recogniser trained on age-specific speech – although they do show a gradual decrease as the children get older [1-7].

The difficulty of automatically recognising children's speech can be attributed to it being acoustically and linguistically very different from adults' speech [1, 2]. For instance, due to their vocal tracts being smaller, the fundamental and formant frequencies of children's speech are higher [1, 2, 7-9]. What is particularly characteristic of children's speech is its higher variability as compared with adults' speech, both within and across speakers [1, 2]. This variability is caused by rapid developmental changes in their anatomy, speech production et cetera, and manifests itself, for example, in speech rate, the degree of spontaneity, the frequency of disfluencies, fundamental and formant frequencies, and pronunciation quality [1, 2, 7-11].

In the context of education, speech interfaces have been developed, for instance, for interactive reading and pronunciation tutoring [1, 2, 12-14]. However, in the case of less-spoken languages, sufficiently large corpora of children's speech are often not available for developing speech-driven educational applications. This has, for instance, been the case for European Portuguese (pt-PT). In this paper, we describe our speech-related work in the Contents for Next Generation Networks (CNG) project, whose end product is a multimodal educational game for 3-10-year-old Portuguese children: developing of a 21-hour corpus of prompted children's speech, and modelling children's speech for automatic speech recognition (ASR) purposes.

The paper is further organised as follows. In Section 2, we introduce the CNG project and the educational game that the project partners are developing. We present the design, collection, transcription and annotation of the CNG Corpus of European Portuguese Children's Speech in Section 3 and, in Section 4, describe the children's speech recognition experiments that we have carried out so far and intend to perform in the future. Finally, in Section 5, we formulate our conclusions.

2. The CNG project and educational game

The CNG project is an ongoing Portuguese industry-academia collaboration that studies speech and gesture as natural alternatives for child-computer interaction in the education of 3-10-year-old children. To this end, the project partners are developing a multimodal educational game that can be played in an immersive virtual environment (a CAVE, [15]) or with a desktop computer using Kinect for Windows, a motion sensing input device by Microsoft [16]. The game addresses two main areas of development: motor skill development (physical coordination skills) and cognitive development (attention, problem solving, mathematical and musical skills). In the game, the children will use their voice and gestures to complete different kinds of puzzles and tasks in 3D and 2D scenarios with educational themes (e.g. the Age of Discovery, dinosaurs). An intelligent virtual assistant with a synthesised voice will guide and help them throughout the game. In one of the themed scenarios, the Age of Discovery, the assistant might, for instance, say, "*Cheer up the sailors by counting from one to*

five and moving your body. With each number you say, you will need to move a part of your body." Speech input will be enabled for tasks related to mathematics (e.g. counting objects, simple mathematical operations) and music (e.g. completing musical note sequences). The expected speech input includes isolated cardinals, sequences of cardinals, and musical notes. The difficulty level of the game can be set up manually before the game starts but will also be adjusted automatically based on the children's performance.

The CNG game is developed for pt-PT. To the best of our knowledge, the only other speech-driven educational application for pt-PT is a speech therapy system that uses games to identify the phones that 5-6-year-old children have problem pronouncing, and to help them overcome their pronunciation problems [17]. For developing that system, a 158-minute corpus of isolated words was collected from 111 children belonging to the targeted age group. pt-PT children's speech is also available in the Portuguese Speecon Database [18]. However, it only contains speech from 52 children, of which only 15 children aged 8-10 belong to the age group targeted in the CNG project; the rest are 11-14-year olds.

3. The CNG Corpus

When it comes to speech material, the goal of the CNG project was to develop a corpus of about 20 hours of children's speech suitable for training and testing acoustic models (AMs) for the speech-driven parts of the CNG game. The resulting corpus is called the CNG Corpus of European Portuguese Children's Speech. The following subsections describe the design, collection, transcription and annotation of the corpus, as well as the details of the full corpus and the datasets used for the ASR experiments.

3.1. Corpus design

3.1.1. Speaker selection

As the CNG game is aimed at 3-10-year-old children, we only collected speech from speakers in that age range. Based on the children's capabilities (see Section 3.1.2), we split them into two age groups that are considered homogenous populations for the purposes of the CNG project: 3-6-year-old and 7-10-year-old children.

We collected speech from children attending nurseries and schools in and around the Portuguese cities of Lisbon, Leiria and Aveiro. The cities were chosen for practical reasons; the time and budget available for the data collection campaign were tight, so we had to concentrate on places where the project partners had existing contacts at nurseries and schools and were able to easily attend recording sessions. We tried to keep the ratio of girls and boys as even as possible but were not, for instance, able to aim at a specific ratio of speakers from the different areas.

3.1.2. Prompt design

Collecting speech from children poses some special challenges. First, children's attention span depends on their age [19]; they might get distracted from a prolonged recording task. Second, they may have difficulty reading or repeating long, complex words or sentences. Taking these challenges and the requirements of the CNG game into account, we designed four types of prompts to record: 292 phonetically rich sentences, musical notes (e.g. *dó*), isolated cardinals (e.g. *44*), and sequences of cardinals (e.g. *28, 29, 30, 31*). The phonetically rich sentences originated from the CETEMPúblico corpus of Portuguese newspaper language [20]. They were short (~4 words/sentence) and did not include any difficult words. In the case of 3-6-year-olds, the cardinals ranged from 0 to 30, and the sequences of cardinals consisted of 2-3 numbers. In case of 7-10-year-olds, the cardinals ranged from 0 to 999, and the sequences of cardinals consisted of 4 numbers.

The younger children produced a set of 30 prompts selected across the different types of prompts in a balanced way. This resulted in a bit more than one minute of speech per speaker. The older children read out a set of 50 prompts resulting in about 3 minutes of speech per speaker.

The differences in the contents and targeted number of prompts between the two age groups were designed based on our experiences from pilot recording sessions with children of different ages. They take into account the differences in the attention span and linguistic capabilities between the two age groups.

3.2. Data collection

We used the *Your Speech* online speech data collection platform [21] for collecting the speech data and some biographical information (age group, gender and region of origin) about the speakers. The platform was operated by recording supervisors trained for managing the recording sessions. The platform's web interface presented the speakers with each of the prompts to record; the recording of an utterance started when the recording supervisor clicked the *Record* button and ended when (s)he clicked the *Stop* button, or when no more speech was detected by the system. The recorded utterance was then uploaded to the web backend of the system and automatically checked for the presence of speech and clipping; if speech was indeed detected and if the utterance did not contain any samples of clipping, the recording supervisor could proceed to the next prompt using the *Next Phrase* button or, if unhappy with the utterance, have the speaker rerecord the utterance using the *Rerecord* button. If the automatic quality control was not passed, the speaker was requested to rerecord the utterance.

The recording sessions took place in a quiet room. In the case of 3-6-year-olds, as well as the 7-10-year-olds that had problems reading the prompts, the recording supervisors read the prompts out first and the children then repeated them. The speech data were recorded using a noise-cancelling Life Chat LX 3000 USB headset and digitised at 16 bits and 22 kHz.

3.3. Transcriptions, annotations and quality control

Using an in-house transcription tool, a (single) native speaker trained for the task transcribed the corpus orthographically. In addition, she annotated the corpus using the tags listed in Table 1. The annotation scheme was designed to be compatible with the requirements of our in-house AM training tool.

The transcriber discarded sessions that contained recordings with consistently poor audio quality, or speech from non-native speakers or speakers with consistent problems repeating or reading the prompts out. After the transcription and annotation work, we identified typographical errors in the transcriptions by checking them against a large Portuguese lexicon. In addition, we used forced alignment to identify potentially problematic utterances; utterances that cannot be aligned are more likely to have problems in the quality of the speech, the transcriptions and/or the audio. We cast these utterances aside. As the proportion of utterances annotated with the <NPS/> tag was very low, we did not include them in the final version of the corpus, either. We have not carried

out formal, systematic assessment of the reliability of the transcriptions and annotations included in the corpus.

3.4. Overview of the corpus and the datasets for ASR experiments

We collected a total of 21 hours of speech from 510 children – 30% of them aged 3-6, and 70% aged 7-10. This makes the CNG Corpus the largest currently available corpus of pt-PT children's speech, also covering some ages that other resources of pt-PT children's speech [17, 18] do not cover. The imbalance between the amount of speech collected from 3-6-year-olds and 7-10-year-olds is due to the difficulty of recording speech from very young speakers, as well as the limited amount of speech collected from each of them. 56% of the speakers in the corpus are girls and 44% boys. The vast majority (84%) are from the Leiria area.

For the purpose of ASR experiments, the corpus was randomly divided into three speaker-independent datasets that respect the proportions of speaker age and gender in the full corpus: a training set used for training AMs (85% of the data), a development test set for optimisation purposes (5% of the data), and an evaluation test set for the final testing of the AMs (10% of the data). Due to the limited number and type of prompts recorded, it was not possible to rule out the same prompts appearing in the three different datasets; with hindsight, this could have been avoided by using a corpus design that splits the speakers and prompts between the three datasets before the recordings. The main statistics of the corpus and the three datasets are presented in Table 2.

Table 1. *The tags used for annotating the recordings.*

Tag	Meaning
<FILL/>	Filled pauses (e.g. "umm", "er", "ah")
<NON/>	Non-human noises (e.g. mouse clicks, music)
<SPN/>	Human noises (e.g. coughs, audible breath)
<UNKNOWN/>	False starts; mispronounced, unintelligible or truncated words; words with considerable background noise
<NPS/>	Speech from non-primary speakers

Table 2. *The main statistics of the speech material.*

	Train	Devel.	Eval.	Total
#Speakers	432	26	52	510
#Word types	605	482	521	614
Ages 3-6	*557*	*218*	*319*	*560*
Ages 7-10	*585*	*458*	*494*	*591*
#Word tokens	102,537	6229	12,029	121,046
Ages 3-6	*9553*	*676*	*1148*	*11,424*
Ages 7-10	*92,984*	*5553*	*10,881*	*109,622*
hh:mm:ss	17:42:22	01:06:26	02:05:34	20:54:22
Ages 3-6	*02:30:24*	*00:10:22*	*00:18:31*	*02:59:17*
Ages 7-10	*15:11:58*	*00:56:04*	*01:47:03*	*17:55:05*

4. ASR experiments

The ASR functionality of the CNG game is implemented using the Microsoft Speech Platform Runtime (Version 11) [22], which contains a Hidden Markov Model (HMM) -based speech recogniser, and a language pack, which incorporates the language-specific components necessary for ASR: grammars, a pronunciation lexicon, and AMs (cf. [23]). The goals of the ASR-related work in the project are to create a pt-PT language pack that is specifically adapted to the CNG game, and to obtain the best possible recognition performance using existing techniques and tools compatible with the requirements of the Microsoft Speech Platform Runtime. The following subsections provide information about the grammars that we have authored for the CNG game, the pronunciation lexicon that we are using, as well as our first experiments with different acoustic modelling approaches.

4.1. Grammars and pronunciation lexicon

We have authored several grammars for language modelling purposes: a list grammar for the musical notes and structure grammars for the isolated cardinals and the cardinal sequences. The grammar for the isolated cardinals allows cardinals from 0 to 999, whereas the grammar for the cardinal sequences allows sequences of 2-4 cardinals ranging from 0 to 999. In the experimentation phase, the phonetically rich sentences are simply recognised using a list grammar consisting of the 292 prompts (see Section 3.1.2); the CNG game itself will not include this type of speech input. Our pronunciation lexicon contains an average of 1.04 pronunciations for the words in the task, represented using a set of 38 phone labels.

4.2. Feature extraction

We carried out feature extraction of the children's speech data at a frame rate of 10 ms using a 25-ms Hamming window and a pre-emphasis factor of 0.98. We calculated 12 Mel Frequency Cepstral Coefficients (MFCCs) and log-energy with corresponding first, second and third order time derivatives, and reduced the total number of features to 36 using Heteroscedastic Linear Discriminant Analysis (HLDA).

4.3. Acoustic modelling

The following subsections describe our baseline recogniser, as well as the different recognisers built so far to test different acoustic modelling approaches. We omit detailed information about the baseline recogniser and our training techniques as commercially sensitive information.

4.3.1. Baseline recogniser

For establishing a baseline, we used the acoustic models from the pt-PT language pack [23]. They comprise a mix of gender-dependent (GD) whole-word models and cross-word triphones trained using several hundred hours of read and spontaneous speech collected from adult speakers. The baseline recogniser also includes a silence model, a hesitation model for modelling filled pauses, and a noise model for modelling human and non-human noises. In addition to recognising children's speech using the female and male AMs of the baseline recogniser in parallel (BL), we also recognised it using the female AMs (BL_F) and the male AMs (BL_M) separately.

4.3.2. Experimental recognisers

We trained and tested several different kinds of AMs to investigate the effect of different variables on speech recognition performance: the type of training technique, and the gender and age group of the children. The two main types of AMs were 1) the BL AMs retrained with the children's speech in the training set and 2) cross-word triphones trained using the children's speech only. The experimental AMs included GD AMs, gender-independent (GI)

AMs, and age group -dependent AMs. Due to the limited amount of training data from 3-6-year-olds, we were only able to train age group -dependent AMs for 7-10-year-olds with the types of training techniques used in our experiments so far.

We retrained the BL AMs using several different set-ups:

- BL_F and BL_M retrained with the training data from girls and boys, respectively (BL_{GD})
- BL_F retrained with the training data from both girls and boys (BL_{GI})
- BL_F and BL_M retrained with the training data from 7-10-year-old girls and boys, respectively ($BL_{GD7\text{-}10}$)
- BL_F retrained with the training data from 7-10-year-old girls and boys ($BL_{GI7\text{-}10}$)

The hesitation and noise models of the baseline recogniser were retrained utilising the <FILL/>, <NON/> and <SPN/> tags available in the CNG Corpus. The idea of only retraining the adult female AMs (BL_F) with children's speech stems from the fact that the acoustic characteristics of children's speech are more similar to adult female speech than to adult male speech [7-9, 24]. In fact, in the ASR experiments reported in [7], the WER obtained for children's speech using adult male AMs is more than twice higher than the WER achieved using adult female AMs. Furthermore, [7] reports better recognition performance with adult female models adapted with children's speech using maximum likelihood linear regression (MLLR) adaptation than with adult male and GI models adapted the same way.

We used a standard procedure with decision tree state tying (see e.g. [25]) to train several different kinds of cross-word triphone recognisers using children's speech only:

- GD triphones trained with the full training set (CNG_{GD})
- GI triphones trained with the full training set (CNG_{GI})
- GD triphones trained with the training data from 7-10-year-olds ($CNG_{GD7\text{-}10}$)
- GI triphones trained with the training data from 7-10-year-olds ($CNG_{GI7\text{-}10}$)

Similarly to the baseline recogniser and its derivatives, the cross-word triphone recognisers also included a silence model, a hesitation model and a noise model – the last two trained utilising the <FILL/>, <NON/> and <SPN/> tags available in the CNG Corpus. To find the optimal number of Gaussian mixtures per state, we trained and tested triphone recognisers with 4-16 Gaussian mixtures per state.

Table 3. *WERs (%) with a 95% confidence interval for all, for 3-6-year-old, and for 7-10-year-old speakers in the evaluation test set.*

	All Eval.	Ages 3-6	Ages 7-10
BL	18.1 ± 0.7	49.2 ± 3.0	14.9 ± 0.7
BL_F	18.1 ± 0.7	49.2 ± 3.0	14.9 ± 0.7
BL_M	32.8 ± 0.9	69.9 ± 2.7	28.9 ± 0.9
BL_{GD}	11.9 ± 0.6	29.4 ± 2.7	10.1 ± 0.6
BL_{GI}	11.5 ± 0.6	27.8 ± 2.6	9.8 ± 0.6
$BL_{GD7\text{-}10}$	13.1 ± 0.6	37.8 ± 2.9	10.5 ± 0.6
$BL_{GI7\text{-}10}$	13.0 ± 0.6	36.8 ± 2.8	10.5 ± 0.6
CNG_{GD}	10.5 ± 0.6	30.9 ± 2.7	8.3 ± 0.5
CNG_{GI}	**10.0 ± 0.5**	**27.1 ± 2.6**	**8.2 ± 0.5**
$CNG_{GD7\text{-}10}$	12.1 ± 0.6	36.5 ± 2.8	9.4 ± 0.6
$CNG_{GI7\text{-}10}$	11.5 ± 0.6	35.0 ± 2.8	9.1 ± 0.6

Table 4. *The WERs (%) of the best-performing recogniser (CNG_{GI}) per prompt type.*

	All Eval.	Ages 3-6	Ages 7-10
Phonetically rich	10.4	25.6	6.6
Musical notes	4.2	13.3	2.2
Isolated cardinals	6.3	27.4	3.9
Sequences of cardinals	10.6	33.3	9.7
Overall (excl. phon. rich)	9.8	29.3	8.7

4.4. Speech recognition results

Table 3 reports the WERs for the most relevant/best-performing recognisers: the baseline recognisers (BL, BL_F and BL_M), the baseline recognisers retrained with children's speech (BL_{GD}, BL_{GI}, $BL_{GD7\text{-}10}$ and $BL_{GI7\text{-}10}$), 14-Gaussian triphone recognisers trained using the training data from both 3-6-year-olds and 7-10-year-olds (CNG_{GD} and CNG_{GI}), and 12-Gaussian triphone recognisers trained using the training data from 7-10-year-olds only ($CNG_{GD7\text{-}10}$ and $CNG_{GI7\text{-}10}$).

All models that had specifically been adapted for children's speech significantly outperformed the baseline recogniser comprising GD AMs trained using adult speech (BL). The error rates obtained with the adult female AMs (BL_F) were identical to those obtained with the combination of the adult female and adult male AMs (BL). In other words, the adult female AMs were always chosen to recognise children's speech. This further illustrates that children's speech is more similar to adult female than to adult male speech. Recognition performance with adult male AMs (BL_M) was significantly worse. Similar to other studies (e.g. [3-5, 7]), the WERs were considerably higher in the case of the younger children. The best-performing recogniser for both age groups was the GI cross-word triphone recogniser trained using all children's speech in the training set (CNG_{GI}). Its performance did not significantly differ from that of the corresponding GD recogniser (CNG_{GD}), however. Overall, its performance was 45% (relative) better than that of the GD baseline recogniser (BL). The improvement was 45% also when calculated separately for both 3-6-year-olds and 7-10-year-olds.

In the case of 7-10-year-olds, training cross-word triphones using children's speech led to significantly better recognition performance than retraining the otherwise similar baseline AMs with children's speech (CNG_{GD} vs. BL_{GD}, CNG_{GI} vs. BL_{GI} etc.). This was the case also when retraining the adult female AMs with children's speech (CNG_{GI} vs. BL_{GI}). However, in the case of 3-6-year-olds, there were no significant differences between the two types of training techniques. For instance, in the case of 7-10-year-olds, the performance of the best-performing cross-word triphone recogniser (CNG_{GI}) was 16% (relative) better than that of the best-performing recogniser comprising adult female AMs that had been updated with speech collected from both girls and boys (BL_{GI}). The improvement was only 3% (relative; not significant) in the case of 3-6-year-olds. These findings might be related to the fact that we had much less training data from 3-6-year-olds than from 7-10-year-olds, and suggest that training children's speech models from scratch is a more promising approach than retraining adult speech models using children's speech especially when a sufficient amount of training data is available from the targeted age group.

The WERs of the experimental recognisers also illustrate that GD AMs do not lead to improved recognition performance in the case of 3-10-year-old children; the performance of otherwise similar GD and GI recognisers (BL_{GD} vs. BL_{GI}, $BL_{GD7\text{-}10}$ vs. $BL_{GI7\text{-}10}$

etc.) did not differ from each other significantly. Although the effect was not significant, the GI recognisers did seem to have a tendency for better recognition performance than the GD recognisers. This finding could be expected based on the fact that the differences in the fundamental and formant frequencies of girls' and boys' speech only become more pronounced at around 11 years of age [7, 9].

Recognition performance deteriorated in the case of 7-10-year-olds when we only used training data from their own age group to retrain or train the AMs, although the effect was only significant in the case of the cross-word triphone recognisers (CNG_{GD} vs. $CNG_{GD7\text{-}10}$ and CNG_{GI} vs. $CNG_{GI7\text{-}10}$). Even though training AMs with data from the targeted age or age group might be expected to lead to improved recognition performance [1, 2], this result supports the view that a broad diversity of the training data aids recognition performance [4]. Unsurprisingly, the WERs of 3-6-year-olds increased significantly with AMs that had been retrained or trained from scratch using the training data from 7-10-year-olds only (BL_{GD} vs. $BL_{GD7\text{-}10}$, BL_{GI} vs. $BL_{GI7\text{-}10}$ etc.).

Table 4 lists the WERs of the best-performing children's speech recogniser (CNG_{GI}) for each of the recorded prompt types. It also includes the overall WERs without phonetically rich sentences, which represent a prompt type that is not relevant for the CNG game. It is clear that the recognition performance of 3-6-year-olds leaves much to be desired. While the recognition performance of the different types of prompts also leave space for improvement in the case of 7-10-year-olds, it is probably already acceptable for the CNG game – in particular in the case of musical notes and isolated cardinals.

4.5. Discussion and future work

For now, the best-performing AMs (CNG_{GI}) have been delivered for use in the first versions of the CNG game, together with the grammars and pronunciation lexicon discussed in Section 4.1. However, we will continue to explore ways to optimise recognition performance.

We are particularly interested in improving on the poor recognition performance of 3-6-year-olds. There are probably at least three reasons for this poor performance. First, as already mentioned earlier, recognition performance correlates with children's age; regardless of the optimisation methods that we will use, the WERs on older children's speech are likely to remain lower than those on younger children's speech. Second, we had much less training data from 3-6-year-olds than from 7-10-year-olds; the AMs we trained using speech from speakers belonging to both age groups were effectively optimised for 7-10-year-olds. Third, many of the 3-6-year-olds recorded for the corpus had difficulty repeating the prompts correctly, especially in the case of prompts containing several words to memorise, and this resulted in a lot of disfluencies and hesitations in the data. In [10], recognition performance did not suffer significantly from disfluencies and hesitations in children's speech. However, we must investigate if this is also the case with our speech data, which also contains speech from speakers younger than those tested by [10]. In addition to the difficulty repeating the prompts correctly, the younger children were very challenging to record speech from because they often reacted to the recording situation with shyness (see also [17]). For all of these reasons, it might be interesting to collect more speech by recording children's verbal interaction with the CNG game itself, and to use those data to adapt the current children's speech AMs. Firstly, the recording situation would be much less intimidating. Secondly, the produced speech would probably be more suitable for training AMs for the CNG game than the speech collected during the data collection campaign: in the case of children, vowel durations are known to be significantly higher and speaking rate lower for read speech than for spontaneous speech, the effect being more pronounced in younger children [1]. Another benefit of collecting more speech is that we could offset the current bias towards regional accents from the Leiria area where most of the speech data was collected (see Section 3.4).

Before embarking on a speech data collection using a preliminary version of the CNG game, however, we will experiment with existing optimisation methods to try to improve on the recognition performance of both 3-6-year-olds and 7-10-year-olds. Some of the factors making children's speech recognition particularly challenging are the higher frequencies of their fundamental and formant frequencies and the high level of variability in these frequencies across children of different ages. Vocal Tract Length Normalisation (VTLN) has been shown to lead to improved recognition performance on children's speech both in the case of AMs trained using speech from another age group [6, 7, 26] and in the case of AMs trained using speech from the same age group [6, 7, 26]. We are currently looking into applying VTLN at both the training and the recognition stages of our experiments (cf. [26]).

Young children may have problems accurately producing particular speech sounds or clusters of speech sounds; they might, for instance, systematically substitute one consonant cluster for another [2]. This is likely to have a negative effect on speech recognition performance [2]. When there is not enough training data for acoustic models to learn such pronunciation patterns, a pronunciation lexicon customised for the pronunciation patterns of the targeted age group might improve recognition performance [1]. For instance, [27] obtained significant decreases in WER in the case of preschool children by studying how they pronounce words with respect to their canonical pronunciations, and by deriving pronunciation rules to add relevant pronunciation variants into their pronunciation lexicon. We are currently studying the pronunciation patterns of the 3-6-year-olds in the CNG corpus, with the goal of incorporating this information into our pronunciation lexicon.

In the case of adults, using restricted grammars may be a good strategy for modelling the kind of structured input (e.g. cardinal numbers) also expected by the CNG game. However, children might not restrict their responses to the utterances vital to the task only; they might, for instance, use utterances expressing excitement or disappointment, or try to interact with the characters on the screen in a way that is not related to the task at hand [10, 28]. The failure to model this kind of extraneous speech is likely to lead to deteriorated recognition performance. Therefore, we intend to analyse children's interaction with a preliminary version of the CNG game to determine the best language modelling approach for the final application.

5. Conclusions

In this paper, we presented the design, collection, transcription and annotation of the largest currently available European Portuguese children's speech corpus, which contains 21 hours of prompted speech recorded from 510 children aged 3-10. The corpus comes with manual orthographic transcriptions and annotations indicating filled pauses, noises and damaged words (e.g. mispronunciations). It was specifically designed for training and testing acoustic models for an educational game that teaches 3-10-year-old children basic mathematical and musical skills. However, it could also prove useful for developing other speech-driven

applications for children and for use in children's speech research. The corpus is available at request for R&D activities. Please contact Miguel Sales Dias (Miguel.Dias@microsoft.com) for further information.

In addition, we presented our first ASR experiments, aimed at finding the best acoustic modelling approach for the aforementioned educational game. Our speech recognition results suggest that, when a sufficient amount of training data is available from the targeted age group, training children's speech models from scratch is a more promising approach than retraining adult speech models using children's speech. This finding also holds for adult female models retrained using children's speech. Our recognition results also show that gender-dependent models do not lead to increased recognition performance in the case of 3-10-year-old children. As compared with a baseline recogniser comprising gender-dependent adult speech models, the best-performing children's speech models that we have trained so far – gender-independent cross-word triphones trained with 17.5 hours of speech from 3-10-year-old children – result in a 45-percent (relative) decrease in word error rate in a task expecting isolated cardinal numbers, sequences of cardinal numbers or musical notes as speech input.

6. Acknowledgements

The QREN 7943 CNG – Contents for Next Generation Networks project is co-funded by Microsoft, the Portuguese Government, and the European Structural Funds for Portugal (FEDER), through COMPETE and QREN. The authors are indebted to the children, recording supervisors, nurseries and schools that took part in the data collection.

7. References

[1] Gerosa, M., Giuliani, D., Narayanan, S. and Potamianos, A., "A Review of ASR Technologies for Children's Speech", in *Proc. WOCCI*, Cambridge, MA, USA, 2009.

[2] Russell, M. and D'Arcy, S., "Challenges for Computer Recognition of Children's Speech", in *Proc. SLaTE*, Farmington, PA, USA, 2007.

[3] Potamianos, A. and Narayanan, S., "Robust Recognition of Children's Speech," *IEEE Speech Audio Process.*, 11(6):603-615, 2003.

[4] Wilpon, J. G. and Jacobsen, C. N., "A Study of Speech Recognition for Children and the Elderly", in *Proc. ICASSP*, Atlanta, GA, USA, 1996.

[5] Elenius, D. and Blomberg, M., "Adaptation and Normalization Experiments in Speech Recognition for 4 to 8 Year Old Children", in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[6] Gerosa, M., Giuliani, D. and Brugnara, F., "Speaker Adaptive Acoustic Modeling with Mixture of Adult and Children's Speech", in *Proc. Interspeech*, Lisbon, Portugal, 2005.

[7] Gerosa, M., Giuliani, D. and Brugnara, F., "Acoustic Variability and Automatic Recognition of Children's Speech", *Speech Commun.*, 49(10-11):847-860, 2007.

[8] Huber, J.E., Stathopoulos, E. T., Curione, G. M., Ash, T. A. and Johnson, K., "Formants of Children, Women and Men: The Effects of Vocal Intensity Variation", *J. Acoust. Soc. Am.*, 106(3):1532-1542, 1999.

[9] Lee, S., Potamianos, A. and Narayanan, S., "Acoustics of Children's Speech: Developmental Changes of Temporal and Spectral Parameters", *J. Acoust. Soc. Am.*, 10:1455-1468, 1999.

[10] Narayanan, S. and Potamianos, A., "Creating Conversational Interfaces for Children", *IEEE Speech Audio Process.*, 10(2):65-78, 2002.

[11] Eguchi, S. and Hirsh, I. J., "Development of Speech Sounds in Children", *Acta Otolaryngol. Suppl.*, 257:1-51, 1969.

[12] Beck, J., Jia, P. and Mostow, J., "Automatically Assessing Oral Reading Fluency in a Computer Tutor that Listens", *TICL*, 1:61-81, 2004.

[13] Hagen, A., Pellom, B., Vuuren, S. V. and Cole, R., "Advances in Children's Speech Recognition within an Interactive Literacy Tutor", in *Proc. HLT/NAACL*, Boston, MA, USA, 2004.

[14] Russell, M. J., Series, R. W., Wallace, J. L., Brown, C. and Skilling, A., "The STAR System: An Interactive Pronunciation Tutor for Young Children", *Comput. Speech Lang.*, 14(2):161-175, 2000.

[15] Soares, L. P., Pires, F., Varela, R., Bastos, R., Carvalho, N., Gaspar, F. and Sales Dias, M., "Designing a Highly Immersive Interactive Environment: The Virtual Mine", *Comput. Graph. Forum*, 29(6):1756-1769, 2010.

[16] Kinect for Windows. Online: http://www.microsoft.com/en-us/kinectforwindows/, accessed on 4 Apr 2013.

[17] Lopes, C., Veiga, A. and Perdigão, F., "A European Portuguese Children Speech Database for Computer Aided Speech Therapy", in *Proc. PROPOR*, Coimbra, Portugal, 2012.

[18] The Portuguese Speecon Database. Online: http://catalog.elra.info/product_info.php?products_id=798, accessed 4 Apr 2013.

[19] Unger, H. G., *Encyclopedia of American Education*, Third Edition, Facts on File Inc., New York, USA, 2007.

[20] CETEMPúblico. Online: http://www.linguateca.pt/cetempublico/, accessed 4 Apr 2013.

[21] Freitas, J., Calado, A., Braga, D., Silva, P. and Sales Dias, M., "Crowd-Sourcing Platform for Large-Scale Speech Data Collection", in *Proc. FALA*, Vigo, Spain, 2010.

[22] Microsoft Speech Platform Runtime (Version 11). Online: http://www.microsoft.com/en-us/download/details.aspx?id=27225, accessed 4 Apr 2013.

[23] Microsoft Speech Platform Runtime Languages. Online: http://www.microsoft.com/en-us/download/details.aspx?id=27224, accessed 4 Apr 2013.

[24] Caldas de Oliveira, L., "eCIRCUS: Children Voices Against Bullying in Schools", keynote at *1st Microsoft Workshop on Speech Technology - Building Bridges between Industry and Academia,* Porto Salvo, Portugal, 2007. Online: http://www.microsoft.com/pt-pt/mldc/news/mldcworkshop.aspx, accessed 8 Apr 2013.

[25] Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. and Woodland, P., *The HTK Book (for HTK Version 3.2.1)*, Cambridge University, Cambridge, UK, 2002.

[26] Giuliani, D. and Gerosa, M., "Investigating Recognition of Children's Speech", in *Proc. ICASSP*, Hong Kong, 2003.

[27] Cincarek, T., Shindo, I., Toda, T., Saruwatari, H. and Shikano, K., "Development of Preschool Children Subsystem for ASR and Q&A in a Real-Environment Speech-Oriented Guidance Task", in *Proc. Interspeech*, Antwerp, Belgium, 2007.

[28] Strommen, E. F. and Frome F. S., "Talking Back to Big Bird: Preschool Users and a Simple Speech Recognition System", *Educ. Technol. Res. Dev.*, 41(1):5-16, 1993.

A Cloud-based Personalized Recursive Dialogue Game System for Computer-Assisted Language Learning

Pei-hao Su [#1], Tien-han Yu [#], Ya-Yunn Su [*], and Lin-shan Lee [*2]

[#]Graduate Institute of Communication Engineering, National Taiwan University
[*]Graduate Institute of Computer Science and Information Engineering, National Taiwan University
[1] r00942135@ntu.edu.tw, [2] lslee@gate.sinica.edu.tw

Abstract

In this paper we present the design and experimental results of a cloud-based personalized recursive dialogue game system for computer-assisted language learning. A number of tree-structured sub-dialogues are used sequentially and recursively as the script for the game. The dialogue policy at each dialogue turn is optimized to offer the most appropriate training sentence for every individual learner considering the learning status, such that the learner can have the scores for all selected pronunciation units exceeding a pre-defined threshold in minimum number of turns. The policy is modeled as a Markov Decision Process (MDP) with high-dimensional continuous state space and trained with a huge number of simulated learners generated from a corpus of real learner data. A real cloud-based system is implemented and the experimental results demonstrate promising outcomes.

Index Terms: Computer-Assisted Language Learning, Dialogue Game, Continuous State Markov Decision Process, Fitted Value Iteration, Gaussian Mixture Model

1. Introduction

Computer-assisted language learning (CALL) systems offer various advantages for language learning such as immersive environment and corrective feedback during the learning process. Thanks to the explosive development of technology in recent years, high performance computers, tablets and even smartphones are common nowadays. It is convenient and useful to embed systems into these devices. Also, The use of speech processing technologies has been considered a good approach to provide effective assistance [1, 2, 3, 4, 5].

"Rosetta Stone" [6] and "byki" [7] are useful applications that provide multifaceted functions including pronunciation evaluation and corrective feedback. However, sentence-level practice lacks opportunities for language interaction and an immersive language learning environment [8, 9]. Spoken dialogue systems [10, 11, 12, 13, 14] are regarded as excellent solutions to provide language interaction scenarios. Recently we presented a dialogue game framework [15] in which proper training sentences at each dialogue turn are selected for each individual learner during the interaction based on the learning status. The dialogue framework was modeled as a Markov decision process (MDP) trained with reinforcement learning [16, 17], and the learning status was based on NTU Chinese [18], a Mandarin Chinese pronunciation evaluation tool. One limitation of this framework is that the discrete state representation was in short of full observation of the learner's learning status. Furthermore, its training assumed a fixed number of dialogue turns; this is impractical and inflexible.

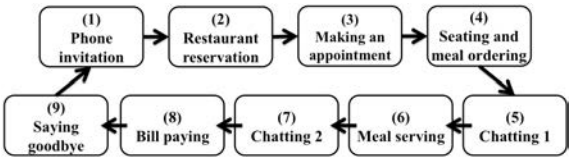


Figure 1: The script of the recursive dialogue game in the restaurant scenario: starting from (1) Phone invitation and (2) Restaurant reservation, after (9) Saying goodbye returning to (1) for next meal.

In a companion paper, we propose a new dialogue game framework for language learning [19]. A number of sub-dialogue trees are used sequentially and recursively. The leaves of the last tree are linked to the root of the first tree, making the dialogue paths infinitely long. At any dialogue turn there are a number of training sentences that can be selected. The goal of the policy is to select the training sentence at each dialogue turn based on the learning status of the learner, such that the learner's scores for all selected pronunciation units exceed a pre-defined threshold in a minimum number of turns. The framework is again modeled as an MDP, but here the MDP is realized in a high-dimensional continuous state space for a more precise representation of the learning status. This framework has been successfully implemented under a cloud-based environment and displayed on the iOS platform. This paper presents the complete design and the preliminary experimental results of this cloud-based dialogue game system.

2. Proposed recursive dialogue game framework

2.1. Recursive dialogue game concept and framework

The progress of the dialogue game is based on the script of a series of tree-structured sub-dialogues cascaded into a loop, with the last sub-dialogue linked to the first. In preliminary experiments, the whole dialogue set contains conversations between roles A and B — one the computer and the other the learner. After each utterance produced by one speaker, there are a number of choices for the other speaker's next sentence. Figure 1 shows the recursive structure of the script in the restaurant scenario. In all, nine sub-dialogues with 176 turns are used in the experiments. The whole dialogue starts with the phone invitation scenario, followed by restaurant reservation and so on, all the way to the last sub-dialogue of saying goodbye. After the last tree, the progress restarts at the first phone invitation sub-dialogue again for the next meal. This makes the dialogue continue infinitely. Figure 2 is a segment of the sub-dialogue

Figure 2: A segment of the dialogue script for the dialogue game example in a restaurant conversation scenario.

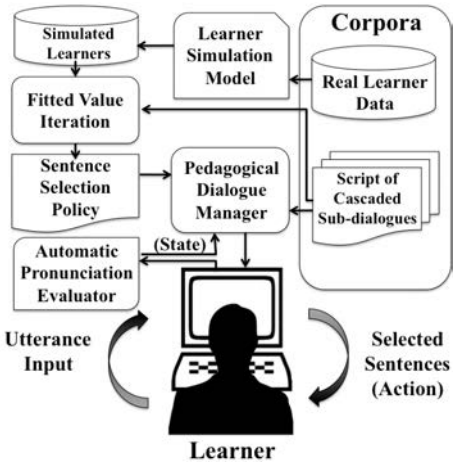


Figure 3: System block diagram of the proposed recursive dialogue game framework.

"Seating and meal ordering", where A is the waiter and B the customer.

Since both the computer and the learner have multiple sentence choices in each turn, every choice influences the future path significantly; this results in a very different distribution of pronunciation unit counts for the learners to practice. The dialogue policy here is to select the most appropriate sentence for the learner to practice at each turn considering the learning status, such that more opportunities are given to practice poorly produced pronunciation units along the dialogue path. In this way the learner can achieve the goal of having the scores of all pronunciation units exceed a pre-defined threshold in a minimum number of turns. Also, they receive pronunciation performance feedback immediately after each utterance pronounced.

The above recursive dialogue game is modeled by an MDP with the desired optimal policy trained with the Fitted Value Iteration (FVI) algorithm. A learner generation model is developed to generate simulated learners from real learner data to be used in the FVI algorithm.

The overall system block diagram of the proposed framework is shown in Figure 3. Interaction between the learner and the system involves Utterance Input from the learner and Selected Sentences from the system. The Automatic Pronunciation Evaluator scores the performance of each pronunciation unit in the utterance. These quantitative assessments are sent to the Pedagogical Dialogue Manager, which is driven by the Sentence Selection Policy for choosing the next sentence for the learner. A set of Real Learner Data is used to construct the Learner Simulation Model, which generates the Simulated Learners to train the Sentence Selection Policy based on the Script of Cascaded Sub-dialogues using the Fitted Value Iteration algorithm.

2.2. Simulated learner generation from real learner data

The real learner data used in these experiments were collected in 2008 and 2009. In total there were 278 Mandarin Chinese learners at the National Taiwan University (NTU) from 36 countries with balanced gender, each pronouncing 30 sentences selected by language teachers. NTU Chinese, a Mandarin pronunciation evaluation tool developed at NTU [18], was used as the Automatic Pronunciation Evaluator in Figure 3. It assigned scores from 0 to 100 to each pronunciation unit in every utterance of the real learner data. The scores of each utterance pronounced by a learner are used to construct a pronunciation score vector (PSV), whose dimensionality is the number of the pronunciation units considered. Every component of the PSV is the average score of the corresponding unit in the utterance; those units unseen in the utterance are viewed as missing data and solved by the expectation-maximization (EM) algorithm [20, 21]. The PSVs from all utterances produced by all real learners are used to train a Gaussian mixture model (GMM), here referred to as the Learner Simulation Model. This is shown in Figure 4.

For MDP policy training, when starting a new dialogue game, we randomly select a Gaussian mixture component as a simulated learner [22, 23, 24]. When a sentence is to be pronounced, a randomly sampled PSV from this mixture yields the scores for the units in this sentence as the simulated utterance. Since the goal of the dialogue is to provide proper sentences for each learner until their pronunciation performance for every unit reaches a pre-defined threshold, we further develop an incremental pronunciation improvement model for the simulated learners. Details about the simulated learners are in the companion paper [19].

2.3. Markov decision process

A Markov decision process (MDP) [25] is a framework that models decision making problems, represented by the 5-tuple $\{S, A, R, T, \gamma\}$: the set of all states S , the set of possible actions A, the reward function R, the Markovian state transition function T, and the discount factor γ which determines the effect of future outcomes on the current state s. When an action a is taken at state s, a reward r is received and the state is transmitted to new state s'. Solving the MDP consists in determining an infinite state transition process called a *policy* that maximizes the expected total discounted reward from state s (or value function) : $V^{\pi}(s) = E[\sum_{k=0}^{\infty} \gamma^k r_k | s_0 = s, \pi]$, where r_k is the reward gained in the k-th state transition, and the policy $\pi : S \rightarrow A$ maps each state s to an action a. The above value function can be further analyzed by the state-action (Q) value function, which is defined as the value of taking action a at state s : $Q^{\pi}(s, a) = E[\sum_{k=0}^{\infty} \gamma^k r_k | s_0 = s, a_0 = a, \pi]$. Thus, the optimal policy π^* can be expressed as $\pi^*(s) = \arg\max_{a \in A} Q(s, a)$ by a greedy selection of the state-action pair. The goal of finding the optimal policy is therefore equivalent to maximizing these Q functions.

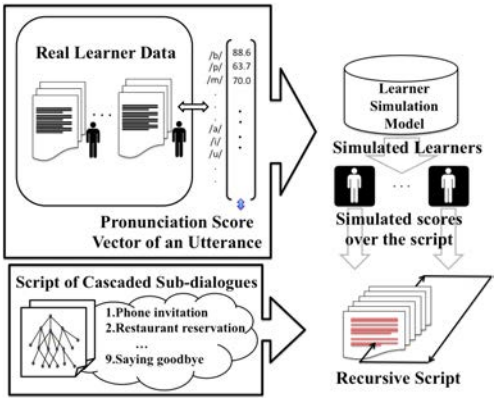


Figure 4: Learner Simulation Model for simulated learner generation.

2.4. MDP framework on dialogue game

We describe how the dialogue game is modeled using MDP.

2.4.1. Continuous state space

The state represents the learner's learning status. It consists of the scores obtained for every pronunciation unit given by the Automatic Pronunciation Evaluator in Figure 3, each a continuous value ranging from 0 to 100 and directly observable by the system. This results in the high-dimensional continuous state space $s \in [0, 100]^U$, where U is the total number of pronunciation units considered. In addition, as the system must determine which dialogue turn the learner is in, the index of dialogue turn t is also included in the state space.

2.4.2. Action set

At each state with dialogue turn t, the system's action is to select one out of a number of available sentence options for the learner to practice. The number of actions is the number of next available sentences to be chosen for the learner at the turn.

2.4.3. Reward definition

A dialogue *episode* E contains a sequence of state transitions $\{s_0, a_0, s_1, a_1, ..., s_K\}$, where s_K represents the terminal state. As mentioned, the goal here is to train a policy that can at each turn offer the learner the best selected sentence to practice considering the learning status, such that the learner's scores for all selected pronunciation units exceed a pre-defined threshold within a minimum number of turns. Hence every state transition is rewarded -1 as the penalty for an extra turn ($r_k = -1, k \leq K-1$), and r_K is the finishing reward gained when the terminal state s_K is reached, where scores of all pronunciation units reach a certain threshold. The final return R is then the sum of the obtained rewards: $R = \sum_{k=0}^{K} r_k$. In addition, a timeout count of state transitions J is used to limit episode lengths.

2.4.4. Fitted value iteration (FVI) algorithm

For the high-dimensional continuous state space, we use the function approximation method [26, 27, 28] to approximate the exact Q value function with a set of m basis functions:

$$Q(s,a) = \sum_{i=1}^{m} \theta_i \phi_i(s,a) = \underline{\theta}^T \underline{\phi}(s,a), \quad (1)$$

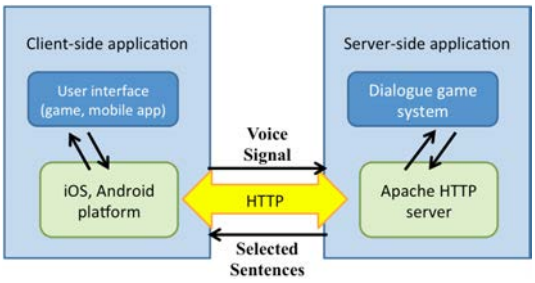


Figure 5: System architecture of the cloud-based system.

where $\underline{\theta}$ is the parameter (weight) vector corresponding to the basis function vector $\underline{\phi}(s,a)$. The goal of finding the optimal policy can then be reduced to finding the appropriate parameters $\underline{\theta}$ for a good approximation $\hat{Q}_\theta(s,a)$ of $Q(s,a)$. A *sampled* version of the Bellman backup operator $\hat{B}$ is introduced for the i-th sampled transition (s_i, a_i, r_i, s_i') as

$$\hat{B}(Q(s_i,a_i)) = r_i + \gamma \max_{a \in A} Q(s_i{}', a_i). \quad (2)$$

With a batch of transition samples $\{s_j, a_j, r_j, s_j' | j = 1, ..., N\}$, least-squares regression can be performed to find the new parameter vector $\underline{\theta}_n$ at the n-th iteration so that $\hat{Q}_{\theta_n}(s,a)$ approaches $Q(s,a)$ as precisely as possible. The parameter vector is updated as

$$\underline{\theta}_{n+1} = \arg\min_{\underline{\theta} \in \mathbb{R}^M} \sum_{j=1}^{N} \left(\hat{Q}_{\theta_n} - \hat{B}(Q(s_i,a_i))\right)^2 + \frac{\lambda}{2}\|\underline{\theta}\|^2, \quad (3)$$

where the second term is the 2-norm regularized term determined by λ to prevent over-fitting.

3. Cloud-based system design and implementation

3.1. System overview

We have implemented the Mandarin Chinese dialogue game core engine as a cloud-based system. To provide good operability, the system is exposed through REST API. Figure 5 shows our system architecture. A web server accepts HTTP requests with a URL mapping to our dialogue system service. The web server then passes the HTTP request including the parameters and translates the request into a corresponding voice signal call to the pedagogical dialogue manager. After the next sentence for practice is selected by the dialogue manager, this selected sentence is packed into a HTTP response. User-specific data, such as pronunciation scores and profile, are stored in a separate database. In this way, developers can build applications for various platforms, such as a web page or a mobile app or a flash game, using any HTTP library that can issue the REST calls.

3.2. Initial user interface

Figure 6 is the initial user interface of our dialogue game showing the fundamental functionalities. The left part shows the dialogue progress, which alternates between the waiter and the customer. The last two sections are the current sentence produced by the waiter (system) and the sentence candidates for the customer (learner) to choose, while the other sections list the past sentences spoken. The "Hide/Show" button on the upper right switch off/on the display of the past sentences. The

Figure 6: An example view of our designed system interface.

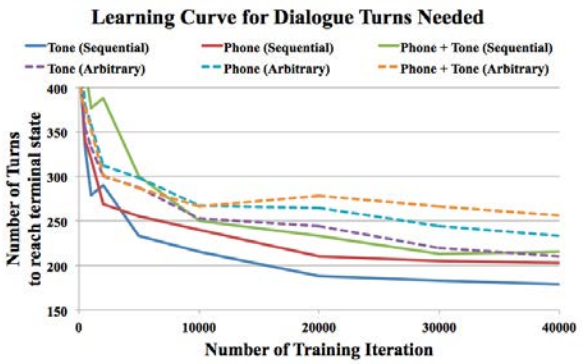


Figure 7: Number of dialogue turns needed with respect to different number of training iterations.

customer chooses to produce one sentence by clicking on the "Start Recording" button. Note that there is a "BEST CHOICE" labeled on one sentence candidate of the customer, it is the one recommended by the optimized sentence selection policy mentioned above. In addition, by clicking on the blue "play" icon we can listen to the sentence spoken by the waiter again. When clicking on the "analysis" icon on the past sentences of the customer, the system shows the evaluation result of each unit within the selected sentence, which is shown on the right part. The evaluation result indicates the pronunciation performance on the whole utterance and on each Mandarin syllable, including scores of Initial/Finals, tone, timing and emphasis. This offers detailed assessment of the learner's pronunciation.

The complete cloud-based system has been successfully implemented and operated in real time. It is also submitted to the demonstration session of SLaTe 2013 [29].

4. Experiments

4.1. Experimental Setup

Experiments were performed on the complete script of nine sub-dialogue trees for Mandarin Chinese learning as described in Section 2.1. The results below are for the computer as role A and the learner as role B. Totally 82 Mandarin pronunciation units including 58 phonetic units (Initial/Finals of Mandarin syllables) and 24 tone patterns (uni/bi-tone) were considered, and three cases were tested: learning tone patterns only, phonetic units only, and both. NTU Chinese [18] was used as the automatic pronunciation evaluator for unit scoring and immediate feedback for the learners. In the MDP setting, the terminal state s_K was defined as the situation that all pronunciation units considered were produced with scores over 75 more than eight times. The reward at the dialogue terminal state r_K was set to 300 and timeout count J was 500. Multivariate Gaussian functions of 82 dimensions served as the basis function $\phi(s, a)$ in (1) to represent the Q value function. The number of the basis functions was set 5, and these Gaussian functions were spread evenly on the state space. The system's initial policy was always to choose the first sentence among the candidate sentences. Five-fold cross-validation was used: in each training iteration, four-fifths of the real learner data were used to construct the GMM to generate simulated learners for policy training, while the rest was saved for another GMM to generate simulated learners in the testing phase. In our work, the MDP testing result was the average of 50 testing simulated learners. Also, Bayesian information criterion (BIC) [30, 31] was employed on GMM to balance the model likelihood and parameter complexity. In the experiment, simulated learners were generated to go through the nine sub-dialogue trees in either sequential and recursive order or arbitrary order until the terminal state s_K was reached.

4.2. Experimental Result

4.2.1. Number of dialogue turns needed

In Figure 7, we plot the number of turns needed to reach the terminal state as a function of the number of training iterations. Clearly the three solid curves (labeled "Sequential") for different sets of target units considered yielded promising results. The number of needed turns for learning tone patterns alone, phonetic units alone, and both converged at 179.88, 203.42, 215.58 turns respectively. Clearly as the number of target units is smaller, the needed turns is smaller. Note that the needed turns of considering phonetic units alone (203.42) is only slightly smaller than considering both phonetic units and tone patterns (215.58), while that of considering only tone patterns (179.88) is much smaller. Different sets of pronunciation units are presented in the training sentences in any case with different distributions. The above results indicate that when considering only phonetic units, the practice may cover many tone patterns as well. That is to say, considering a set of units together as target learning units at a time may result in less total number of training sentences than considering the same set of units in separated times. In addition, since there were 84 turns in all for role B in the nine consecutive sub-dialogues, the results indicated that going through all nine trees and restarting from the first sub-dialogue was necessary for the testing simulated learners here.

The dashed curves (labeled "Arbitrary") show the results of using the nine sub-dialogue trees in a different scenario, in which the learner chose to practice the sub-dialogue trees in an arbitrary order. For example, the learner could jump to sub-dialogue four after finishing sub-dialogue two (after restaurant reservation, the learner wishes to learn how to order meals first). The same three cases (tone patterns only, phonetic units only, and both) tested in this scenario converged at 210.16, 233.64, and 256.82 turns respectively as shown in Figure 7. The extra turns needed compared to the sequential order scenario shows the trade-offvbetween the user's free will to interact with the dialogue game and the dialogue turns needed to learn all target units well enough.

4.2.2. Focused learning for specific sets of pronunciation units

From section 4.2.1 we learned the effectiveness of the system policy. The system provided personalized pronunciation unit practice as efficient as possible to each individual learner. However, some language learners might already know their pronunciation status in advance and wished to focus their learning on a specific set of units using the dialogue game system. We therefore would like to test the learned policies considering different target units as discussed in section 4.2.1. In the experiments below, the simulated learner selected certain number of units randomly as the units to be focused on while ignoring the scores of all other units.

Table 1: Number of dialogue turns needed for focused learning on a specific number of pronunciation units.

Target units	**Number of units focused**	**Number of turns needed**	
(1) Tone patterns	10	Sequential	140.28
		Arbitrary	159.77
(2) Phonetic units	20	Sequential	173.53
		Arbitrary	205.09
(3) Phone + Tone	20	Sequential	179.11
		Arbitrary	209.16

Table 1 shows the dialogue turn needed for focused learning of 10 tone patterns (row(1)), 20 phonetic units (row(2)), and 20 phonetic units or tone patterns (row(3)) using the policies learned in section 4.2.1 respectively, either following the sub-dialogue trees sequentially (labeled "'Sequential') or in arbitrary order (labeled "'Arbitrary'). Each result is the average over 100 simulated learners. This shows different ways of utilizing the dialogue game developed here.

From Table 1 we can see that a significant number of turns were needed even if only 10 units are focused on, but the policy became more efficient when more units were considered. This is obviously because the training utterances automatically carried many different units for practice even if the learner wished to focus on a small number of them. Also, some low frequency units, if selected by the learner, may require more turns to be practice in the dialogue.

5. Conclusions

We presented a cloud-based recursive dialogue game with an optimized policy offering personalized learning materials for CALL. A series of recursive tree-structured sub-dialogues are used as the script for the game. The policy is to offer the proper sentence for practice at each turn considering the learning status of the learner. It was optimized by an MDP with a high-dimensional continuous state space and trained using fitted value iteration. The cloud-based system has been successfully completed and operated in real time. Experimental results of sequential and arbitrary order usage showed promising results and the effectiveness of the proposed approach.

6. References

[1] M. Eskenazi, "An overview of spoken language technology for education," in *Speech Communication*, vol. 51, 2009, pp. 832–844.

[2] C. Cucchiarini, J. van Doremalen, and H. Strik, "Practice and feedback in l2 speaking: an evaluation of the disco call system," in *Interspeech*, 2012.

[3] Y. Xu, "Language technologies in speech-enabled second language learning games: From reading to dialogue," Ph.D. dissertation, Massachusetts Institute of Technology, 2012.

[4] X. Qian, H. Meng, and F. Soong, "The use of DBN-HMMs for mispronunciation detection and diagnosis in l2 english to support computer-aided pronunciation training," in *Interspeech*, 2012.

[5] T. Zhao, A. Hoshino, M. Suzuki, N. Minematsu, and K. Hirose, "Automatic chinese pronunciation error detection using svm trained with structural features," in *Proceedings IEEE Workshop on Spoken Language Technology*, 2012.

[6] (1999) Rosetta Stone. [Online]. Available: http://www.rosettastone.com/

[7] (2013) byki. [Online]. Available: http://www.byki.com/

[8] D. Christian, *Profiles in Two-Way Immersion Education. Language in Education: Theory and Practice 89.*, 1997.

[9] W. L. Johnson, "Serious use of a serious game for language learning," in *International Journal of Artificial Intelligence in Education*, 2010.

[10] S. Young, M. Gasic, B. Thomson, and J. Williams, "Pomdp-based statistical spoken dialogue systems: a review," in *Proceedings of the IEEE*, vol. 99, 2013, pp. 1–20.

[11] A. Raux and M. Eskenazi, "Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges," in *InSTIL/ICALL Symposium 2004*, 2004.

[12] J. D. Williams, I. Arizmendi, and A. Conkie, "Demonstration of AT&T "let's go": A production-grade statistical spoken dialogue system," in *Proc. SLT*, 2010.

[13] Y. Xu and S. Seneff, "A generic framework for building dialogue games for language learning: Application in the flight domain," in *Proc. SLaTE*, 2011.

[14] S. Lee and M. Eskenazi, "Incremental sparse bayesian method for online dialog strategy learning," *Journal of Selected Topics Signal Processing*, 2012.

[15] P.-H. Su, Y.-B. Wang, T.-H. Yu, and L.-S. Lee, "A dialogue game framework with personalized training using reinforcement learning for computer-assisted language learning," in *ICASSP*, 2013.

[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. MIT Press, 1999.

[17] R. Bellman, *Dynamic programming*. Princeton University Press, 1957.

[18] (2009) NTU Chinese. [Online]. Available: http://chinese.ntu.edu.tw/

[19] P.-H. Su, Y.-B. Wang, T.-H. Wen, T.-H. Yu, and L.-S. Lee, "A recursive dialogue game framework with optimal policy offering personalized computer-assisted language learning," in *Interspeech*, 2013.

[20] R. Hogg, J. McKean, and A. Craig, *Introduction to Mathematical Statistics*. Pearson Prentice Hall, 2005.

[21] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," in *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, 1977, pp. 1–38.

[22] J. Schatzmann, K. Weilhammer, M. Stuttle, and S. Young, "A survey of statistical user simulation techniques for reinforcement-learning of dialogue management strategies," in *The Knowledge Engineering Review*, vol. 00:0, 2006, pp. 1–24.

[23] H. Ai and F. Weng, "User simulation as testing for spoken dialog systems," in *SIGdial*, 2008.

[24] J. Schatzmann, M. N. Stuttle, K. Weilhammer, and S. Young, "Effects of the user model on simulation-based learning of dialogue strategies," in *ASRU*, 2005.

[25] A. N. Burnetas and M. N. Katehakis, "Optimal adaptive policies for markov decision processes," *Mathematics of Operations Research*, 1995.

[26] L. Daubigney, M. Geist, and O. Pietquin, "Off-policy learning in large-scale pomdp-based dialogue systems," in *ICASSP*, 2012.

[27] Y. Engel, S. Mannor, and R. Meir, "Bayes meets bellman: The gaussian process approach to temporal difference leraning," in *ICML*, 2003.

[28] F. S. Melo, S. P. Meyn, and M. I. Ribeiro, "An analysis of reinforcement learning with function approximation," in *ICML*, 2008.

[29] P.-H. Su, T.-H. Yu, Y.-Y. Su, and L.-S. Lee, "NTU Chinese 2.0: A personalized recursive dialogue game for computer-assisted learning of mandarin chinese," in *Proc. SLaTE (submitted)*, 2013.

[30] W. Zucchini, "An introduction to model selection," in *Journal of Mathematical Psychology*, vol. 44, 22006, pp. 41–61.

[31] K. Hirose, S. Kawano, S. Konishi, and M. Ichikawa, "Bayesian information criterion and selection of the number of factors in factor analysis models," in *Journal of Data Science*, vol. 9, 2011, pp. 243–259.

POLLI: a handheld-based aid for non-native student presentations

Elizabeth M Davis, Oscar Saz, Maxine Eskenazi

Carnegie Mellon University, Pittsburgh, PA. USA

emd1@andrew.cmu.edu, osaz@cs.cmu.edu, max@cmu.edu

Abstract

Language learning is finally leaving the classroom for the real world. This usually takes the form of a dedicated implementation on a handheld device. In this paper we describe POLLI, an app that helps non-native students prepare scientific presentations.
Index Terms: handheld applications, pronunciation skills

1. Introduction

In the past few decades, Computer-Aided Language Learning (CALL) has emerged, freeing students from limited time in the classroom to unlimited language exposure and learning opportunities. Students can learn at any time they want, spend as much time as they want and interact with educational material in a variety of ways. While in the past one course had to cover all aspects of the target language (L2), CALL software has gradually become specialized, teaching, for example, just pronunciation, or writing. Students learn what they immediately are in need of. CALL has also become more realistic. Often, rather than consisting of a set of exercises using expressions that the student is unlikely to use ("this is a table and that is a book"), CALL systems are starting to address real student needs. While this is great progress, they still fall short at being permanently present for a learner. Students learn best when they need some knowledge and are going to apply it in their everyday lives. A CALL system that requires a laptop or desktop cannot be out in the street with the student when they want to ask for directions or pay their hotel bill. An omnipresent system makes every new event a learning event..

Within the past three years, people have come to depend on mobile devices, such as tablets, personal data assistants, and smartphones, and ubiquitous internet access, in their everyday lives. These handheld devices, with touchscreens and voice-activated software, allow users to complete important day-to-day tasks from virtually anywhere. Moving away from the mouse-and-keyboard interface, touchscreens and voice-activated software allow for a diverse set of natural interactions, more suited to an on-the-go lifestyle. Thanks to this rise in mobile computing, there has been work on mobile learning [1] [2] [3] [4], etc. Past sessions of SLaTE have included demos of hand held systems as well, usually, like [4], for pronunciation training. Such methods of learning come with a different set of assumptions than that of a classroom or desktop environment, especially due to the lack of a fixed, predetermined location. The student can be anywhere. There is always a potential for learning to take place when information has to be exchanged. And a learning situation can be defined as any situation where the student is exposed to new material that is both understandable and usable.

This new learning situation, which has been called Mobile Assisted Language Learning (MALL) by some [5] offers tremendous potential in the pursuit of augmented, on-demand learning, picking up where the desktop technologies leave off in handling specific, contextual situations. We strongly believe that the advent of MALL will dramatically change the way in which L2 is acquired, greatly augmenting the quality and fluency of non-native speech due to the immediacy of need and use. In this paper, we discuss an app that helps non-natives with their pronunciation in presentations. It aims at helping them avoid confusable contexts, or phrases where a subtle mispronunciation, as small as perhaps a single phone, could give a statement an entirely different meaning from the one intended.

2. Background

There are an increasing number of computer-based systems that can teach non-native students the correct pronunciation of an L2 (amongst the many, [6], [7], etc). While these systems give students a basic understanding of how to pronounce phones and give examples students can imitate, they only address the problem in the abstract. They do provide examples that are designed to exercise the new phonetic knowledge in many different contexts. But these contexts are not necessarily related to anything that the student would say in real life. Also, while they do teach many different contexts, most of them are places where a mispronunciation will not cause a misunderstanding. Students then go out in the real world and encounter situations where they are mostly understood, just as long as their pronunciation errors are not in minimal pair contexts situated in sentences where either member of the minimal pair could fit. In these important minimal pair contexts, where meanings can be very different when a pronunciation error occurs, either they get the wrong message across or they confuse the listener as to what they intended to say [8]. The latter case happens when the syntactic context is correct, but no semantic interpretation can be drawn from the utterance ("we take the author's point" "we make the author's point").

In this paper, as in [9], we will call sentences with both types of potential for confusion "confusable contexts". For example, the phrases "We will adopt your proposal" and "We will adapt your proposal" differ by only one phoneme. Both make sense, but have two distinctly different meanings. For speakers of languages where the phonemes /a / and /ä/ either do not exist or are indistinguishable, it would not be difficult to mistakenly say the phonemes of one sentence while intending the other, and a native English speaker would not necessarily pick up on the mistake.

[12] implemented the BICC algorithm that automatically detects confusable contexts using several measures, including minimal pairs (gleaned from CMUDICT), and bigram part of speech models (Stanford POS tagger). The phoneticised version of words in a given sentence is compared to all of the entries in CMUDICT to find all possible minimal pairs. For each minimal pair where both words in the pair are the same part of speech, a new sentence is generated with the minimal pair word substituted

in it. The new sentence is then evaluated for plausibility by the POS tagger. If it passes this test, the original and the newly-generated sentences are then passed through the POS tagger and compared. If both sentences have the same POS string, and if the immediate context of the confusable word has similar POS tri-grams, and similar word tri-grams, then the sentences are considered to be confusable. This generates some more possibilities than are actually plausible and some heuristics can be used to further prune the number of possible confusable contexts. One way to prune the contexts is to show an individual only the confusions that they are likely to make, *given their native language*. Thus, a Japanese native speaker could have issues differentiating "we used crowd computing" from "we used cloud computing" in English while this would not be a problem for a native speaker of Spanish or of French. BICC therefore has representations of confusions between L1 and L2 for many L1s.

3. POLLI: The POcket Language Learning Interface

There are many ways that the BICC algorithm described above could be used in language learning systems. For example, a pronunciation trainer could have BICC generate the sentences that students are asked to pronounce and practice, thus replacing the abstract, useless ones used at present. We decided to concentrate on one specific application, keeping in mind that our goal is to address narrow needs, not the broad abstract ones of the past. We believe that in the future students will concur, choosing their learning software according to their immediate needs. Thus POLLI uses the BICC software in the narrowly-focused application of scientific presentations.

POLLI is also mobile, usable anywhere at any time it was needed. Thus for POLLI, since we chose to help non-native students who were preparing to give a talk at a scientific gathering, they could use it to prepare while on a plane, or sitting at a session of the conference. The goal of the system is to show speakers the confusable contexts that they might produce in their talks. For now, the student is left to decide what to do about these contexts – whether to get pronunciation training on how to say them well, or to avoid using them in the talk.

POLLI exists as an application for the Android platform using the BICC algorithm on a piece of English text. Students can either type the text in themselves (Figure 2) or load in a plain text document (Figure 1). They also can specify their native language in the app by selecting it from the list of supported languages, as shown on Figure 3.

The ability to choose to either type in what the person intends to say or to upload the scientific paper they are presenting gives the students more flexibility, depending on whether they create their presentations in a way that is very close to the text or if they author them on a more global semantic basis.

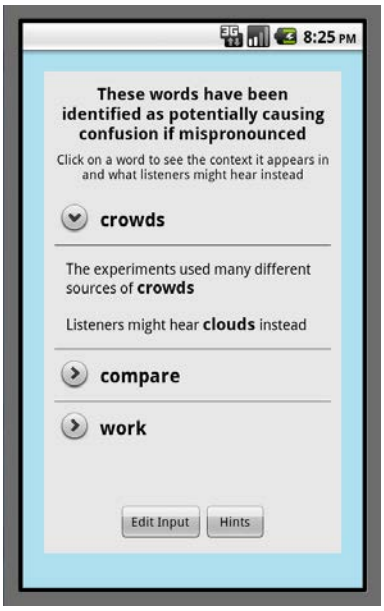
Figure 1 Uploading a text file in POLLI

Figure 2. Text input to POLLI

Figure 3. Selecting a native language in POLLI

POLLI analyzes the text and displays the output as an expandable list of the original words that, if changed by one phoneme, can change the interpreted meaning of the sentences they appear in (or cause considerable confusion as to what the speaker intended). By selecting an element in the list, the speaker can see the sentences that the words appeared in, as well as the word(s) that listeners might infer from the context if the word was mispronounced. (Figure 4).

Figure 4. POLLI presents the text analysis results



4. Assessing user acceptance of POLLI

First POLLI was tested for correct functionality: that documents could be uploaded and analyzed, that screen-input text could be analyzed, that correct minimal pairs were found and displayed and that plausible contexts were also found and displayed. With a fully working app, we proceeded to assess its usefulness for foreign undergraduate and graduate students.

User tests were carried out with a group of 24 non-native English speakers, students recruited from within Carnegie Mellon University in the Language Technologies Institute and the Computer Science Department. They had many different native languages: Mandarin Chinese, Spanish, Japanese, Russian, Thai, Tamil, Marathi, Portuguese, Greek, and French. Before coming to try POLLI, each subject was asked to submit the text of a familiar piece of writing that they felt comfortable talking about. Upon arrival, subjects recorded a practice presentation summarizing the text they had submitted. After the practice session, they tried POLLI on an Android Nexus 7, opening the app, loading their text, and submitting it for analysis. Subjects were then given ten minutes to review POLLI's output and prepare another summary presentation. That new presentation was then recorded. During the ten minutes with POLLI, they were free to do whatever they wanted with its output, for example, making notes on their text or asking questions about how to properly pronounce the words POLLI had flagged for them. After they recorded their second presentation, they were given an exit survey. They were asked to say how strongly they agreed or disagreed with a series of statements about the application, their answers ranging from 1 to 5 (Likert scale), where 1 meant strongly disagree and 5 meant strongly agree. The following are a subset of the questions that were asked.

Statements on effectiveness of the feedback were:

I learned something from this app

The information I got from the app was useful

Statements on the usability of the app were:

The system behaved as expected

It was easy for me to load a file into the app

I found the app particularly easy to use.

Statements on the app's potential for real world use:

There are features I would like to see added to this app that would make me more likely to use it

If the output were presented in a more helpful way, I would be more likely to use this app

I would use this app again if it were available as a paid app

All of the subjects except two were successfully able to run the BICC analysis for their own native language. At the time of this study, phonetic representations of Greek and Russian were not supported for analysis. Thus the Greek speaker chose to use the analysis for Spanish speakers, believing the accent and pronunciation errors to be similar between the two languages. The Russian speaker also chose to use the analysis for Spanish speakers, and answered the questionnaire as if Spanish had been his native language.

5. Results

This section shows subjects' responses to the questionnaire for:

- effectiveness of the feedback
- the usability of the app
- potential for real-world use.

5.1 Effectiveness of the feedback

The results shown in Figures 5, and 6 confirm that the subjects overall found the app to be useful with most answers in the 4-5 range (strongly agreeing with positive statements about the app). We did note that not all inappropriate contexts had been filtered out.

We also gathered verbal comments from the subjects. Several subjects remarked that the "trouble phones" they saw were indeed sounds they have trouble with. Some actually thought that the app had used speech recognition to detect that they had pronounced these phones incorrectly (it did not). Some of the words that the app showed to the users had not been used in their summaries. Subjects seemed to be paying more attention to their speech in the summary they recorded after using POLLI. Some subjects tried to use the flagged words to show that they were paying attention.

Figure 5. POLLI taught me something

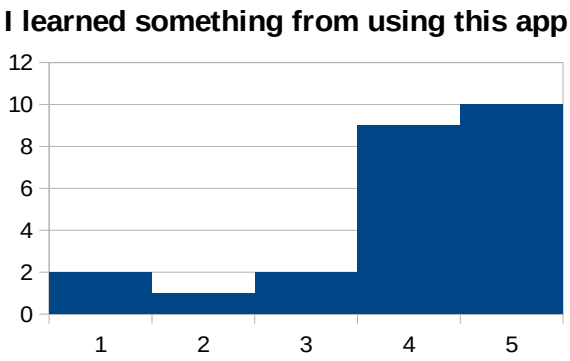
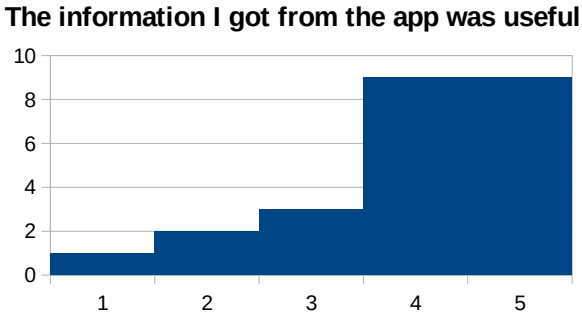


Figure 6. Useful information



5.2 Usability of the app

The results shown in Figures 7. 8 and 9 for usability continue to be positive. The system only crashed the on two texts at the beginning of the study and was always able to give some feedback. Subjects found that the system met their expectations, which not only reflects on the quality of system function, but on the way we chose to present it. The ease of use expressed in Figures 8 and 9 certainly contribute to the overall positive impression.

Figure 7. system meets expectations

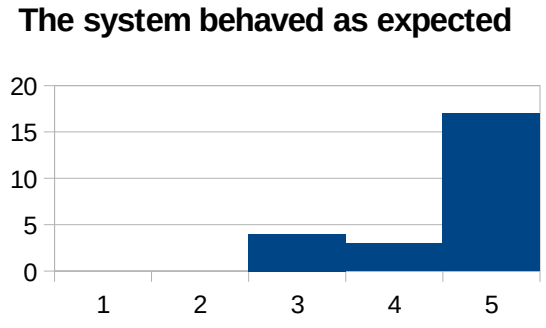


Figure 8. Ease of text upload

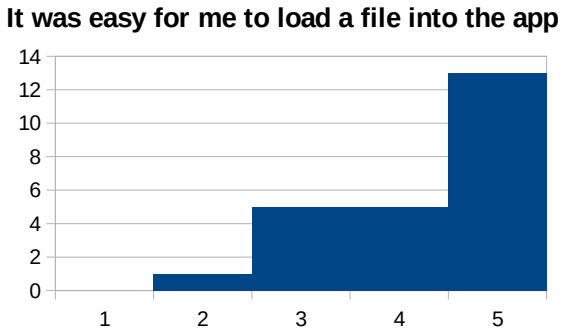
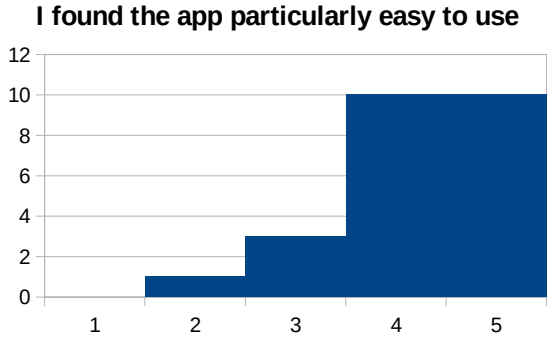


Figure 9. Ease of use



5.3 Potential for real use

Figures 10, 11 and 12 reflect what subjects think of POLLI's potential for real use. Subjects were asked if they could think of add-ons that they would like to see in POLLI that would make it more useful. Many of them told us that they would like to be able to speak to it and to get pronunciation feedback. They seemed to appreciate the way BICC's output is presented on the screen. Some believed, however, that the output could be enhanced with the use of either recorded speech or speech synthesis so that they could hear what a specific context should sound like. Although the subjects did not mention it, they would probably also benefit from hearing their own recordings. Subjects would have liked for the confusable contexts to be ranked in some way (order of appearance, difficulty, probability of making that specific mistake, etc.) and some would have liked to see more text around the confusable contexts.

Presenting subjects with the choice of whether POLLI should be free or paid seemed to be another appropriate way to gauge acceptance. With many apps being offered for free, it was reassuring to find that subjects stated that they would be willing to pay around $1 or $2 for POLLI.

Figure 10. Should POLLI have more features

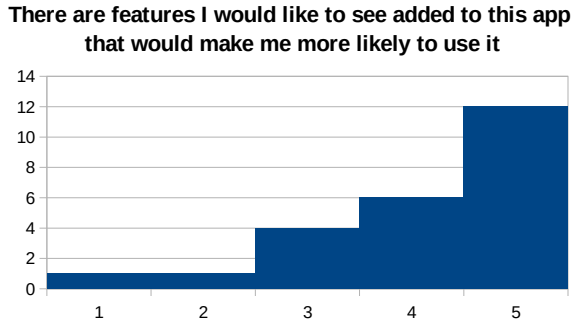


Figure 11. Presentation

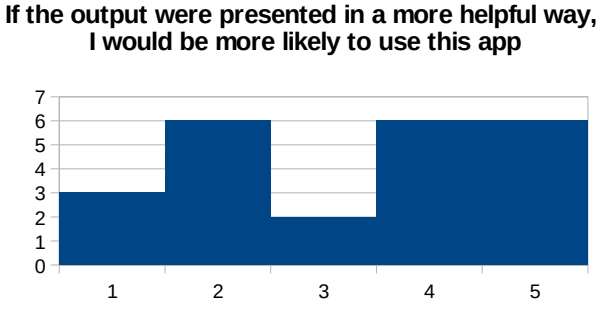
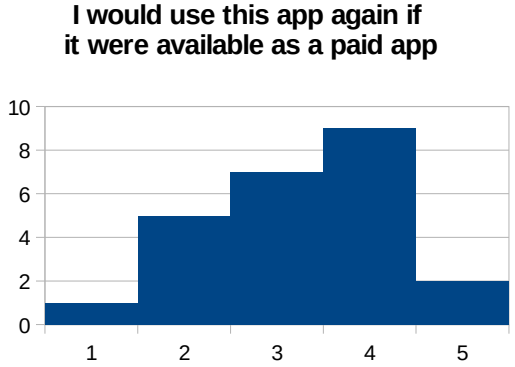


Figure 12. Recurrent use



6. Conclusions

We have described POLLI, an Android app that helps non-native students preparing presentations. POLLI embodies both the everpresence that is desirable in language learning and the focus that we believe reflects what students will most want in the future. While we cannot yet conclude that this presentation format affords more learning than a regular laptop interface (and we will want to test this hypothesis in the future), we do observe that the subjects who have tried POLLI have positive feedback and seem very motivated to use it. Since motivation has been shown to have a positive effect on learning [10], we believe that POLLI can make a positive impact on learning.

Acknowledgements

The first author Is sponsored by the Semiconductor Research Corporation as part of their Undergraduate Research Opportunities program.

References

[1] Rosetta Stone 2013, http://www.rosettastone.com/mobile-apps (accessed 4-9-13)

[2] British Council 2013, http://learnenglish.britishcouncil.org/en/apps (accessed 4-9-13)

[3] Viberg, O., Gronlund, A., 2012, Mobile Assisted Language Learning: a literature review, http://ceur-ws.org/Vol-955/papers/paper_8.pdf

[4] SOUNDS, https://itunes.apple.com/us/app/sounds-pronunciation-app-free/id428243918?mt=8, (accessed 4-9-13)

[5] Wikipedia MALL http://en.wikipedia.org/wiki/Mobile_Assisted_Language_Learning (accessed 4-9-13)

[6] Carnegie Speech, http://www.carnegiespeech.com (accessed 4-9-13)

[7] Catia Cucchiarini, Joost van Doremalen, Helmer Strik: Practice and feedback in L2 speaking: an evaluation of the DISCO CALL system. INTERSPEECH 2012

[8] Saz, O., Eskenazi, M., 2012, Addressing confusions in spoken language in ESK pronunciation tutors, Proc. Interspeech 12 Portland.

[9] Saz, O., Eskenazi, M., 2011, Identifying confusable contexts for automatic generation of activities in second language pronunciation training, Proc. SLaTE 2011, Venice.

[10] Dela Rosa, K., Eskenazi, M., 2011, Self-Assessment of Motivation: Explicit and Implicit Indicators in L2 Vocabulary Learning, Proceedings of the 15th International Conference on Artificial Intelligence in Education (AIED 2011).

GOBL: Games Online for Basic Language Learning

Helmer Strik[1,2], *Polina Drozdova*[1], *Catia Cucchiarini*[1]

[1]Centre for Language and Speech Technology, Radboud University, Nijmegen, the Netherlands
[2]Department of Linguistics, Radboud University, Nijmegen, the Netherlands
w.strik@let.ru.nl, p.drozdova@let.ru.nl, c.cucchiarini@let.ru.nl

Abstract

In the GOBL project we develop and test small web-based mini games for low-educated and disadvantaged beginning learners of Dutch, English, and French. An innovative aspect of this project is that we incorporate speech and language technology (SLT) to practice speaking skills. The present paper explains the notion of mini games employed in the project and the advantages of their use with respect to the target group. We then present the first results of the project concerning pedagogical and user requirements, based on the literature and user-based research. We then introduce our plans for the immediate future.

Index Terms: serious gaming, speech and language technology, second language acquisition

1. Introduction

Worldwide millions of people have to learn foreign languages or a second language that they need to integrate in new socio-economic contexts. Still, some groups participate to a lesser extent, or lose interest in formal language education altogether. Moreover, it may be argued that in foreign language curricula speaking practice receives relatively little attention.

Adults and young people with low foreign language skills are commonly marginalized in European society. More specifically, a lack of basic communicative proficiency in foreign languages is an obstacle for (re-)integration into the labor market, especially in SMEs [1]. Also, adults and young people who lack basic communicative skills in foreign languages are less mobile in Europe, in educational as well as in economic contexts [3].

Although European member states invest significant effort in foreign language education, a number of social groups display a low degree of participation in language learning programs. First, young people who drop out of formal education before attaining a degree from upper secondary education do not acquire the foreign language skills needed to integrate into the labor market and European society. Although Early School Leaving has decreased throughout Europe in the period spanning 2000-2009, more effort is needed to reduce ESL under 10% by 2020 [10]. A second group comprises adults, whose participation in lifelong learning programs is still relatively low, despite considerable efforts at European level [11]. This results in inadequate language skills, especially among socially disadvantaged groups. A third group consists of immigrants in particular. The European Union considers acquiring at least basic knowledge of the language of the host country a fundamental instrument for successful integration into European society [9], as well as into the European labor market [18]. For these three groups, i.e. early school leavers, adults with low participation in language learning programs, and immigrants, low proficiency in a foreign language seems to be related to motivational factors and a lack of awareness, and it may be argued that these groups need an approach which transcends the national education systems.

With respect to the foreign language curriculum, two aspects deserve more attention. First, practicing speaking skills, in general, receives little attention in language classrooms, because of lack of time. However, speaking is a crucial skill in a second or foreign language and is relevant for all learners, independent of their background and educational or professional objectives. Research has indicated that practicing speaking skills is essential to learn to speak the target language [7]. Second, it may be argued that attention to formal and explicit knowledge of grammar and vocabulary has waned in recent years. Research shows that formal aspects of language acquisition are just as important to achieve proficiency in a second or foreign language as fluency-related aspects [14], and that grammar errors are known to persist also after years of immersion in the language of the host country [12]. The provision of feedback on errors is crucial in order to remedy this situation [20].

In the present paper we first describe mini games in relation to language learning (Section 2). We then introduce our project GOBL, describing its objectives and the approach adopted. Section 4 is about the design of GOBL., we first present the pedagogical needs analysis, the user-based needs analysis, and the GOBL implementation. In Section 5 we explain how speech and language technology is employed in GOBL.

2. Mini games and language learning

Educational mini games are small and self-contained games which are highly reusable, cost-effective, and motivating and focus on specific well-defined learning topics.

Games have been put to use for language teaching purposes over the last few years. The main advantage of using games for language learning is that the user tries to achieve a non-linguistic goal: reaching a new level, obtaining more points. It makes the process motivating for language learners. It has been shown that educational mini games lead to fast gains in L2 vocabulary and to increased speed of lexical access [4].

Virtual world-based (2D/3D) language learning games have been developed [15], [16], [23], but these products typically target advanced language learners. Moreover, many of these games remain in a phase of prototyping, are only used for research, or stay within the confines of the academic or military world. This may be due to practical reasons, such as the availability of expensive hardware, but also because these products are mainly technology-driven. As a result, many of these virtual world language learning games are not accessible to the

large mass of (low-skilled) language learners, who need them most.

Mini games, on the other hand, require only basic technical skills and hardware and are easy to use. They are typically embedded in other websites, such as social networking services. It makes them easily accessible and particularly fit for low-skilled and/or disadvantaged language learners. Research [13] indicates that resource-deprived language learners make more use of the web (including games) as a medium for entertainment than highly educated people, and prefer and profit more from mini games than from complex strategic games.

A number of mini-game products exist, both via commercial licenses [24], [25] and through freely accessible websites [26]; [27]. The latter free websites, however, do not offer tracking and logging capabilities which do exist in non-game-based free e-learning environments. As a result, these websites offer a one-size-fits-all approach for the learning content, and take learner interests or characteristics only to a limited extent into account. In summary, existing mini games for language education are not adapted to the needs of the learner.

Opportunities to practice speaking skills through mini games are missing altogether. Automatic speech recognition (ASR) technology has until now especially been integrated only in full-immersive avatar-based games [16], [17].

First releases of games seldom reach the market success that is obtained by game titles that have several versions. This is mainly due to the complex design process of games, which needs to keep a close eye on technological, content-related and motivational aspects. For educational games, this problem is exacerbated by the fact that game design needs to be leveled with instructional design and teaching methods, and that this medium still has to catch on in formal education. This not only requires a strongly inter-disciplinary and cross-disciplinary approach, but also a methodology that takes into account the complex interplay of user research, instructional design, content development, and technological development.

In a nutshell, what is missing today are easily accessible web-based mini-game content for practicing basic language, including oral skills, integrated into a platform which motivates learners to keep practicing, also outside of formal language teaching contexts. The GOBL project aims to fill this gap in the state-of the-art so as to increase the proficiency of low-skilled language learners, both in formal and in informal learning contexts. The description of the project together with the solution it provides for the above mentioned problems is given in the following section.

3. Games Online for Basic Language Learning

The 'Games Online for Basic Language Learning' (GOBL) project started in January 2012. Participants of the project include the University of Nijmegen, the University of Leuven, the University of Newcastle upon Tyne, Televic Education and Council of Scientific and Industrial Research of Meraka Institute. The project aims at the providing youths and adults learning French, Dutch, or English with access to on-line mini games to improve their speaking proficiency, grammar and lexicon. In this section we present project objectives and project approach employed to reach these objectives and address the issues mentioned in the previous section.

3.1. The GOBL objectives

In the GOBL project we are aiming to develop mini games in a user-centered way for teaching of grammar in use, vocabulary and basic communicative skills in French, English and Dutch as a foreign or second language. As research has shown [14], a good command of these linguistic aspects is as necessary to achieve proficiency in a second language as fluency-related aspects. Learning materials target the A2 level of the CEFR. Dedicated speech recognition technology is employed for stimulating speaking practice in Dutch and English, and mini games and accompanying materials will be made available online.

In this project we address the needs of learners with low language skills, with a special attention to low-educated and disadvantaged youths and adults, because there is a high demand for qualitative, motivating learning materials for the lower levels of the CEFR. We choose gaming (instead of other tuition methods) because it has been shown to be an appealing medium of learning for learners in this category [13] and because we believe it can assist in overcoming the lack of motivation that is often observed in the social categories mentioned.

The materials proposed for the project specifically address such components in the foreign language curriculum that deserve more attention and can be easily incorporated into mini games. It means focusing on well-defined topics and communicative situations relevant for the target groups that take their expectations and goals into account. The advantages are that these aspects can relatively easily be addressed through human language technology while the extra teacher time that is freed up when these aspects are addressed in mini games can be employed for practicing other linguistic aspects that do require interaction with a teacher.

From a societal point of view, a crucial element is that the mini-game content developed within the project will be easy accessible for low-skilled language learners. In this way, we aim to bridge the gap between formal and informal learning for this particular group, to help them (re-)integrate into the labor market and society.

An additional, innovative aspect of this project is that speech and language technology, and especially ASR technology, is incorporated in many exercises, which makes it possible for learners to practice speaking skills and receive corrective feedback from the system on their speaking performance.

3.2. The GOBL approach

To make the complex design process easier, a number of important decisions have been taken at the beginning stage of the project, namely, to:

- focus on small and self-contained mini games;
- identify and isolate potential problems early on in the project by making lists of technical, user- related and pedagogical requirements, in order to limit the design space and reduce risks in development;
- rely on existing platforms and technology as offered by Televic Education and by the Centre for Speech and language technology (especially ASR).

An iterative and user-centered methodology, which will allow to design, develop and test the product in small but well-organized steps, is employed in the project. Target users are involved in several stages of the design and testing, so that we can take into account their needs, and can deliver a product which

appeals exactly to these audiences. Independent external experts are asked to give advice and feedback on pedagogical and gaming aspects.

Target groups that benefit the most from the project include:

1. low-skilled adult and young learners of English, French and Dutch as a foreign and second language in various European countries and beyond, who can use the language learning mini games developed in this project.
2. teachers of these languages in various countries who can use the mini games in the courses;
3. language teaching institutions which will be able to use the mini games and evaluate their use by adopting the various evaluation instruments developed in this project;
4. publishers of language learning materials which can use the developed mini games as examples;
5. companies that develop ' computer-assisted language learning' (CALL) technology and applications, which will be given the opportunities of using the speech data and relative annotations made available by the project partners;
6. academic institutions that carry out research on language learning and teaching, lifelong learning, CALL and speech technology, which will have the opportunity of using the learning material, the protocols, the data collected and the evaluation instruments for their own research.

Language teachers and learners have been involved in focus groups and in the evaluation to provide input for system design. Moreover, contacts have previously been established with the consortium partners and their networks of language teaching institutions, publishers and CALL companies. At the beginning stage of the project focus groups interviews were conducted and questionnaires distributed among the language learners and instructors, which formed the basis for the list of user-related and pedagogical requirements.

4. The GOBL Design

For the design process we followed a procedure similar to the one used in the DISCO project [22]. To inform the design process a list of pedagogical, user-based and technological requirements was drawn up in the beginning stage of the project. The following section reports the results of pedagogical and user-based needs analysis, together with the steps which were undertaken within the project to answer these needs.

4.1. Pedagogical needs analysis

The list of pedagogical requirements was formulated on the basis of a literature review. It was decided that all mini games should be embedded in tasks, eliciting real language use. The task-based approach [19] thus formed the pedagogical framework for the design. Since the learners are trying to reach a particular goal while playing the game, language becomes the resource to reach this goal, and language use acquires an additional meaning.

Moreover, while performing the game tasks, users should receive sufficient support in the form of corrective feedback, which is one of the central elements of games [1]. The feedback can be provided in the form of right/wrong responses during the game, but should be more detailed in the end of the game.

Finally, mini games should primarily target development of fluency and accuracy, because of their fast pace and focus on particular aspects of language form. Since the aim is to develop materials for low-proficient language learners, the sentences used should not be too long and complex, so that a balance can be achieved between the speed of the game and its complexity.

4.2. User-based needs analysis

In May – June 2012 the user-based need analysis was conducted among language learners at the target levels of proficiency (A2-B1 according to CEFR) in the form of a task-based focus group. Additionally, their language instructors were interviewed.
The focus group with learners included the following phases:

1) a warm-up phase, during which the learners reported on the difficulties they have in learning the language;
2) a phase where the learners were introduced to and could play some of the existing mini games (Article Wolf [29], Frog Verbs [30], Beat the Keeper [27], Mindsnacks [25]), and share their experience and opinions afterwards;
3) a phase in which the overall scenario and first mock-ups of the games under development were introduced;

Most participants evaluated grammar and speaking skills as the most difficult and necessary to acquire. For the games to be practically useful they should be relevant for the target users and contain communicative situations the learners come across in real life. A number of such topics was mentioned: going to the doctors, getting a citizenship document, job interview, etc. Moreover, some of the participants mentioned that the scenario should be adapted to the needs and gender of the players, and that they would like to be able to choose the topics themselves.

In general, the participants were positive about the games which were demonstrated to them. It was mentioned that mini games can be used for additional practice outside the classroom. The fast pace of the game was found motivating, but the learners would like to be able to adapt it to their needs. Motivation for playing the game can also be facilitated by providing rewards for correct answers, or through a competition with others. Another important motivating aspect mentioned by both teachers and learners is providing comprehensive feedback. Seeing that your answer was right or wrong is not enough, the users would like to know why it was incorrect.

Graphics and music were found to be important factors: the learners in the Dutch group, for example, did not like the first games presented to them, since the graphics were too simple. At the same time one of the reasons why the Mindsnacks game was preferred to the others was the good-matching music.

Finally, the participants commented on the possible scenario of the GOBL mini games. Both the learners and the instructors mentioned the necessity of full immersion into the game, which can be achieved through presenting the tasks in a particular scenario. As suggested by the learners from Belgium it would be interesting to play the role of a detective and solve some mysteries or murders while playing mini-games. At the same time, the idea about uniting the developed mini games in one scenario received mixed feedback from the learners. It was considered important that the games can be played separately from the scenarios as well.

4.3. Implementation in GOBL

The results of the analyses were taken into account in the development of the first demos of the mini games. Following the comments of the learners, the initial scenario around the reporter was changed to the scenario of a detective story, where the

learner plays the role of a detective and has to find the stolen cookbook with chocolate recipes.

Three game types were developed within the project:

1) The lie-detector game, where the learner has to decide whether the sentence pronounced by the suspect is correct or incorrect;
2) The finger print-collector, where the learner has to collect finger prints in the form of the words, which can be used to fill in the gap in the sentence.
3) The roof-surfing parrot, where the learner has to move a parrot and save it from smog by pronouncing or clicking on the sentence to continue the dialogue.

All the games are time-paced, so that the learner has to provide an answer at a given time. In the case of an incorrect answer, the learner gets the "right-wrong" feedback from the system, and can see the result of his/her activity on the screen: the suspect is happy, the finger-print is destroyed, the parrot is suffocated by the smog. An example of one of the forms of such feedback is given in Figure 1. Moreover, after playing the game, the learner receives a feedback screen showing which items he answered correctly or incorrectly. The need for more detailed feedback mentioned in the needs analysis will be considered after evaluation of the first prototypes.

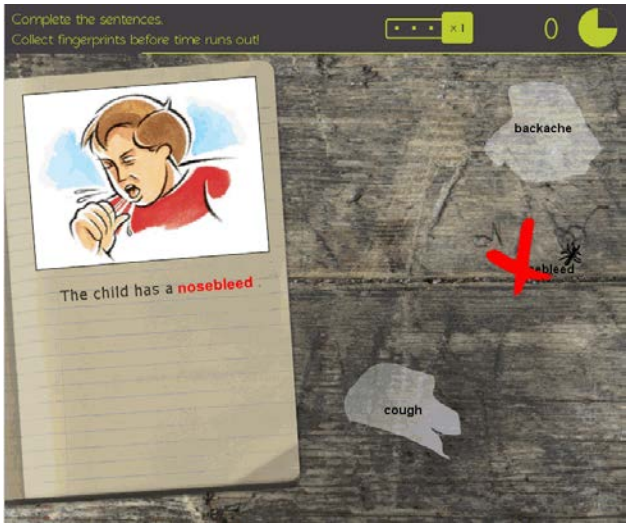


Figure 1: *A screen-shot from the finger-print collector game. The learner has chosen the wrong answer and destroyed the finger-print. The number on the top of the screen shows how much time the learner still has to complete the task.*

The following motivational strategies mentioned by the second language learners and teachers during the needs analysis stage have been incorporated in the current design of the mini-games: timing (the learner has to answer at least half of the questions correctly before time elapses) and scoring (the learner receives points for each correct answer and extra points for a number of consecutively correct answers). The scoring board is planned to be implemented in the future where the users will be able to compare their own scores after different attempts and/or to compare their result to the results of their peers.

Since music and graphics were mentioned to be important for the target users, background music was incorporated in all the games. The appropriateness of the chosen music and graphics will be evaluated during the evaluation stage.

To answer the needs of the learners at the target levels of proficiency, a list of relevant linguistic topics and themes was created for developing the content. Vocabulary and grammar can be practiced with the help of the first two games, while the "roof-surfing parrot" mini-game is in the form of a dialogue. Two of the games presented will be powered with ASR to practice speaking skills. The incorporation of speech and language technology (SLT) in the exercises is discussed in the following section.

5. Speech and language technology

An innovative aspect of GOBL is that SLT and in particular automatic speech recognition (ASR) technology is incorporated in many exercises, which makes it possible for learners to practice speaking skills and receive corrective feedback from the system on their speaking performance.

There are 3 possible versions of the CALL system:

1. no ASR, completely text-based
2. with ASR, but it will not determine the flow of the game
3. ASR is decisive for the game

Not all exercises are suitable for ASR. Since good (acoustic) conditions are required for the optimal functioning of ASR, it should be possible for the user to practice with non-ASR versions of the mini games if these conditions are not available (see below). Then there are two options: the user is warned that the conditions are not optimal for ASR, but still can use ASR, or if the conditions are not good enough ASR cannot be used at all.

LST will be employed to analyze the learners' language output. Therefore, LST is developed for Dutch and English versions of the mini games, to recognize and further process the spoken utterances produced by the learners. At first no assumptions are made regarding language pairs (all L1's). Later we can look into specific, frequent L1-L2 pairs (as an extension). In developing technology for GOBL, we build on technology developed for other projects at the Centre of Language and Speech Technology (CLST) of the University of Nijmegen, such as the projects Dutch-CAPT [32], DISCO [33], MPC [34], FASOP [35], DigLIn [36] (for an overview see also [21]).

The integration of ASR technology in CALL programs needs to be done with specific care. ASR technology has reached a level of maturity sufficient for language learning applications, but still has a number of limitations, which need to be taken into account in the design process. First, in the context of web-based mini games, ASR technology needs to be optimized for web-based delivery, and the software design has to take into account several platforms, browsers, and contexts in which it will be used. Second, ASR for foreign language learning needs to be adapted to the speech of non-native speakers, and to the kind of mistakes they make.

Recognition of non-native speech is more complex than the recognition of native speech [2]; [5]; [8]. In order to deal with this increased complexity, specific LT modules are developed and optimized. The exercises are organized in such a way that the LST modules can handle the user's spoken output. Spoken utterances are elicited such that the possible (correct) answers by the users are restricted. The examples of the mini-games within the GOBL project which can be powered by an ASR are given in Figure 2 and 3.

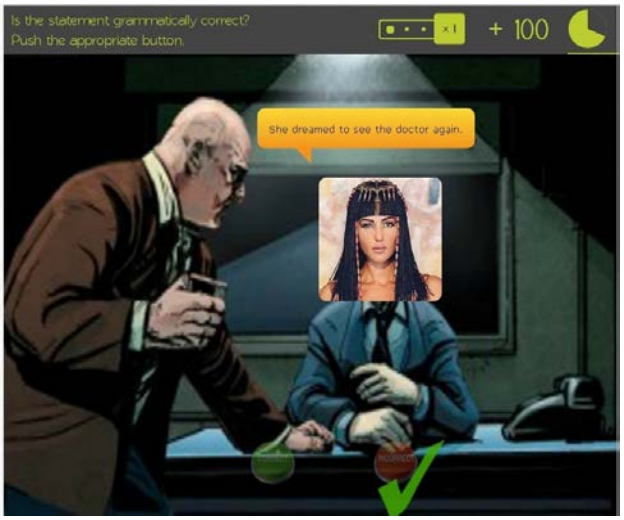


Figure 2: *A screenshot from the lie-detector mini-game. In the version powered by an ASR the learner has to repeat the sentence if it is correct, otherwise press "lie".*

Figure 3: *A screenshot from the roof-surfing mini-game. In the version powered by ASR the learner has to save the parrot from smog and move him from one roof top to another by repeating the appropriate sentences to build a coherent story.*

For each exercise a list of possible and probable correct and incorrect responses is drawn up. This list is then employed to build a specific language model for each individual exercise. The language model is created automatically. For every exercise the CALL system has to choose among these correct and incorrect responses. The most suitable decoding technique is applied.

After the LST modules have determined what was spoken and where mistakes have been made, the appropriate feedback has to be generated. For this purpose, the study of the most appropriate feedback moves is necessary.

Two types of feedback are possible:

1. Implicit feedback during the game
2. Explicit feedback after the game

The game has to be fast-paced. Therefore, during the game only feedback is provided that does not hinder the speed of the game. If an exercise is not carried out correctly, or not within the time limits, the user gets negative feedback (e.g. points are deducted). The user thus gets implicit feedback, which increases the 'game feeling', and supports learning. Still, it might also be possible to give some explicit feedback during the game, such as highlighting errors (e.g. denote pronunciation errors by underlining or coloring the corresponding graphemes).

After the game, the user can get explicit feedback, e.g. an overview of (language) errors made. This kind of feedback is language learning supporting, and thus probably not part of the game play (penalty, reward, etc.). The game keeps track of errors made during the game, and after the game the user can get feedback on these errors in different ways. In some cases users can listen to correct examples, model-answers. These model answers are recorded speech utterances. Speech synthesis is also an option provided the quality is sufficiently high.

The application is web-based, accessible in various browsers. A client-server architecture is used, and obviously the use of a microphone should be supported. The server performs the computationally most demanding tasks, while the ('thin') client performs more simple computational routines. The server consists of three components: (1) the course software; (2) the LST software, and (3) the content.

ASR of non-native speech is already challenging, as was already mentioned above. However, there are also other problematic issues such as noise (especially background speech), and end-point detection (EPD).

We intend to use head-sets for better performance. In addition we employ standard noise reduction and adaptation techniques such as 'spectral subtraction', 'cepstral mean subtraction' (CMS), and 'vocal tract length normalization' (VTLN). The question is whether calibration is necessary. If is it going to be used, it has to be a short procedure, probably at the start, possibly combined with an automatic (background) procedure during the game. Or calibration is done only when there are problems. The experiments we have carried out have made it clear that at start it is preferable to show the voice level (VU meter), and an option to choose between different microphones on the computer. Furthermore, users also prefer to see an indication of the voice level during use, mainly to be sure that the microphone is still working correctly (especially if the system tells them they have made an error).

The question is also what kind of end-point detection (EPD) is best suited for a fast-paced game. Options are:

1. push-to-talk, i.e. user specifies begin and end
2. user only specifies begin (e.g. with space bar), and automatic detection of the end
3. automatic detection of begin and end.

In options 2 and 3, for automatic detection of the end, the length of the correct answer(s) might be useful information. Option 1 is probably less error-prone. However, for an optimal game feeling option 3 might be better. In practice, we have to find the compromise that works best, maybe option 2. To start recognition immediately, while the utterance is being pronounced, streaming of the user speech seems the best option.

A first version of the system will be ready in April-May 2013. In May-June 2013 it will be evaluated with language learners.in the Netherlands, Belgium, the UK, and South-Africa. At the SLaTE 2013 workshop, the system will be shown and results will be presented. The results and the feedback will be taken into account, and used to develop a second, improved version of the system, which again will be evaluated.

6. Acknowledgements

This project has been funded with support from the European Commission under project number 519136-LLP-1-2011-1-NL-KA2-KA2MP. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.
We are indebted to the other members of the GOBL team for their contributions, in alphabetical order: Frederik Cornillie, Piet Desmet, Febe de Wet, Johannes De Smedt, Andrew Grenfell, Marijn Huijbregts, Ann-Sophie Noreillie, Thomas Snell, Sylvie Venant, and Scott Windeatt.

7. References

[1] Becker, K. (2007). Pedagogy in commercial video games. In M. Prensky, C. Aldrich, & D. Gibson (Eds.), *Games and simulations in online learning: research and development frameworks* (pp. 21–47). Hershey: Information science.

[2] Benzeghiba, M., R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, Automatic speech recognition and speech variability: a review, Speech Communication, vol. 49, no. 10-11, pp. 763-786, 2007.

[3] Bonin, Holger, Eichhorst, W., Florman, C., Hansen, M.O., Skiöld, L., Stuhler, J., Tatsiramos, K., Thomasen, H., Zimmerman, K.F. (2008). Geographic Mobility in the European Union: Optimising its Economic and Social Benefits. IZA Research Report No. 19.

[4] Cobb, T., & Horst, M. (2011). Does Word Coach Coach Words ? CALICO Journal, 28(3), 639-661.

[5] Compernolle, D. van (2001) Recognizing speech of goats, wolves, sheep and..non-natives, Speech Communication, vol. 35, no. 1-2, pp. 81-79, 2001.

[6] Davignon, E., Albrink, W., Dyremose, H., HJanssen, M., Jenner, C., Gomes de Pinho, A., Hussain, W., Klimek, S., Legernes, L., Mathews, P., Kostoris Padoa Schioppa, F., Proszeky, G. (2008). Languages Mean Business. Companies work better with languages. Recommendations from the Business Forum for Multilingualism. Luxembourg: Office for Official Publications of the European Communities.

[7] DeKeyser, R. (2007). Practice in a second language, chapter Introduction: Situating the concept of practice. New York: Cambridge University Press.

[8] Doremalen, J. van, Cucchiarini, C. & Strik, H., Optimizing automatic speech recognition for low-proficient non-native speakers. EURASIP Journal on Audio, Speech, and Music Processing, Article ID 973954, 13 pages, 2010.

[9] European Commission (2005). Communication from the Commission to the Council, the European Parliament, the European Economic and Social committee and the Committee of the Regions - A Common Agenda for Integration - Framework for the Integration of Third-Country Nationals in the European Union.

[10] European Commission (2010). Reducing early school leaving. Accompanying document to the Proposal for a Council Recommendation on policies to reduce early school leaving. Brussels. Retrieved from http://ec.europa.eu/education/school-education/doc/earlywp_en.pdf .

[11] European Commission, Action Plan on Adult Learning. It is always a good time to learn. 2007. Retrieved from http://ec.europa.eu/education/policies/adult/com558_en.pdf

[12] Han, ZhaoHong. (2004). Fossilization in Adult Second Language Acquisition. Clevedon: Multilingual Matters.

[13] Herselman, M. E. (1999). South African Resource-Deprived Learners Benefit from CALL through the Medium of Computer Games. Computer-Assisted Language Learning, 12(3), 197-218.

[14] Housen, A. & Kuiken, F. (2009) Complexity, accuracy and fluency in second language acquisition. Applied Linguistics 30, 4, 461-473.

[15] Hubbard, P. (2002). Interactive Participatory Dramas for Language Learning. Simulation & Gaming, 33(2), 210-216.

[16] Johnson, W. L., Vilhjalmsson, H., & Marsella, S. (2005). Serious games for language learning : How much game, how much. AI ? AIED 2005. IOS, Amsterdam.

[17] Li, R.-C., & Topolewski, D. (2002). ZIP & TERRY: A New Attempt at Designing Language Learning Simulation. Simulation & Gaming, 33(2), 181-186.

[18] Lodovici, M.S. (2010). Making a success of integrating immigrants into the labour market. Retrieved from http://www.peer-review-social-inclusion.eu/peer-reviews/2010/making-a-success-of-integrating-immigrants-into-the-labour-market/synthesis_report_no10/download

[19] Purushotma, R., Thorne, S. L., & Wheatley, J. (2008). 10 key principles for designing video games for foreign language learning. Retrieved from http://lingualgames.wordpress.com/article/10-key-principles-for-designing-video-27mkxqba7b13d-2/

[20] Sheen, Y. (2010). The Role of Oral and Written Corrective Feedback in SLA. Studies in Second Language Acquisition, 32, 169-179.

[21] Strik, H., ASR-based systems for language learning and therapy, Proc. of IS-ADEPT: International Symposium on Automatic Detection of Errors in Pronunciation Training, KTH, Stockholm, Sweden, 6-8 June, pp. 9-14, 2012.

[22] Strik, H., Cornillie, F., Colpaert, J., van Doremalen, J., & Cucchiarini, C., Developing a CALL System for Practicing Oral Proficiency: How to Design for Speech Technology, Pedagogy and Learners. Proceedings of the SLaTE-2009 workshop. Warwickshire (England), 2009.

[23] Sykes, J. M. (2008). A dynamic approach to social interaction. Synthetic immersive environments & Spanish pragmatics.

[24] http://mywordcoach.us.ubi.com/

[25] http://www.mindsnacks.com/

[26] http://www.digitaldialects.com/

[27] http://learnenglish.britishcouncil.org/en/games

[28] http://beta.visl.sdu.dk/games_gym.html

[29] http://www.english-online.org.uk/games/articlesframe.htm

[30] http://www.english-online.org.uk/games/pasttense.htm

[31] http://www.gobl-project.eu/

[32] http://hstrik.ruhosting.nl/wordpress/dutch-capt/

[33] http://hstrik.ruhosting.nl/wordpress/disco/

[34] http://www.ru.nl/arts/mpc/

[35] http://hstrik.ruhosting.nl/wordpress/fasop/

[36] http://diglin.eu/

Enhancing Speech Recognition in Fast-Paced Educational Games using Contextual Cues

Carrie J. Cai, Robert C. Miller, Stephanie Seneff

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, Massachusetts 02139, USA
{cjcai,rcm}@mit.edu, seneff@csail.mit.edu

Abstract

Arcade-style games like Tetris and Pacman are often difficult to adapt for educational purposes because their fast-paced intensity and keystroke-heavy nature leave little room for simultaneous practice of other skills. Incorporating spoken language technology could make it possible for players to learn as they play, keeping up with game speed through multimodal interaction. To date, however, it remains exceedingly difficult to augment fast-paced games with speech interaction because the frustrating effect of recognition errors highly compromises entertainment. In this paper, we design a modified version of Tetris with speech recognition to help students practice and remember word-picture mappings. Using utterances collected from learners interacting with the speech-enabled Tetris game, we present and evaluate several techniques for leveraging contextual cues to increase recognition accuracy in fast-paced game environments.

Index Terms: speech recognition, education, serious games, user interfaces

1. Introduction

The pervasive spread of computer games has made a significant impact on game-based learning as a serious topic in the field of education. Research evidence has shown that fun and enjoyment are central to the process of learning because they increase learners' intrinsic motivation [2,9]. Good games can motivate players to learn through repeatedly doing the game itself until they have virtually automatized the new skill [4].

Although the highly engaging, repetitive nature of existing arcade-style games makes them natural settings for embedding learning through rehearsal, most adaptations of existing games emerge from turn-based frameworks like card games [10] or from complex virtual environments [14], perhaps due to less time pressure on learners and greater amenability to structural changes. However, arcade-style games such as Tetris and Pacman are advantageous in that they are much simpler to manipulate by developers, have open source code bases, and allow a wider range of time commitment from players. Just as flashcards enable students to review vocabulary on the run, arcade games allow players to either indulge in short spurts or stay indefinitely.

Augmenting games with speech interaction offers multiple advantages for adapting such games for learning. Not only does speech production strengthen memory by providing learners with phonological input back to the mind [8], but speech is also a typically unused input channel during traditional arcade gameplay. It could therefore enable users to keep up with the original game speed more so than text input. Previous work has further indicated that embedding motivations for *retrieval practice*, the act of repeatedly attempting recall from memory, could improve long-term retention in a speech-augmented game environment [3]. However, fast-paced games offer an unusual challenge in that their motivational effectiveness depends heavily on the rhythm and flow of the game, along with clear accountability for progress [12]. The thrill of playing a fast-paced game could be seriously dampened by the frustrating effect of speech recognition errors, a reason that perhaps explains the limited adoption of speech technology in this area.

Recent work has explored using dialogue context to enhance speech understanding, both in standard information-access systems [13][16] and in dialogue systems for second language learning [15]. However, less research is devoted to enhancing speech recognition systems in time-sensitive settings for rapid gameplay. Fast-paced arcade style games may offer the advantage of providing even more fine-tuned contextual information, due to simpler game logic, fewer possible states, and a more granular trial-by-trial structure.

In this paper, we investigate useful techniques for enhancing speech recognition performance by using in-game context to provide additional information to the recognizer. We use Tetris as a prototypical example for evaluating these approaches. Tetris is classic arcade-style video game in which players prevent falling blocks from stacking to the top by rotating and maneuvering the blocks to form rows.

2. System Design

Building on an existing open source web implementation of Tetris[1], we modified traditional Tetris rules to offer an incentive for learning any set of associations, such as capitals and countries or names and faces. Our specific implementation teaches word-picture associations to help users learn and remember the meaning of words.

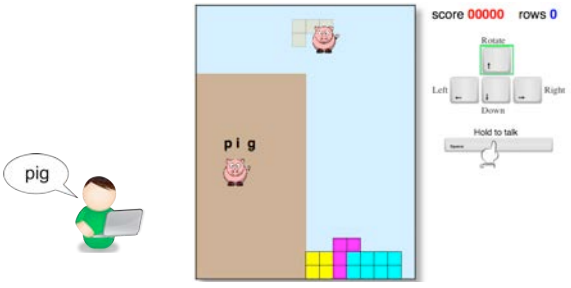


Figure 1: *Modified Tetris game interface. Saying the correct word unlocks block rotation.*

Each player sees a Tetris block attached to the picture and must correctly speak the word associated with the picture

[1] http://codeincomplete.com/posts/2011/10/10/javascript_tetris
© Jake Gordon

before block rotation can be unlocked for the trial (Figure 1). As in traditional Tetris, a block can only be maneuvered while it is still falling. Once it has dropped, the next block with a new picture immediately appears. Although our specific implementation allows learners to rehearse word-picture associations, the framework is not limited to pictorial cues and can be applied to learning any set of paired associations, in either the first or second language. For example, learners could practice recalling historical events and the dates on which they occurred, or scientific terminology and their definitions.

The game can furthermore be configured in three different modes: 1) In study mode, the word associated with the picture is presented each time the picture appears. 2) In retrieval mode, learners see the word-picture pair only the first time it appears, and in subsequent trials only see the picture displayed. The word is revealed if the learner says nothing after four seconds, or as soon as the learner records a response regardless of correctness. 3) Multiple choice mode is similar to retrieval mode, except in subsequent trials learners are aided by the display of two word options to choose between rather than having to exercise free recall. In all three modes, the learner hears the pronunciation of the word when the word-picture pair is first introduced.

To recognize speech input, we used the WAMI (Web-Accessible Multimodal Interface) software [6], a framework that allows audio to be captured at the web page and transmitted to the SUMMIT speech recognizer [5] running remotely. To enhance user input efficiency, we implement the voice recording functionality via a spring-loaded hold-to-talk spacebar (Figure 1) rather than the more traditional two step process of push-to-record followed by push-to-stop.

3. Speech Data Collection

We collected speech on Amazon Mechanical Turk by inviting remote participants to play the fully speech-enabled Tetris game multiple times, in different modes. Due to poor quality microphone hardware in many older computers, only participants who passed a pre-qualifier microphone test were allowed to complete the tasks. Within each game, learners were first introduced to a word-picture pair and then rehearsed the mapping four times, totaling 35 trials for the seven words per game.

In real-life situations, a learner may wish to learn or review words that may be missing from the recognizer's existing vocabulary, such as scientific terminology or proper nouns like *peroxisome* or *Nowocin*. To model these situations, our game presented an artificial vocabulary rather than existing words in the English language. The novel word-picture mappings also precluded any user from having a learning advantage due to prior exposure. We pre-generated the artificial vocabulary using a probabilistic model[1] on English phonemes. The final vocabulary consisted of 28 English-like words (Table 1) mapped to pictures of familiar animals and household objects. The lexicon for speech recognition used an English letter-to-sound model [1].

During gameplay, each user's utterances and game activity were logged to a database. In total, we collected 2584 utterances, at a sample rate of 8 kHz, from 16 users (12 male, 4 female) between the ages of 21 and 51 (mean=31.6). All participants were native English speakers located in the United States. Data for two sessions were not evaluated due to technical difficulties expressed in the user comments in a follow-up questionnaire. We thus perform evaluation on a total of 2351 utterances.

Vocabulary Words	
wug	blicket
speff	dax
pimwit	zigant
nanose	gazzer
tusket	toma
intess	fendle
priole	moffer
unty	illo
rint	del
mata	blas
pos	omma
tranco	atter
musker	corros
henne	barnel

Table 1: *The artificial vocabulary that users learned while playing the speech-enabled Tetris game. These words were randomly mapped to common animals and household objects.*

4. Evaluation of Static Recognizer

The recognizer's performance depends critically on its letter to sound (L2S) model used to generate lexical pronunciations for each out-of-vocabulary word. To evaluate the robustness of our L2S model, we utilized different pronunciation models ranging from one to twenty-best pronunciation hypotheses.

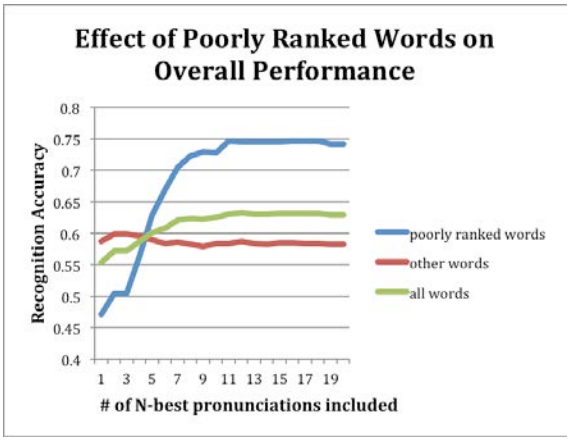


Figure 2: *Poorly ranked words (9 of 28) account for increased recognition accuracy when more pronunciations are included in the L2S.*

These N-best pronunciations were produced from the SUMMIT L2S model applied to the 28 artificial words. We configured a static recognizer with the full 28-word vocabulary and evaluated it on all utterances in which the speaker had produced any one of the 28 vocabulary words. When only one pronunciation per word was included in the L2S, recognizer performance was surprisingly low at 55%, but accuracy increased to 63% when 20 pronunciations were

[1] ibbly.com/Pseudo-words.html

included per word. Although performance for the majority of words peaked at a small number of included pronunciations, for 9 of the 28 words the most common pronunciation was ranked very low, causing overall performance on the 28 words to suffer in lexicons using only a limited number of L2S pronunciations (Figure 2). Hence, the total corpus benefited from an expansion of the lexicon to include more N-best pronunciations. The high risk of missing a key pronunciation commonly produced by users thus appears to outweigh the diluting effect of including greater pronunciation variety.

We also examined the extent to which performance could be enhanced by including L2S confidence scores for each pronunciation (Figure 3). Confidence scores [7] are used to weigh pronunciations based on their likelihood of being correct. For a benchmark comparison, we also evaluated the same corpus on a lexicon built using 1-best pronunciations manually created by an expert. Regardless of the number of pronunciations included, the expert lexicon performed better than an L2S lexicon with no confidence scoring, illustrating the disadvantage of poor pronunciations in the lexicon. However, the inclusion of L2S confidence scores produced a recognizer whose performance surpassed expert lexicon performance when the L2S model included at least ten-best pronunciations, illustrating some tangible benefit to including pronunciation variety on untrained words, particularly if confidence scores are available to down-weigh less likely pronunciation occurrences. In line with this notion, letter-to-sound confidence scores kept performance relatively steady even at the inclusion of a high number of potentially irrelevant pronunciations – a point at which lexicon performance without confidence scores had begun to drop.

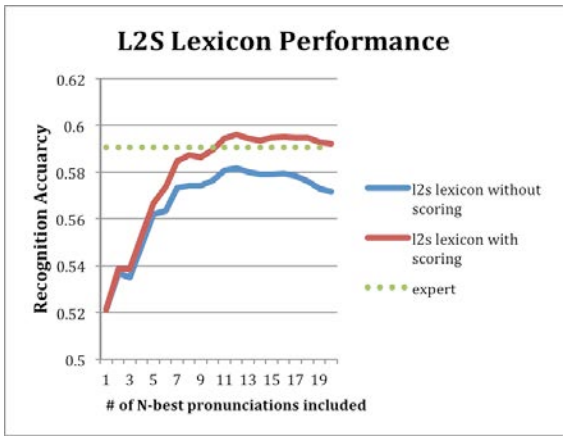


Figure 3: *Comparison of L2S performance with and without confidence scoring to an expert L2S.*

Although average recognition performance on a static 28-word recognizer was surprisingly low, recognition accuracy for the highest performing speaker was 94%, and it was above 85% for the top four speakers (Figure 4). As our user study was strictly a remote task, the remarkably wide spread among different speakers is partly due to substantial differences in microphone and hardware quality on different computers. To better understand the low average performance and high variance among speakers, we further categorized misrecognitions by false negative and false positive recognition errors. We found that the vast majority of errors were due to false negatives (85%), and only a small number were false positives (2%). The remaining errors (neither false positive nor false negative) were situations in which the learner produced the wrong utterance, but the recognizer hypothesized a third word that was neither the learner's utterance nor the target word.

Interestingly, the alarmingly high false negative rate was partially a function of in-game user behavior. Many users tended to repeat the same utterance multiple times upon experiencing a false negative error, in an attempt to resolve the recognizer's mistake. These repeated false negatives widened the performance gap between speakers because a single false negative error would almost always be exacerbated by an ensuing sequence of more false negative errors. This behavior may manifest particularly strongly in fast-paced game settings with short target utterances; the urgency associated with game incentives (i.e. Tetris blocks dropping) is complemented by the fact that one-word utterances are easy to repeat incessantly and thus worth the attempt. To discover the impact of repeated false negatives, we re-evaluated the corpus without false negatives that had been purely due to repetition, and found a 14% increase in overall recognition performance.

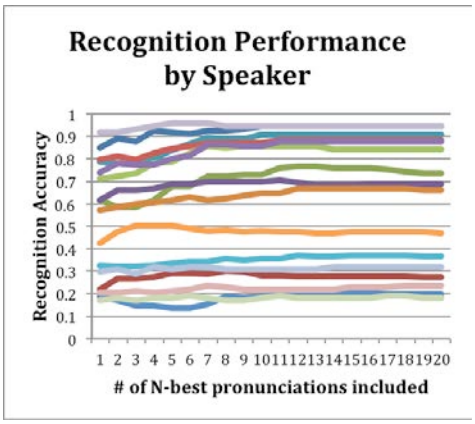


Figure 4: *Comparing recognizer performance across the 16 different speakers.*

False negative speech recognition errors also appeared to have an asymmetric impact on user enjoyment. In a post-study questionnaire on Mechanical Turk, some users reported that false negative errors inhibited their enjoyment of the game. For example, one user wrote that false negatives "made me less engaged, because I felt like [the game] was counting off for something I knew." On the other hand, false positive errors seemed to have a less detrimental effect on user enjoyment. Observations from local pilot testing revealed that false positive errors were more rare because users tended to speak only when they had some confidence or inkling of the correct answer. Moreover, because the target answer was revealed whenever the user succeeded, users often appeared amused rather than misdirected by the small number of false positives that they experienced.

The combination of time pressure and playful exploration inherent in gameplay may also have contributed to more anomalous utterances, which further increased the number of recognition errors. Anomalous utterances (Figure 5) accounted for 15% of the speech corpus and 10% of all recognition errors. For example, because we had changed the input method to be spring-loaded to optimize efficiency, some recordings

were partially cut-off due to the player releasing the record button prematurely. At other times, recordings were silent because the user hesitated to speak or accidentally pressed the record button. On occasion, game sounds such as row-completion ringing tones were also captured in the recording, even though they were designed not to overlap temporally with recorded speech. Furthermore, some users uttered nonsense phrases or English labels for the pictures, perhaps in a playful attempt to test the recognizer or in order to trigger the display of a hint, which is designed to appear once the user has attempted any utterance in a trial. More rarely, users conflated two vocabulary words and spoke a hybrid of two words.

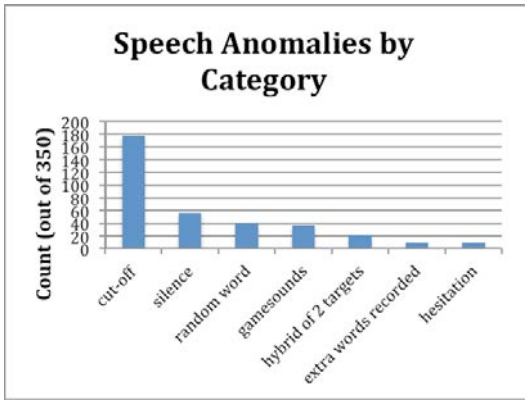


Figure 5: *Count of anomalous utterances by category.*

Overall, the most common anomalous cases were cut-off words and silent recordings (51% and 16% of anomalies, respectively). Cut-off recordings could be addressed by having the system constantly listen for speech and pad recorded utterances with extra time on both ends before sending them to the recognizer. Silent recordings could be better handled by incorporating silence into the recognizer's language model such that silence is a competing hypothesis in addition to the existing vocabulary words. In cases where the recognizer hypothesizes silence, the game interface can give feedback to the user to try again or speak louder. We leave these improvements for future work and instead focus on improving overall performance regardless of anomalies.

5. Strategies to Improve Performance

The disheartening effect of false negative recognition errors on user enjoyment suggests that relaxing the constraints of speech recognition to be more lenient could benefit engagement. The difficulties inherent in optimizing a letter-to-sound model for out-of-vocabulary words might also be alleviated by training lexicons on user-produced pronunciations mid-game that are detected to be likely correct. To this end, game-based constraints could be leveraged to provide strong contextual clues for maintaining high recognition accuracy in the face of greater leniency. To explore the viability of this approach, we identify several potential techniques for modifying the speech recognizer and re-evaluate the collected speech corpus on alternative recognizer configurations.

5.1. Dynamic vs. Static Vocabulary

Effective educational approaches tend to focus the learner's attention on only a few words or concepts at a time until their meanings have been internalized by the learner through repeated practice. In an intense and time-sensitive game setting, the gradual introduction of small sets of words is also critical for reducing the learner's cognitive load imposed by existing simultaneous interactions. Unlike typical speech interactions in which the set of possible user utterances may be large and uncertain, speech interactions amidst a learning game have implicit constraints that can be leveraged for enhancing speech recognition. Specifically, the game environment enables us to both constrain the recognizer vocabulary size and dynamically add additional words to the vocabulary as they are introduced to the learner. Constraining the vocabulary size can hopefully decrease the likelihood of false negative errors by preventing the recognizer from hypothesizing a word that the learner is unlikely to produce.

To determine the potential impact of this approach, we compare recognition accuracy between a static vocabulary of 28 words and a dynamic vocabulary (Figure 6), at varying numbers of pronunciations included in the lexicon. In the dynamic condition, we add a new word to the vocabulary only once it has appeared in the game, and constrain the maximum vocabulary size to only the words that the learner has seen within any particular game session (seven words maximum). The dynamic vocabulary demonstrated a 27% increase in accuracy over the static vocabulary when 10-best L2S pronunciations were included, and this benefit appeared fairly consistent across different numbers of N-best pronunciations included. The benefits were largely due to the substantial reduction in false negative errors, which were the source of most recognition errors.

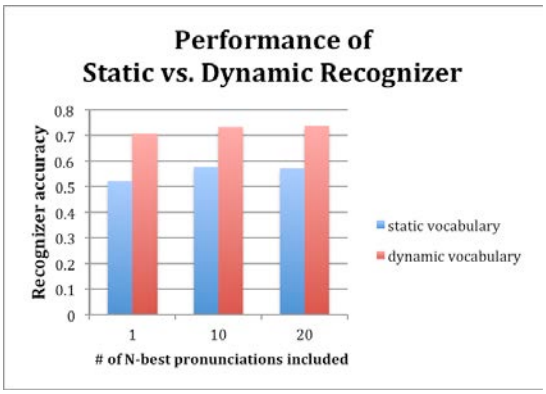


Figure 6: *Performance of static vs. dynamic recognizer at 1, 10, and 20 included pronunciations. The advantage of the dynamic recognizer remains fairly consistent across different numbers of N-best pronunciations.*

5.2. Deepening N-best Hypotheses

Game-based settings also provide strong contextual information about the target item on a trial-by-trial basis. Because the game keeps state of which target item is being presented to the user at every turn, a more lenient system could deem the learner correct if the target word appears in *any* of the top-N recognition hypotheses. This approach

assumes that the recognizer has some room for error and that, because the learner is likely to have spoken the target word, it is safer to check the top few hypotheses for the correct response before deeming the utterance incorrect. Figure 7 illustrates a substantial increase in overall word accuracy simply by expanding the N-best depth from one (59%) to four (73%), all with a static vocabulary of 28 words. In practice, even though recognition accuracy can be further boosted with more hypotheses accepted, it would be preferable to set a limit on this number so that the user does not assume that the recognizer will accept any response.

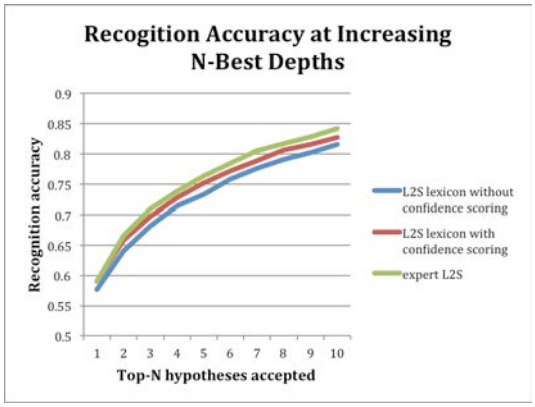


Figure 7: *Recognition performance, varying the number of N-best hypotheses accepted. Utterance is deemed correct when any top-N hypothesis matches the target word. Uses 10-best L2S pronunciations.*

A primary concern surrounding N-best depth expansion is the increased risk of false positive recognition errors. In the case of false positive errors, learners may mistakenly believe they have correctly recalled the word for a particular picture, with the consequence of strengthening an incorrect mapping. Hence, a trade-off may exist between decreasing frustration due to false negatives and increasing incorrectly learned mappings due to excessive leniency.

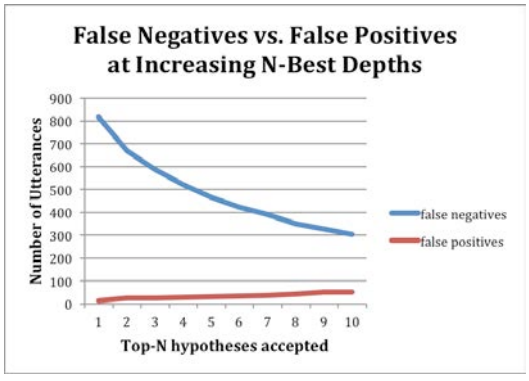


Figure 8: *Comparing the number of false negative and false positive utterances at an increasing number of N-best hypotheses accepted.*

To examine this potential trade-off, we measure the number of misrecognized utterances due to false negative and false positive errors at increasing N-best depths. Figure 8 shows that, as the number of accepted hypotheses increases, the number of false negative errors decreases dramatically, with only a minor increase in false positives. The significant decrease in false negatives is magnified by the elimination of repeated false negative errors due to learners re-attempting the same utterance after experiencing a false negative. Nevertheless, we find a very similar trend even after removing such repetitions from the dataset.

We further analyze false negatives and false positives among anomalous utterances, and find that anomalous recordings account for a substantial 80% of all false positive errors, compared to only 24% of all false negative errors. Because the majority of false positives are anomalies, and because a sizeable number of those are due to users producing random utterances, learners may find false positives more transparent and potentially less impenetrable than false negatives. In general, false positives are also less frustrating because they do not unfairly hinder the player's in-game progress. After a false positive, the player immediately focuses his or her attention on block rotation rather than being forced to re-attempt the utterance, making those experiences potentially more forgettable. These patterns lend support to the notion of adapting in-game speech recognition systems to be more lenient.

5.3. Training on high confidence user utterances

Lastly, out-of-vocabulary terminology can be detrimental to recognition accuracy and game enjoyment. Unlike acoustic and language models that learn the values of their parameters from training data, word pronunciations in a recognizer's lexicon are typically specified manually, often by an expert. Hence, a user wishing to review out-of-vocabulary words might encounter frequent recognition errors due to a letter-to-sound model that has been trained using only existing lexicons.

Recent work on pronunciation mixture models (PMM) has made it possible for experts to specify a set of pronunciations, but leave the weighting of these pronunciations to the PMM using speech data collected on the fly [11]. Yet, in a game-based learning context, it is unclear how unlabeled utterances can be used for training a PMM live, due to a chicken or egg problem of learners being unreliable agents for speaking the correct target item.

Nonetheless, we make a key insight that players are first introduced to the word-picture pair before the word is withheld for memorization practice. Because the learner sees both the word and cue on the first trial by way of introduction, the first utterance the player produces for any word has a high likelihood of being correct. In the Tetris game we have designed, the learner also hears the word pronounced out loud when it is first introduced, making it more likely that the learner will speak the target word correctly, particularly in the case of second language learning. On the other hand, first utterances may also be riskier for training since they could contain more anomalies such as hesitation and silence due to the user's unfamiliarity with the new item.

We thus evaluate speech recognition using pronunciations obtained by training a pronunciation mixture model solely on the user's first utterance of each word as a replacement lexicon (Figure 9). As a benchmark, we compare these results against lexicons produced using the letter-to-sound model. Because the test set for the PMM condition does not include any of a user's first utterances, we similarly remove all first utterances

when evaluating recognition on the normal letter-to-sound lexicons.

Remarkably, the PMM trained purely on the users' first utterances demonstrated a 3% improvement over the L2S lexicon (averaged over results from one to twenty pronunciations included), despite having no ground-truth labeling of any first-trial utterances. A PMM trained on other learners' first utterances produced no significant advantage over the L2S lexicons, suggesting that speaker-dependent characteristics may be critical to effective recognition.

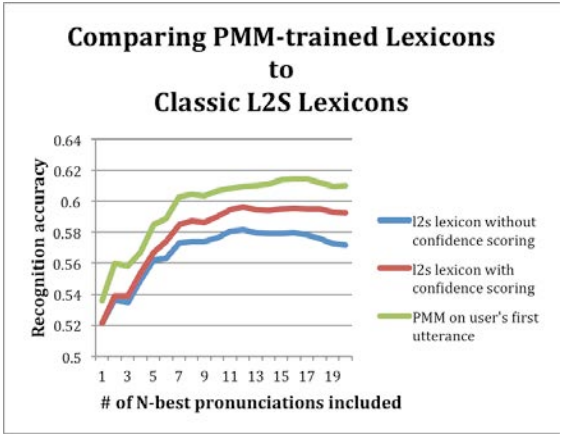


Figure 9: *Comparing performance of a PMM trained on the first utterance of each word to that of normal L2S lexicons.*

The promising speech recognition enhancement obtained by training only a small number of high confidence user utterances suggests further exploration of opportunities to perform user-specific PMM training using high confidence in-game scenarios. For example, starting a game in study mode before transitioning to retrieval mode could not only give the learner more time to develop familiarity with new items, but also offer an advantage for speech recognition enhancement. One could imagine collecting utterances during the study phase to produce a true mixture of multiple utterances produced by the same user for each word.

6. Conclusion

Our work has shown that a speech recognizer designed for traditional purposes may be unnecessarily strict when placed in a fast-paced game context, particularly because false negative recognition errors are both self-perpetuating and detrimental to learner enjoyment. We have proposed several techniques for improving performance, such as using a small and dynamic recognizer vocabulary, expanding the set of N-best accepted hypotheses, and using high confidence in-game utterances to retrain out-of-vocabulary words. Although a more lenient recognizer may run the risk of accepting learner errors, we found these occurrences to be surprisingly rare, and well worth the trade-off of decreasing the significant frustration associated with false negatives. It would be worthwhile to evaluate whether first utterances remain advantageous for PMM training in a second language learning context, despite learner inexperience in the target language.

While speech recognition has experienced limited adoption in fast-paced educational games compared to alternatives such as adventure style games, our results suggest that tailoring the recognizer to the unique needs of time-sensitive game environments could be key to increasing adoption. Future work should explore methods for handling speech anomalies specific to learning amidst rapid gameplay, such as using voice activity detection or time padding to prevent cut-off speech, and a silence model to handle accidental or hesitant recordings. Finally, automatic detection of words that are likely to be poorly ranked by the recognizer's letter-to-sound model, perhaps by comparing PMM scores to default L2S rankings of out-of-vocabulary items, would be a worthwhile venture for future research.

7. Acknowledgements

This research was funded by MIT Lincoln Laboratory. Special thanks to Ian McGraw for his valuable input and mentorship.

8. References

[1] Bisani, M. and Ney, H. (2008). Joint-sequence models for grapheme-to-phoneme conversion. Speech Communication 50(5), 434-451.

[2] Bisson, C. and Luckner, J. (1996). Fun in learning: The pedagogical role of fun in adventure education. Journal of Experiential Education 19(2), 108-12.

[3] Cai, C. (2013) Adapting arcade games for learning. Proc. CHI 2013, 2665-2670.

[4] Gee, J. (2003). What video games have to teach us about learning and literacy. Computers in Entertainment 1(1), 20-20.

[5] Glass, J. (2003) A probabilistic framework for segment-based speech recognition. Computer Speech and Language 17(2), 137-152.

[6] Gruenstein, A., McGraw, I., and Badr, I. (2008). The WAMI Toolkit for Developing, Deploying, and Evaluating Web-Accessible Multimodal Interfaces. Proc. ICMI, 141-148.

[7] Hazen, T. J., Seneff, S., and Polifroni, J. (2002). Recognition confidence scoring and its use in speech understanding systems. Computer Speech and Language, 16(1), 49-67.

[8] Kumar, A., Reddy P., Tewari, A., Agrawal, R., and Kam A. (2012). Improving literacy in developing countries using speech recognition-supported games on mobile devices. Proc. CHI 2012, 1149-1158.

[9] Malone, T. (1980). What makes things fun to learn? A study of intrinsically motivating computer games. Pipeline 6(2), 50-51.

[10] McGraw, I., Yoshimoto, B. and Seneff, S. (2009). Speech-enabled card games for incidental vocabulary acquisition in a foreign language. Speech Communication 51(10), 1006-1023.

[11] McGraw, I., Badr, I., and Glass, J. (2013). Learning Lexicons from Speech Using a Pronunciation Mixture Model. IEEE Transactions on Audio, Speech & Language Processing 21(2): 357-366.

[12] Nakamura, J. And Csikszentmihalyi, M. (2009). Flow theory and research. In C. R. Snyder & S. J. Lopez (Eds.), Handbook of positive psychology, 195-206.

[13] Seneff, S., Adler, M., Glass, J. Sherry, B., Hazen, T., Wang, C., & Wu, T. (2007). Exploiting Context Information in Spoken Dialogue Interaction with Mobile Devices. Proc. Intl Workshop on Improved Mobile User Experience, Toronto, Canada.

[14] Van der Spek, E.D. Wouters, P., & Van Oostendorp, H. (2009). Code Red: Triage. Or, Cognition-based Design Rules Enhancing Decisionmaking Training in a Game Environment. In Games and Virtual Worlds for Serious Applications, 2009. VS-GAMES'09. Conference in (pp. 166-169). IEEE.

[15] Xu, Y. and Seneff, S. (2012). Improving Nonnative Speech Understanding Using Context and N-Best Meaning Fusion. Proc. ICASSP, 4977-4980. IEEE.

[16] Xue, W. and Rudnicky, I. (2000). Language Modeling for Dialog System. Proc. ICSLP.

Spoken Grammar Practice in an ASR-based CALL system

Bart Penning de Vries, Stephen Bodnar, Catia Cucchiarini, Helmer Strik, Roeland van Hout[1]

[1]Centre for Language and Speech Technology, Radboud University Nijmegen, the Netherlands

{B.Penningdevries, S.Bodnar, C.Cucchiarini, W.Strik, R.vanHout}@let.ru.nl

Abstract

In this paper we present a computer assisted language learning (CALL) system that is developed to practice grammar in spoken language. To enable this, the system uses Automatic Speech Recognition (ASR) to process the L2 learner's responses. We investigate the possibility of providing corrective feedback (CF) on learner errors, and compare that with self-monitored language learning through output practice. In this paper we present the comparison of the two conditions 1) one group of learners received oral practice and CF on spoken performance, and 2) the other group received oral practice and no CF on spoken performance. We found that our system is successful for L2 speaking practice. The main finding is that both groups show learning after treatment. Between the groups, we did not find a learning difference, but the groups' sessions proceeded differently. Additionally we found that the CF group was more positive about the system than the NO CF group.

Index Terms: second language acquisition, corrective feedback, speech recognition, CALL

1. Introduction

A requirement for improving second language (L2) spoken proficiency is to practice speaking the language. For L2 learners, the opportunities for spoken practice are often not optimal: in language classrooms practice is limited, and the instruction is not tailored to individual needs.

CALL systems can create more suitable learning situations by offering individualized instruction, increased learner control, reduced anxiety and no time limits. Existing CALL systems, however, focus mainly on written language, whereas the benefits of CALL practice could also be applied to improve spoken practice. This requires use of automatic speech recognition (ASR). Currently, there are no systems that practice grammar in the spoken modality. For this reason we investigated the possibility of developing and testing such a system. In our study we investigate two versions of this system: spoken practice with and without automatic corrective feedback (CF). We present the results of an experiment on grammatical accuracy in oral production, and discuss the results.

2. Research background

2.1. Computer Assisted Language Learning (CALL)

Several studies have shown that CALL can be effective for language learning, and may outperform classroom instruction (e.g. [1]). The reasons for its effectiveness are availability, the amount of individual practice, and the learner's level of control over the learning experience. Practice with a computer is also found to result in reduced anxiety about making mistakes -which can be a restraining factor in classroom or face to face interaction- and this results in the learners producing more, and more varied, output.

When examining the CALL literature, we find that the majority of studies address written production [2]. However, the differences between written production and spoken production, for instance the increased cognitive load required for speaking and the involvement of the articulatory system, make it interesting to examine oral proficiency.

The rapid advances in the field of ASR technology make it possible to implement focused exercises for spoken production [3]. A review of CALL systems using ASR reveals that these generally address communicative skills or pronunciation, while no systems offer grammar practice in the oral modality [4]. Since developing grammar is important for improving spoken proficiency, there are reasons to develop systems that push learners to practice their grammar in the spoken modality.

2.2. Second Language Acquisition (SLA)

In our study we adopt the interactionist view on SLA, which assumes that interaction and the accompanying feedback serve an important function in developing the interlanguage (e.g. [5], [6]). This is in contrast with proponents of Universal Grammar-based theories (e.g. [7], [8]), who argue that exposure to input is sufficient for language acquisition. These opposing viewpoints have implications for language learning, particularly with respect to the role of output and CF. In designing the learning conditions in our CALL system, we took these issues into account.

2.2.1. Output

The role of output in L2 learning is subject to discussion, with most of the evidence pointing toward a positive effect of output [9] (but see [10] for an opposing view). A theoretical basis for the role of language production for SLA is outlined in the Output Hypothesis [6], which indicates three possible functions of output: 1) a noticing function, 2) a hypothesis testing function and 3) a metalinguistic function. By referring to skill-acquisition theory [11], [12] added the function of enhancing fluency through practice, thus stressing the importance of producing spoken output to improve speaking proficiency.

2.2.2. Corrective Feedback (CF)

Though the role of CF is still controversial, a considerable body of research has shown that CF has an effect on language learning (e.g. [13], [14]). Factors that are found to influence CF effectiveness are educational setting, type of CF [14], and learner differences [15].

In an overview of CF research [16] lists the conditions for CF to be effective for interlanguage development: CF needs to be systematic and consistent, clear enough to be perceived as CF, it

should allow for time and opportunity for self-repair and modified output, the intention of CF should be clear, and the learner should be ready for the feedback. This overlaps with the objections that [17] raises against grammar correction in oral practice, who claims that it is impossible to achieve the ideal characteristics of individualized CF in classroom environments and therefore argues for abandoning grammar correction altogether. He admits at the same time that "the possibility remains that some untested combinations of these variables could produce successful feedback, while avoiding (or minimizing) the accompanying problems." (451)

2.3. Research Questions

Since grammatical accuracy is an important part of proficiency, we designed a system that provides learners with the opportunity to practice and internalize grammar rules and possibly receive CF based on ASR. The demands on our ASR are high, since it has to parse non-native speech accurately and provide accurate CF. Our design takes the limitations of ASR into account, but we need to test whether the system is effective in improving learners' proficiency. By implementing two learning conditions, one based on practicing with spoken output and opportunities for self-correction, and the other based on practicing with spoken output and automatic CF, we can test the working of the system and compare its performance in two learning conditions. Thus we address the following research questions:

- Is it possible to develop an ASR-based CALL system that can detect grammatical errors in spoken performance and provide appropriate corrective feedback?
- Is there a difference of effect between practicing with spoken output and self-monitoring (NO CF group), and practicing with spoken output and automatic CF (CF group)?

3. Method

This section describes our ASR-based CALL system that offers grammar practice in spoken production. The learning exercise we developed is embedded in an experiment to measure learning outcomes by applying pre- and post-tests, and learner appreciation through a post questionnaire.

3.1. The GREET system architecture

In Figure 1, we give a schematic overview of the practice system. The learner interacts with the system through the GUI (a screenshot is given in 3.5, Figure 2), and is presented with a task from the courseware database. The learner is given a question and 'word blocks' to construct a sentence. This restricts the number of possible responses and contributes to higher ASR accuracy.

For each question in the exercise, the language model contains all answer possibilities (i.e. all possible sentences) that can be created with the word blocks. These are tagged with meta-information, which in the current experiment is simply whether that answer-sentence is grammatically correct or incorrect. In later experiments, this setup can include more detailed messages.

When the learner records an utterance, the speech recognizer outputs a recognition result and a confidence level. If the confidence level is below a preset threshold, the system assumes that the learner did not record a valid attempt: it does not try to detect errors, but instead the learner is asked to re-record. Otherwise the recognizer maps the utterances to a sentence in the language model, and error detection sends to the courseware engine the appropriate message regarding the sentence's grammaticality (for more detail see [18]). The final step is the presentation of a feedback message to the learner. Taking the input from error detection, the definition in the courseware engine determines what type of feedback is presented, and how it is presented (see section 3.5.1).

The experiment is run through a website and the learner interacts with the system through a web browser. All recognition is performed on the web server. This allows us to run more experiments at the same time and at different locations, and to store all data centrally.

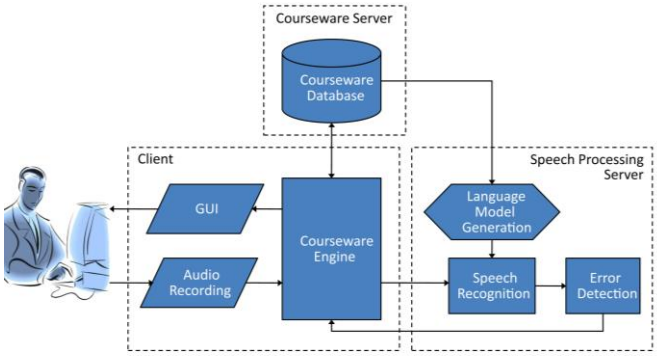


Figure.1. A schematic overview of the GREET system

In our experiment the system behaves differently in the two conditions. For both groups, the speech is processed through the recognizer, and the recognition result is logged. However, for one group the answer is explicitly evaluated on the screen, and they can advance or retry according to the system; the other group receives no information on the recognition result, and can advance or retry according to their own preference.

Throughout the experiment, our system logs the learner-system interactions: this allows us to look in detail at learner behavior, and inspect the logs for irregular behavior (cf. [19]). Relevant for the current paper are the interactions with respect to questions, the number of attempts, and when a correct response was given after an error (i.e. repair).

3.2. Procedure

Participants signed up for two sessions of one and a half hours, at times of their choosing. In the first session, they completed a pre-test questionnaire on personal data, two pre-tests (described in section 3.5), and a treatment session (section 3.4). Within a week of the first test day they completed the second session: the second part of the treatment, two post-tests, and finished with a post-test questionnaire (section 3.6).

The experiment was run through a website. Participants logged in and the website gave them their task in a step-by-step fashion. Each time they completed a task, a new one was shown. Before the proficiency tests, instructions were shown in a slide show webpage, and the participants did three practice questions before moving on to the real test.

3.3. Participants

We recruited 29 adult participants from Dutch language courses at A2 and B1 (CEF level) at the Radboud University Nijmegen in the Netherlands. They were offered 15 Euros for participating. The participants were randomly assigned to the CF group, or to the NO CF group. Random assignment and drop-outs resulted in an uneven division in groups: 12 participants in the control group (6 male, 6 female), and seventeen in the experiment group (5 male, 12 female).

In total, there were 13 different L1s in our sample: Arabic, Chinese, Dari, English, French, German, Indonesian, Italian, Russian, Luganda, Polish, Romanian, Spanish (NO CF group: 7 L1s, CF group: 11 L1s). The participants' education was all over eight years of formal education after primary school, with most at university level. Their mean age was 31 years with a range of 22 to 48 years old.

3.4. Target structure

The syntactic feature under investigation was inversion as a result of Dutch verb second (V2). Dutch V2 means that the finite verb appears in the position following the first constituent in the main clause. In Dutch, the subject precedes the verb. Inversion of this order occurs when the first constituent of the sentence is not the subject: the V2 principle requires the verb to remain in second position, thereby forcing the subject to take the position following the verb.

a. Subject-initial main clause

Melvin	koopt	morgen	bloemen
S	V	A	O
Melvin	buys:3SG	tomorrow	flowers

'Melvin buys flowers tomorrow'

b. Inversion clause

Morgen	koopt	Melvin	bloemen
A	V	S	O
Tomorrow	buys:3SG	Melvin	flowers

'Tomorrow Melvin buys flowers'

The acquisition of inversion is problematic for L2 learners of Dutch [20]. A reason for this may be that violation of inversion does not necessarily affect meaning. This makes inversion an appropriate feature to study the effect of CF, as CF is likely to benefit errors that do not affect meaning, do not typically lead to communication breakdown, or that lack a clear form-meaning relationship [21].

3.5. Treatment design

Participants practiced implicitly with the target structure. After watching a short (approximately 35s) clip of an ongoing story, a 'teacher' on screen asked them questions about the content. To answer, the participants had to construct a sentence using 'word-blocks': parts of a sentence that need to be combined to form one sentence. The participants recorded their answer by speaking into the microphone.

The treatment was spread over two sessions of 45 minutes each. In total there were 120 questions (63 questions in session 1, 57 questions in session 2), of which 37 were target questions (19 in session 1, 18 in session 2). If the participant completed all the questions within 45 minutes, they went back to start at the beginning, to control for time-on-task.

Figure. 2. screenshot of the GREET treatment exercise

3.5.1. CF in treatment

In a meta-analysis of studies comparing CF types [14] conclude that prompt feedback is the most effective, since it allows and triggers self-repair. This is in line with skill-learning theory, because uptake following prompts and containing repair enables learners to strengthen their control over linguistic forms that they have only partially acquired [22]. As a result, for the current experiment we used the prompt as the type of CF.

If the learner's response was correct, the learner saw a green checkmark pop-up, and was advanced to the next question. If the learner's response was incorrect, (s)he saw a red screen saying "That is incorrect. Try again", and one word-block was put in the correct place to serve as a hint. If the ASR could not confidently process the learner's response, (s)he saw a neutral light-grey screen, containing a message saying "I'm sorry, I could not understand. Please try again".

For the NO CF group, the message is always the same: a white screen saying "Your answer has been saved. You can only save one answer. Do you want to keep this one and move to the next question, or try again?" The learner then had the choice to progress, or retry.

3.6. Proficiency tests

Two proficiency tests were selected to measure knowledge of the target grammatical feature (accuracy). We selected two tests for cross-task comparison and validation [23]: a timed grammaticality judgment task (GJT) and a discourse completion test (DCT). The tests were selected based on the psychometric study by [24]. The tests are distinct in that the one is a receptive reading task, and the other a spoken production task, but at the same time they are complementary because they measure the same aspect of language competence.

3.6.1. Grammaticality judgment task (GJT)

In the GJT participants judged 40 sentences within a time limit, set at 12 seconds based on pilot versions. The test had an equal number of target and filler sentences, and grammatical and ungrammatical sentences were also equally distributed.

Pre- and post-test versions were counterbalanced, and the order of item presentation was randomized per subject. Correct

judgments scored 1, incorrect judgments scored 0. A response outside the time limit was scored as incorrect.

3.6.2. Discourse completion task (DCT)

The discourse completion task (DCT) elicits oral production. The target structure under investigation is easily avoided in Dutch, so we restricted the possibilities for output, and modified the design to make inversion obligatory in target sentences (see [25] for a similar task design for written production). Participants saw the beginning of a sentence which they were required to complete. To establish some context for the task, they were given a lead-in sentence, one or two hint words, and a picture. To answer, the participant pressed the record button and spoke a full sentence. In this task, there was a time limit of 30 seconds.

The test was counterbalanced, and the order of item presentation was randomized for each participant. Each test version was made up of 32 items of which half were targets, and half were fillers. The recordings were transcribed and scored for correct use of inversion.

3.7. Post-test questionnaire

After the test, the participants were asked their opinion of the system. Subjects indicated on a five-point Likert scale whether they agreed or disagreed with statements about the experiment system. Questions concerned if they felt that their level of Dutch had improved, their self-confidence to speak Dutch had grown, whether it was a good system for learning Dutch, and if they had enjoyed using it. They were also asked open questions to elicit opinions and suggestions about the program and CF.

4. Results

4.1. System performance

In a pilot study we found the feedback accuracy rate of our system to be 96% (accepts) and 97% (rejects) [26]. In this experiment, we examine several indicators such as practice logs and questionnaire responses to see if the CF provided by the ASR was as expected.

Analysis of individual sessions revealed that for two subjects several questions elicited more attempts than normal. They were receiving many 'I don't understand' feedback messages, which does not allow you to advance. When inspecting these cases, we found that for these subjects the sound recording was inferior, which was caused by either the wireless connection, or hardware sound recording issues. Some other subjects also experienced network related sound problems, but for them it did not affect CF accuracy. There seemed to be a particular combination of accent, sentence content, and a female voice with the sound problem that resulted in the ASR being unable to confidently process the response. Overall, we did not see evidence of proficiency level as measured by the pre-test (or CEF level), or L1 background on system behavior.

In total, 23 participants were able to complete the full experiment without assistance from the experimenter, which meant they could practice with the system individually. The only issues that caused disruptions for the other six were some wireless network related problems, learner errors using the recording interface or microphone, and one server crash. These problems are unfortunate, but can largely be avoided in future experiments. However, though unrelated to the ASR technology, they affected the learner's experience with the system.

4.2. Proficiency tests

As a first step in our analysis we inspected the item scores in both proficiency tests. The internal reliability scores (Cronbach's alpha) of the GJT was 0.86, and for the DCT 0.95. The two tests show a high correlation (pre- plus post-test: r=.781, p=.000). However, they are both qualitatively different and measure competence in a different modality, so we take the tests to be complementary in providing us a better picture of the learner's language level. As a result, we present the test results separately and combined.

Table 1 shows the data from the pre- and post-tests. The CF group received immediate CF after each utterance; the NO CF group had the possibility to re-record their answer after each utterance.

Group	DCT Pre		DCT Post		GJT Pre		GJT Post	
	M	SD	M	SD	M	SD	M	SD
CF group	.64	.35	.68	.36	.64	.22	.70	.19
NOCF group	.71	.27	.77	.31	.66	.22	.71	.20

Table 1. Mean and Standard Deviation of proficiency tests

In Table 1 both groups seem to improve. There is no difference immediately visible between the two groups. We turn to an analysis of variance to see if there was a difference of interaction of groups with the treatment. The filler data of the GJT is also included to see if non-target items were learned.

	DCT (targets)	GJT (targets)	GJT (fillers)	DCT&GJT combined
Treatment	F(1,27) =3.856 p=.060	F(1,27) =5.209 p=.031	F(1,27) =1.501 p=.231	F(1,27) =8.512 p=.007
Interaction Group, Treatment	F(1,27) =.298, p=.589	F(1,27) =.01, p=.947	F(1,27) =.03, p=.564	F(1,27) =.146 p=.706
Group	F=.468, p=.500	F=.047, p=.830	F=.036, p=.851	F=.284, p=.598

Table 2. ANOVA results of the proficiency tests. It shows scores on targets and on the fillers items for the GJT, for comparison.

In Table 2 we see that treatment had a significant effect $p<.05$ in the GJT test, and for the DCT and GJT combined, and a $p<.1$ for the DCT. We did not find an effect for the fillers of the GJT, which shows that the subjects improved only on the target sentences, and have improved their accuracy of Dutch V2 as a result of treatment.

There is no evidence of an interaction of group with treatment. This indicates that the CF group did not have an additional learning effect as a result of prompt feedback.

4.3. Log data

The system logs interactions with the participant, which is inspected for information on learner behavior during the test. First we excluded ceiling participants, because they have already mastered the rule, and the log data cannot show us learning behavior among these participants. We excluded participants who scored higher than .93 on the combined pre-tests. Then we

proceed to look at the difference in learning behavior between the CF and the NO CF group.

The logs show us the differences in the treatment for the participants. In the bar graph in Figure 3, the data for the two sessions are shown. Figure 3 shows the number of questions practiced (Q), and the number of attempts (A), where a learner can have several attempts per question. When a learner attempts a question again after making an error, there is chance for successful repair of the error. The number of times this happens is the number of repairs (R).

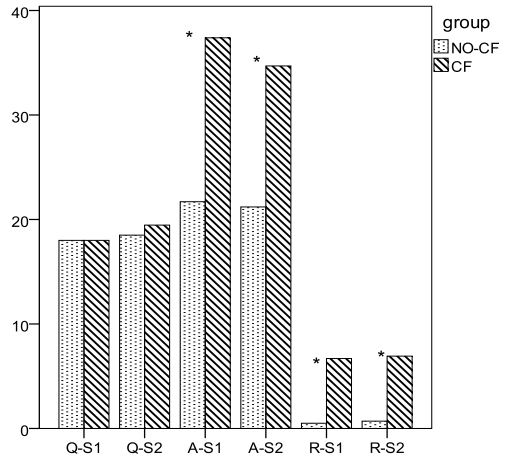


Figure 3. Bar graph of GREET treatment sessions. It shows the number of questions practiced (Q), attempts (A), and repair(R). An asterisk '' indicates a p<.05 significant difference*

The CF group had significantly more attempts. When comparing the NO CF and the CF groups, we found that there is a significant difference for the number of attempts in session 1 (t (df=27) = -4.285, p=.000), and for session 2 (t (df=27) = -3.027, p=.006). The CF group also had significantly more counts of successful repair in both sessions: session 1: t (df=27) = -5.496, p=.000, session 2: (t (df=27) = -4.666, p=.000. (If we take a strict count and exclude the problem instances mentioned in (4.1), we find that in session 1 for 86% of the attempts, and in session 2, 99% of the attempts, the system is performing as expected. If we take these numbers, there is still a significant difference between the CF and the NO CF group with respect to Attempts and Repair). Interesting to note is the almost identical number of questions for both groups, where we expected that the NO CF group could practice more questions, since they are not slowed by the CF.

4.4. Post-test questionnaire

In a post-test questionnaire we asked the participants about their practice session in 17 questions. Participants responded using a 5-point Likert scale (1= very negative, 5=very positive), from which we calculate their attitude towards the system. The data is given in Table 3 below.

The mean for both groups is above three, which shows that the participants were moderately positive about using the system (3 = neutral). The CF group is more positive about the system: they show a higher appreciation score (t (df=26) = 2.193, p=0.037).

The participants were also asked open questions. Analysis of these questions revealed that some participants in the CF group indicated a wish to receive more specific CF, or to receive CF on pronunciation. In the NO CF group, most participants indicated they would like to receive feedback, though they did not specify what kind of feedback they would prefer.

Group	Mean	N	Std. Deviation	Minimum	Maximum
NO CF	3.34	12	.347	2.50	3.79
CF	3.68	16	.445	2.94	4.31
Total	3.53	28	.434	2.50	4.31

Table 3. Descriptive statistics of the post-test questionnaire

5. Discussion

We found that both the CF group and the NO CF group show improved accuracy after practice with GREET. We did not observe an additional effect of providing automatic CF. Though there is a larger number of utterances (attempts) and repairs found in logs for the CF group, we did not find a higher overall improvement. An explanation for this may be that the NO CF group had to pay more attention to their utterance, because they had to self-monitor their answer, making the task was more cognitively demanding. Additionally, we may assume an effect of structured input, since the logs show that both groups received an equivalent amount of input (cf. [25]).

Output practice without CF seems to be effective only when the learner has prior knowledge of the grammatical structure practiced in the treatment. A closer investigation of pre-test proficiency level indicates that NO CF group participants with a low entry level were unable to improve their accuracy on the target structure. This suggests that CF might be necessary at lower levels of proficiency [cf 27].

It is important to note that the NO CF group was told that their answers were recorded and evaluated by the system, and they would receive their score after the experiment. In a pilot experiment where we did not promise a score, the learners indicated that they felt they were not practicing effectively. This leads us to the conclusion that a valuable contribution of ASR in our system is that learners feel that somebody (i.e. the CALL system or an interlocutor) is listening to (the accuracy of) their utterances. The learner seems to need a sense of interaction to produce output that is beneficial for L2 development.

In addition, we found that the learners indicated a preference for receiving CF from the system. This may tie in with the idea that the learner appreciates the feeling of interaction: the immediate response of the system on their utterance. Relevant may also be the role of the positive feedback, in the form of a green check mark.

It is also likely that the (positive and negative) CF had an effect on learner confidence. They may feel more confident that they are learning, because they can only proceed when they provide a correct Dutch sentence. This is also an important aspect of the learning system, because it may prompt learners to use their language more often, and thus practice more.

Though the GREET system worked well overall, we found that there were some improvements possible for the experiment design and the CF design. With respect to experiment setting, a reliable internet connection is required; and we intend to include a microphone test for the learner to check if his/her recording sounds proper before commencing the experiment. With respect

to CF design, it seems that the check for the valid utterance is set too strict for the ASR (confidence level threshold), resulting in too many 'I cannot understand' messages. This may confuse the CF reception by the learner. A reason why we can lower the threshold is because learners are generally found to be motivated and cooperative and trying to record a valid utterance.

6. Conclusions

To improve L2 learning possibilities, we developed an ASR-based CALL system that offers spoken grammar practice. The system worked successfully. L2 learners who practiced their spoken grammar by using the system improved their grammatical accuracy. We obtained learning gains for a group of 29 participants practicing for 90 minutes with our system. In the group that received CF, the ASR component of our system successfully interacted with learners with eleven different L1s. Moreover, we found that the group receiving CF evaluated the system more positively than the group that did not receive CF. The learners of the CF group were positive that the system was a good way to learn Dutch, and that their Dutch had improved as a result of working with the system. This suggests that our current setup is a good way to practice grammar for oral proficiency.

In developing this system, we succeeded in creating learning situations in which we can monitor the learner's behavior during practice. Additionally, the system interacted with the learner by providing CF on spoken grammar.

The system we have developed lends itself well to further language learning research, as we can continue to run experiments under various learning conditions and with an increasing number of participants. Experiments with more learners will eventually shed light on the role and impact of individual learner differences.

7. Acknowledgements

We would like to thank our colleague Joost van Doremalen for developing the ASR component of the CALL system used in this experiment. This work is part of the research program `Feedback and the acquisition of syntax in oral proficiency' (FASOP), which is funded by the Netherlands Organisation for Scientific Research (NWO).

8. References

[1] Torlakovic, E. and Deugo, D. "Application of a CALL system in the acquisition of adverbs in English". Computer Assisted Language Learning, 17 (2):203–235, 2004.

[2] Gamper, J. and Knapp, J. A review of intelligent CALL systems. Computer Assisted Language Learning, 15(4):329–342, 2002.

[3] Clifford, R. and Granoien, N. "Applications of Technology to Language Acquisition Processes: What can Work and Why" In Holland, M., Fisher, P. [Eds], The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Practice. Routledge : New York. 2008.

[4] Bodnar, S., Cucchiarini, C., Strik, H. "Computer-assisted Grammar Practice for Oral Communication". Proceedings of the International Conference on Computer Supported Education..355-361, 2011.

[5] Long, M. "The Role of the Linguistic Environment in Second Language Acquisition". In Ritchie, W. and Bhatia, T. [Eds], Handbook of Second Language Acquisition, 413-68. San Diego: Academic Press, 1996.

[6] Swain, M. "Communicative competence: some roles of comprehensible input and comprehensible output in its development", In Gass, M., Madden, C. Rowley [Eds] Input in Second Language Acquisition. Newbury House, 235-53, 1985.

[7] Krashen, S. Principles and practice in second language acquisition. Oxford: Pergamon, 1982.

[8] Schwartz, B. "On explicit and negative evidence effecting and affecting competence and 'linguistic behavior'". Studies in Second Language Acquisition (15), 147–63, 1993.

[9] Muranoi, H. "Output practice in the L2 classroom". In DeKeyser, R. [Eds], Practice in second language learning: Perspectives from applied linguistics and cognitive psychology. Cambridge, UK: Cambridge UP. 51-84, 2007.

[10] Krashen, S. "Comprehensible Output.", System (26), 175-82, 1998.

[11] Anderson, J., Fincham, J., "Acquisition of procedural skills from examples". Journal of Experimental Psychology: Learning, Memory, and Cognition (20): 1322-40, 1994.

[12] De Bot, K., "The Psycholinguisitics of the Output Hypothesis", Language Learning (46) ,529-55, 1996.

[13] Norris, J., Ortega, L., "Effectiveness of L2 instruction: A research synthesis and Quantative Meta-analysis", Language Learning(50), 417-528, 2000.

[14] Lyster, R., Saito, K. "Oral Feedback in Classroom SLA: A Meta-Analysis", Studies in Second Language Acquisition (32). 265–302, 2010.

[15] Dörnyei, Z. The Psychology of the Language Learner: Individual Differences in Second Language Acquisition. Lawrence Erlbaum Associates, Publishers: New Jersey, 2005.

[16] El Tatawi, M., "Corrective feedback in second language acquisition", Working papers in TESOL and Applied Linguistics (2), 1-19, 2002.

[17] Truscott, J. "What's wrong with oral grammar correction". Canadian Modern Language Review (55), 437-56,1999.

[18] Strik, H., Cornillie, F., Colpaert, J., van Doremalen, J., Cucchiarini, C. "Developing a CALL System for Practicing Oral Proficiency: How to Design for Speech Technology, Pedagogy and Learners.", Proceedings of SLaTE workshop, Warwickshire, England, 2009.

[19] Collentine, J. "Insights into the Construction of Grammatical Knowledge Provided by User- Behavior Tracking Technologies." Language Learning Technology(3), 44-57, 2000.

[20] Jordens, P. "The acquisition of word order in Dutch and German as L1 and L2". Second Language Research (4), 41-65, 1988.

[21] Sauro, S. "Computer-Mediated Corrective Feedback and the Development of L2 Grammar". Language Learning and Technology (13). 96-120, 2009.

[22] Ellis, R. "Epilogue: A Framework for Investigating Oral and Written Corrective Feedback". Studies in Second Language Acquisition (32), 335–49, 2010.

[23] Norris, J., and Ortega, L. "Defining and measuring SLA. In C. Doughty and M. Long [Eds] The handbook of second language acquisition. Malden : Blackwell. 717-61, 2003.

[24] Ellis, R. "Measuring Implicit and Explicit Knowledge of a Second Language: A Psychometric Study". Studies in Second Language Acquisition (27). 141-72, 2005.

[25] Sanz, C., Morgan-Short, K. "Positive Evidence Versus Explicit Rule Presentation and Explicit Negative Feedback: A Computer-Assisted Study". Language Learning(54), 35–78, 2004.

[26] Bodnar, S., Penning de Vries, B., Cucchiarini C., Strik, H., van Hout, R. "Feedback in an ASR-based CALL system for L2 syntax: A feasibility study". Proceedings of SLaTE workshop, Venice, Italy, 2011.

[27] Ammar, A., Spada, N. "One size fits all? Recasts, Prompts, and L2 Learning. Studies in Second Language Acquisition (28). 543-74, 2006.

Learners' situated motivation in oral grammar practice with an ASR-enabled CALL system

Stephen Bodnar, Bart Penning de Vries, Catia Cucchiarini, Helmer Strik, Roeland van Hout[1]

[1]Centre for Language and Speech Technology, Radboud University Nijmegen, the Netherlands
{B.Penningdevries, S.Bodnar, C.Cucchiarini, W.Strik, R.vanHout}@let.ru.nl

Abstract

Advances in speech recognition (ASR) technology have resulted in computer applications that provide compelling forms of speaking practice to learners of a second language (L2). Evaluation of such applications typically does not include an analysis of how learners' situated motivational states fluctuate during language practice. In connection with recent developments in L2 motivation theory, this paper investigates situated learner motivation in practice with an ASR-enabled system. An experiment was conducted in which our system provided oral grammar practice for Dutch L2 under two learning conditions: with and without corrective feedback. We report on learners' motivational experiences by triangulating an analysis of 1) motivational trajectories from a periodic motivation questionnaire, 2) post-practice reflective questionnaires, and 3) behavioural log data recorded by the system during practice. Our analysis shows that learners maintained positive attitudes towards the system throughout practice and became increasingly confident over time.

Index Terms: CALL, motivation, ASR, corrective feedback, grammar

1. Introduction

Advances in automatic speech recognition (ASR) technology have resulted in computer applications that provide interactive speaking practice to learners of a second language (L2). Different types of interaction are possible, including simulated conversational exchanges and pronunciation training. Typically, evaluation of such applications has concentrated on the performance of the technology, or the impact on L2 proficiency, rather than on learner motivation. When motivation has been included, it has typically been addressed through pre- and post-training evaluation questionnaires. Thus, little is known about how learners' situated motivational state fluctuates during practice with ASR-enabled systems.

Connected to this gap in the research is work by [1] which advocates viewing motivation as a learner state that emerges from a series of interactions between learner and context. Furthermore, it has recently been suggested to complement subjective methods with objective records of events that occur during practice [2, p.68].

In this paper we investigate situated learner motivation in practice with an ASR-enabled system that provides oral grammar practice for L2 learners of Dutch. The system can be configured in different ways to allow different forms of learner-system interaction. For instance, the system elicits spoken output from the learners, which later can be evaluated through ASR to provide an overall score. Another configuration option is to employ ASR to provide immediate corrective feedback on each utterance. This makes it possible to investigate situated motivation under different learning conditions. An important issue in second language acquisition (SLA) research is the effect of corrective feedback on L2 development. Corrective feedback can affect learner proficiency, but is also related to learner motivation [3]. Our system employs a logging module that records learner and system behaviour with temporal precision and consistency; in our view, this practice environment presents compelling opportunities to study situated motivation in a way that closely corresponds with current views in the L2 motivation literature. Recently, we conducted an experiment with our system which aimed at investigating situated motivation under different learning conditions, specifically with and without immediate corrective feedback. We report on the motivational experiences of the participants by triangulating an analysis of 1) motivational trajectories from a periodic motivation questionnaire, 2) post-practice reflective questionnaires, and 3) behavioural log data recorded during practice.

2. Situated motivation and Corrective Feedback in L2 learning

In SLA research, motivation is a multi-faceted concept. Early views of motivation took a social-psychological approach where research emphasised the role of learners' reasons or motivations for learning an L2 [4]. Though considered pioneering for the inclusion of social context in the study of motivation [5, p. 67], later research would criticise this macro perspective, which dominated the study of L2 motivation for many years. A criticism particularly relevant here was that the approach did not

allocate much of a role to motivation in the classroom or other learning contexts [6], in contrast with the views of L2 educators and educational psychology researchers.

More recently, situated perspectives that involve 'a more fine-tuned and situated analysis of motivation as it operates in actual learning situations' [5, p.74] have begun to receive attention in the literature. Particularly relevant here are the views of Ushioda, who has described a person-in-context approach to studying situated L2 motivation [2]. Important in this view is the acknowledgement that motivation in a situated context can change over time and that motivation should be studied in relation to interactions or events that take place during language practice. Thus, Ushioda [2] can be seen as calling for fine-grained objective practice data to complement subjective data gathered in situated motivation studies.

In our view, this approach to motivation is well-suited for research into language learning in a computer environment. The reason is that CALL applications can support a variety of learner-system interactions. A second reason is that they have the capability to log in great detail these interactions. Despite these capabilities, few studies have researched situated motivation in a CALL environment. Notable exceptions are work by [7], who studied the effect of a personalisation strategy in an L2 vocabulary tutor, [8], who used computer logs to study motivation in an online language course, and [9], who investigated the automatic detection of learners' motivational state from computer log files.

As a step towards employing computer capabilities to study situated motivation in a CALL environment, we built and tested a new motivation component in our CALL system. In an experiment with Dutch L2 learners, we tested the use of our new component in the context of an issue that has received considerable attention in the recent literature: the role of corrective feedback in L2 learning. A considerable body of literature has indicated that one of the problems in research on corrective feedback is the impossibility of studying this phenomenon under tightly controlled conditions [10]. Studying learner motivation in relation to corrective feedback is particularly relevant as various researchers have pointed to a possible demotivating effect [11, 12, 13, 14]. Our system makes it possible to create conditions in which corrective feedback is provided instantaneously, systematically and intensively, with opportunities for self-repair on the part of the learner. The new motivation component, together with the logging capabilities of our system, provide very large quantities of subjective and objective data that contribute to our understanding of motivation and corrective feedback in L2 learning.

3. Materials and Methods

This section presents materials used in the experiment: an ASR-based CALL system for practicing grammar in Dutch L2, situated mini-motivation questionnaires, computer practice logs, and a post-practice questionnaire gathering participants' impressions of the training. A description of the experimental procedure is also included.

3.1. An ASR-based CALL system for Dutch L2 Oral Grammar Practice: GREET

We use ASR in a system that helps learners practice aspects of Dutch word order. A common difficulty encountered by Dutch learners is inverting the position of the subject and verb when required. Our system provides training for this aspect of Dutch with a collection of video clips and question and answer exercises. Learners first watch a video clip and are then quizzed on the contents of the clip by the system. In the quiz, they respond to questions by recording themselves speaking their answer aloud. For experimental purposes, we have built two versions of the system. In the corrective feedback system (CF), ASR is used to provide immediate corrective feedback. The feedback was designed to be of a prompt type (for a detailed description see [15]). First, the system notified the learner that it detected an error by displaying a message, formulated as 'That was incorrect. Please try again'. Second, the system provided the learner with a hint by incrementally revealing the correct word sequence. In the no-feedback system (NOCF), the system displayed a neutral message notifying the learner that their recording has been saved. Prior to beginning practice, the NOCF group received a message stating that their scores would be evaluated at a later time. All other aspects of practice were identical.

3.2. Mini-Motivation Questionnaires

A key component we added for this experiment is a 3-item 'mini-motivation' questionnaire we use to track motivational changes. A screenshot of the questionnaire is shown in Figure 1. The questionnaire consists of three semantic differential scales designed to survey *attitude*, learners' general attitudes towards practice with the system, *motivation*, learners' motivation level as the desire to continue practice with the system or stop practice and do something else, and *self-confidence*, as the learner-estimated level of difficulty of future practice with GREET. To aid learners' understanding of the questionnaire items, English translations of the texts are accessible from an on-screen hyperlink.

3.3. Computer practice logs

The GREET system maintains a detailed log of events that occur during practice. The events recorded by the system include page views, number of video clips viewed, number of questions viewed, time on different types of pages, number of recordings, ASR recognition results, type of feedback returned, and others. When a

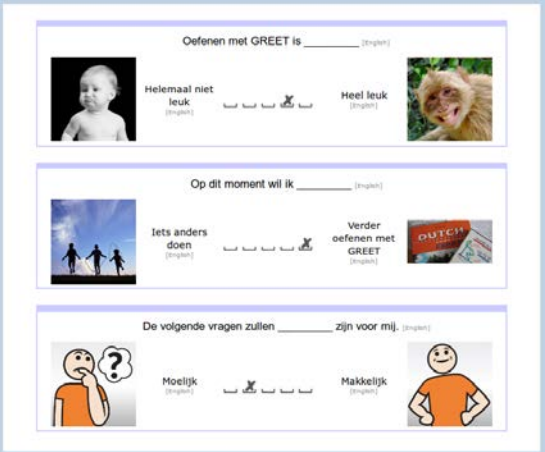


Figure 1: The new mini motivation component in our system. Question 1 (top): Practising with GREET is . . . (Not very nice - Very nice). Question 2: At this moment I want to . . . (Do something else - Continue practising with GREET). Question 3: The next questions will be . . . for me (Difficult - Easy). Learners indicate their opinion by marking one of the empty boxes with their mouse.

learner begins an activity, the system creates a practice session object to store the events that occur. In later analyses these sessions serve as records of the interactions that took place during practice. In this experiment, we focus on the number of video clips viewed, the number of questions completed, and the number of attempts made at answering a question.

3.4. Post-practice questionnaire

We used a post-practice questionnaire to obtain learners' subjective evaluations of the system. The questionnaire consisted of seven Likert items (using a 5-point scale):

- One item asked learners to rate the efficacy of training with the system.
- One item asked for learners' opinions on whether they felt their Dutch had improved as a result of practice.
- Three items asked learners to rate the video clips, question exercises and the system as a whole.
- Two items evaluated learners' difficulty in understanding the Dutch dialogs in the video clips and answering the practice questions.

3.5. Experimental Procedure

In the fall of 2012, we recruited 31 participants from Radboud In'to Languages, the Nijmegen university language centre, for an experiment with our system (for a description of an earlier experiment, see [16]). We focused on students studying at the A1 or A2 level of the Common European Framework (CEF). Participants were randomly assigned to one of two groups: a group who practiced without CF (NOCF group) or a group which practiced with CF (CF group). All other activities were equivalent. Each participant completed two sessions. In session 1, learners logged into the system and completed a background questionnaire and two proficiency tests before beginning practice. In each session they practiced for a total of 45 minutes. Practice was divided into three 15-minute *micro sessions* (for a total of 6 micro sessions) in which learners practiced spoken Dutch with our system. Before beginning the micro sessions, participants completed the mini-motivation questionnaire for the first time. They then practiced with the system for 15 minutes, three times in a row in each session, resulting in three micro sessions. At the end of each micro session, they completed the mini-motivation questionnaire. That means that each session has four data points and that we have eight data points overall (two sessions).

4. Results

In our analysis of the situated motivation data, we refer to the collection of motivation ratings recorded by the participant at the beginning (S1-0 for session 1 and S2-0 for session 2) or after finishing the three micro sessions (S1-1 to S1-3 and S2-1 to s2-3). Excluded from the following analyses are the data of three subjects who did not complete all eight mini-motivation quizzes.

4.1. Situated motivation

Differences in situated motivation between the NOCF and CF group were analyzed using a repeated measures ANOVA design. In our tests each micro session is one level in the variable time in the analysis. We tested for effects of group, time (all eight mini quizzes) and the interaction effect between the two. The analysis showed significant effects for time, but not for group or the group-time interaction. We observed a time effect for all three situated motivation items (Attitude: $F(7,189)=2.21$, $p = 0.0349$; Motivation: $F(7,189)=7.09$, $p= 0.000$; Self-confidence: $F(7,189)=3.76$, $p=0.001$). As there was no group effect (the CF vs the NOCF group) at all, we proceeded with analyzing the changes over time by including session (the first and second session) and micro session (the four mini quizzes with each session) as independent within-subjects variables. Changes in learners levels of attitude, motivation and self-efficacy can be seen in Figure 2.

Analysis of variance for attitude returned a near-significant effect for session ($F(1,28) = 3.332$, $p = .079$), a significant effect for micro session ($F(3, 84) = 3.465$, $p = .020$), and no effect for the interaction between session and micro session ($F < 1$). Although the factor session was not significant, the mean scores in session 2 were higher (4.22) than in session 1 (4.01), indicating that the

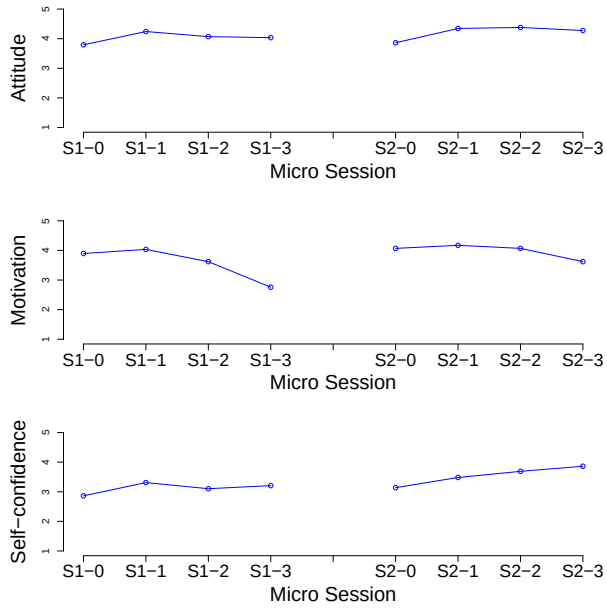


Figure 2: Changes in motivational state for the combined CF-NOCF group. The first graph is attitude, the second motivation, and the third self-confidence. Within-session motivation is sampled approximately every 15 minutes over a 45-minute period.

scores were not decreasing in any way by doing similar exercises again in the second session. It seems an important conclusion in favour of the system we used. Looking at the micro sessions (using the Sidak post-hoc procedure) we see that the measurements at the start (S1-0, S2-0) have the lowest mean scores (3.83) which differ significantly from micro sessions S2-1 and S2-2, which have higher scores, though not from micro session S2-3 (mean scores respectively 3.83, 4.22, 4.16). Learners seem to enjoy the training sessions more than they expected at the start, both in sessions 1 and 2. It can be taken as an indication that practice contributed to enhancing attitude.

For motivation to practice, both the factor session ($F(1,27) = 7.234$, $p = .012$) and the factor micro sessions ($F(3,84) = 9.727$, $p = .000$) turn out to be significant, whereas their interaction is just not ($F(3,84) = 2.663$, $p = .053$). The session effect suggests that scores in the second session are higher than in the first, a positive outcome implying that practice with the system does not evoke negative experiences. The micro session effect indicates a falling tendency as practice time increases within the same session. This effect is supported by the outcomes of a post-hoc test (Sidak procedure) which points to micro session S1-3 as the most conspicuous one (significantly different from all other micro sessions in session 1, no other pair-wise differences being significant; also significantly different from micro sessions S2-1 and S2-2 in session 2, again no other pair-wise differences being significant). The nearly significant interaction effect seems to point to a stronger falling trend for micro session 3 in session 1. The motivation returns at a higher level again at the beginning of session 2.

For situated confidence, the statistical analysis shows both a session effect ($F(1,28) = 4.222$, $p = .045$) and a micro session effect ($F(3,84) = 5.9906$, $p = .001$), with interaction between the two not significant ($F(3,84) = 1.307$, $p = .277$). The session effect means that the scores in session 2 are higher than in session 1, a result in favor of our system. The post-hoc analysis (Sidak procedure) of the micro session effect points out that micro session S1-0 stands out as the condition with the lowest score within the two sessions (significantly different from all micro sessions in session 2). Micro session S2-3 is significantly different from micro sessions S2-1 and S2-2, suggesting a rising increasing trend. The differences in session 1 are too small to deliver significant post-hoc results). The higher score in session 2 and the rising trend within the same session seem to evidence that the learner gets more confident during practicing.

4.2. Post-practice questionnaire

The values for the CF group were systematically higher for all seven relevant questions with an exception for an item concerning learners' evaluations of the practice questions (Q13. The questions were ... (Boring - Nice)) which had identical means (3.8). The overall mean of all seven questions is just not significant for a group effect between the NOCF (mean = 3.74, SD = .693) and CF group (mean = 4.14, SD = .404). It is important to note that both mean scores are high, indicating a positive post-practice evaluation. We found a significant difference in favor of the CF group in a previous experiment [16]. Removing question 13 makes the group difference significant ($F(1,27) = 5.153$, $p = .031$). Given the result of the previous experiment our cautious conclusion is that the CF group tend to be more positive than the NOCF group.

4.3. Practice logs

To complement the subjective analysis above, we checked in the practice logs to look at how practice differed for the two groups by conducting a number of repeated ANOVA tests on three measures: the number of video clips watched, questions completed and attempts made per question. Our analysis of practice behaviours showed that there were significant main effects and interactions. The graphs are displayed in Figure 3.

For number of video clips watched, we did not find an effect for group, indicating that both groups watched an equivalent number of videos in each of the micro sessions. We continued the statistical analysis for the effects of session and micro session. There was a clear effect

for session (F(1,28) = 38.628, p = .000), no effect for micro session (F(2,56) = 1.577, p = .216), and a just not significant interaction between session and micro session (F(2,56) = 2.709, p= .075). So, the learners watched more videos in the second session, with perhaps a slight tendency to raise watching frequency during session 2.

For questions completed, a number of trends are visible. Three effects are significant: session ((1,27) = 14.694, p = .001, the interaction between session and group (F(1,27) = 8.819, p = .006, and micro session (F(2,54) = 13.664, p = .000). The last finding, given the trend visible in figure 3, indicates that the number of questions completed by learners increased across practice.

The interaction effect for group by session shows that the increase in questions per session was different for each group. The trend depicted in figure 3 shows a rise in questions completed for the CF group. The difference between both groups can be made visible when the statistical analysis is done for each of the groups separately. In the NOCF group, a significant effect remains for micro session only (F(2,24) = 3.563, p = .04. A post-hoc analysis (Sidak procedure) reveals a difference between micro session S1-1 (lower scores) and micro sessions S1-2 and S1-3 (higher scores). In the CF group, both session (F(1,15) = 27.604, p = .000) and micro session (F(2,30) = 15.212, p =.000) effects are present. The three micro sessions show a constant rising trend, all differences between the micro sessions being statistically significant (Sidak procedure) for the CF group.

For attempts per question, we have five significant effects, starting with a main effect for group (F(1,27) = 12.704, p=.001). Figure 3 makes clear that the CF group has more attempts per question than the NOCF group in all micro sessions. This seems to be an outcome related to the differences in conditions. We see at the same time a group by session effect (F(1,27) = 12.065, p = .002) in combination with a session effect F(1,27) = 8.165, p = .008), with a smaller difference between the two groups in session 2. A separate analysis of the two groups returns no effect for session in the NOCF group (F(1,12) = 1.061, p = 323) and a significant effect for the CF group (F(1,15) = 13.381, p = .002). This means that there is a change in attempt behaviour in the CF group but not in the NOCF group. The overall analysis gives an effect for both micro session (F(2,54) = 4.226, p = .020) and the interaction between micro session and session (F(2,54) = 4.049, p = .023). The differences between the two sessions in the second micro session seem to be the source of these effects. A post-hoc analysis (Sidak procedure) for session 1 reveals a significant difference between micro session S1-2 and the other two micro sessions, whereas no significant differences turn up at all in a post-hoc analysis of micro sessions in session 2.

The two group by session interaction effects above, for questions and attempts per question, suggest that the CF group, which had to formulate the correct answer before proceeding to the next question, became more proficient over time.

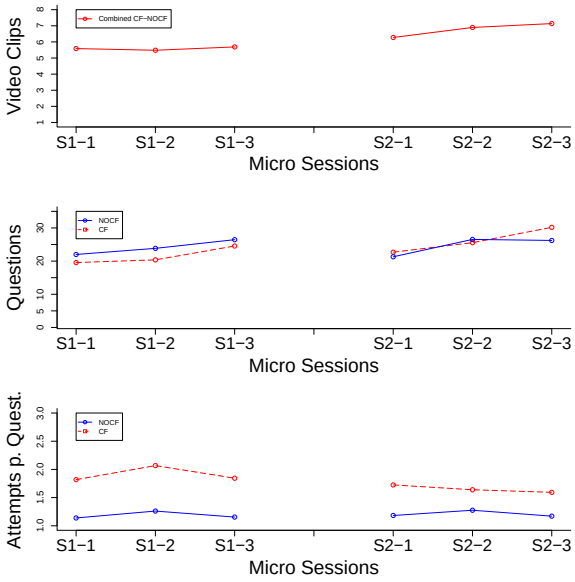


Figure 3: Changes in practice behaviour. The first graph depicts mean video clips viewed for all participants (there is no group effect) for each 15-minute micro session. The second one depicts mean questions completed and the third mean attempts per question for each 15-minute micro session in each group.

5. Discussion and Conclusions

Our analyses indicate that participants in both groups benefited from practice. Learners became increasingly confident about their ability to answer questions in the practice exercise and reported positive attitudes towards practice with the system throughout practice. However, their motivation levels tended to fall as they approached the end of each practice session. Given their positive attitudes and increasing confidence throughout the session, it seems plausible to attribute learners' diminishing desire to practice fatigue. It is encouraging, however, to see that, although motivation decreased at the end of the session, their motivation returned to higher levels at the beginning of the second session. Post-questionnaire analysis point to a similar conclusion, as the results indicate that both groups had similarly positive views on individual questionnaire items surveying learners' attitudes towards practice, opinions concerning utility of practice, and perceived difficulty of the videos and questions. Analysis of the practice logs suggests that participants in the CF group had to produce many more utterances, but that this requirement did not cause significant differences in their

evaluation of the system or their motivation during practice.

Our findings here suggest that CF in our particular computer environment, in the context of grammar practice, did not have a significant effect on learner motivation. This stands in contrast to some views in the SLA literature which suggest it would have a negative effect (e.g. [13]). It may be the case that learning context has a role to play. In cases where learners communicate in meaningful exchanges frequent CF may be perceived as disruptive to the natural flow of exchange. In some social situations, such as the language classroom, CF may also cause a learner to feel embarrassed or to lose face and negatively impact learner motivation. In oral grammar exercises with a computer, which lacks a meaningful exchange or social element, CF may not have the same negative effect.

Other possible explanations include that two 45-minute sessions of practice may not have been enough time to see a group or group-time effect emerge, or that our NOCF group condition may not have been as poor as we assumed: This might be in part due to us telling participants that their recorded answers would be scored and that these scores would be provided to them at a later time. The fact that someone would listen to their utterances and score them, even if not immediately, may have been enough to motivate learners.

A final possibility is that the results were influenced by the experimental design. In the first session, learners completed a number of activities before commencing practice. For a non-native Dutch learner, the experience of filling out a background questionnaire, reading through instruction materials, and completing two proficiency tests before beginning practice may have a large and similar effect on both the CF and NOCF groups which hides any smaller CF-related effects. An interesting future possibility would be to structure the experiment so that questionnaires and proficiency tests are completed on separate days, with two days consisting only of practice activities, so that learners begin in more similar motivational states at S1-0 and S2-0.

Taken together, the results are encouraging. Participants in both groups maintained positive attitudes towards our system over time, and seemed to become increasingly confident about their ability to do well in the practice exercises. Their views after practice confirm this conclusion. Based on our results here, we believe this setup is suitable for situated motivation, by means of mini-questionnaires and practice log analysis, and that these data can be combined with pre- and post-test data to provide a compelling account of what occurs during practice, with links to outcomes. An interesting topic for future experiments would be to use our system to investigate different motivational strategies in the system to explore how learners' motivation can be sustained over time.

6. Acknowledgements

We thank our colleague Joost van Doremalen for developing the ASR component of the CALL system used in this experiment, and the three anonymous reviewers for their comments. This work is part of the research program 'Feedback and the acquisition of syntax in oral proficiency' (FASOP), which is funded by the Netherlands Organisation for Scientific Research (NWO).

7. References

[1] E. Ushioda, *Motivation, Language Identities and the L2 Self: A Theoretical Overview*. Multilingual Matters, 2009, ch. Motivation, Language Identity and the L2 Self, pp. 1–8.

[2] ——, *Psychology for Language Learning - Insights from Research, Theory and Practice*. Palgrave Macmillan, 2012, ch. Motivation: L2 Learning as a Special Case?, pp. 58–73.

[3] R. Ellis, "Corrective feedback and teacher development," *L2 Journal*, vol. 1 (1), pp. 3–18, 2009.

[4] R. C. Gardner, *Social Psychology and Second Language Learning: The Role of Attitudes and Motivation*, R. C. Gardner, Ed. Edward Arnold, 1985.

[5] Z. Dörnyei, *The Psychology Of The Language Learner: Individual Differences In Second Language Acquisition*, Z. Dörnyei, Ed. Routledge, 2005.

[6] G. Crookes and R. W. Schmidt, "Motivation: Reopening the research agenda," *Language Learning*, vol. 4, pp. 469 – 512, 1991.

[7] M. Heilman, K. Collins-Thompson, M. Eskenazi, A. Juffs, and L. Wilson, "Personalization of reading passages improves vocabulary acquisition," *International Journal of Artificial Intelligence in Education*, vol. 20(1), pp. 73–98, 2010.

[8] E. Ushida, "The role of students' attitudes and motivation in second language learning in online language courses," *CALICO Journal*, vol. 23 (1), 2005.

[9] A. de Vicente and H. Pain, "Informing the detection of the students' motivational state: An empirical study," in *Proceedings of the 6th International Conference on Intelligent Tutoring Systems*, ser. ITS '02. London, UK, UK: Springer-Verlag, 2002, pp. 933–943.

[10] M. E. Tatawy, "Corrective feedback in second language acquisition," *Working papers in TESOL and Applied Linguistics*, vol. 2(2), pp. 1–19, 2002.

[11] S. Krashen, *Principles and practice in second language acquisition*. Pergamon Press Inc., 1982.

[12] B. D. Schwartz, "On explicit and negative data effecting and affecting competence and linguistic behavior," *Studies in Second Language Acquisition*, vol. 15, p. 147163, 1993.

[13] J. Truscott, "The case against grammar correction in L2 writing classes," *Language Learning*, vol. 46(2), pp. 327–69, 1996.

[14] ——, "What s wrong with oral grammar correction," *The Canadian Modern Language Review*, vol. 55 (4), pp. 437–456, 1999.

[15] S. Bodnar, B. P. de Vries, C. Cucchiarini, H. Strik, and R. van Hout, "Feedback in an asr-based call system for l2 syntax: A feasibility study," in *Proceedings of the SLaTE-2011 workshop*, 2011.

[16] B. Penning de Vries, S. Bodnar, C. Cucchiarini, H. Strik, and R. van Hout, "Spoken grammar practice in an asr-based call system," in *Proceedings of the SLaTE-2013 workshop*, 2013.

Filtering-based Automatic Cloze Test Generation

Kyusong Lee[1], *Soo-Ok Kweon*[2], *Hae-Ri Kim*[3], *Gary Geunbae Lee*[1]

[1]Department of Computer Science and Engineering,
[2]Division of Humanities and Social Sciences,
Pohang University of Science and Technology, Korea
[3]Department of English Education,
Seoul National University of Education, Korea

{[1]Kyusonglee, [2]soook, [1]gblee}@postech.ac.kr, [3]hrkim@snue.ac.kr

Abstract

We propose a method to generate high-quality cloze test questions using a computational approach. Previous methods for automatic cloze test generation have contained some problems; specifically, there can be multiple correct answers. We found that approximately 50% of the generated answers have such errors with previous methods, which requires human post-editing was necessary in previous research. We propose an N-gram filtering method that can detect the answer to a given question. We compare the errors of the generated questions before and after applying the filtering methods. We found that our filtering method can select quality distractors by reducing errors in the generated questions. Moreover, when we generate cloze tests using semantic similarity, non-native speakers are very hard to answer the questions.

Index Terms: Cloze Test Generation, Sentence Completion Task, Vocabulary Question Generation

1. Introduction

A cloze test is an activity in which students must fill in the blanks in a text with appropriate words. Although this type of test has been widely used in language learning to assess learner's proficiency in the target language, it requires significant labor to create because writing test questions by hand is a laborious task even for experienced teachers. To lessen the burden of test development, automatic question generation techniques [1];[2];[3];[4] have been sought. The goal of these techniques is to provide questions with constant quality and appropriate difficulty that, lead to an objective assessment. Mitkov and Ha proposed an NLP-based methodology for the construction of test items from instructive texts such as textbook chapters and encyclopedia entries [5]. The system for generation of multiple-choice test described in Mitkov and Ha [6] and in [7] was evaluated in a practical environment in which the user was offered the option to post-edit and in general to accept, or reject the test items generated by the system. The formal evaluation demonstrated that a significant fraction of the generated test items needed to be discarded. The primary motivation of our work is that a considerable number of generated questions are not sufficiently good or practical use in language learning. Human post-editing processing is still needed in automatic cloze test generation. Our purpose in this paper is to reduce the cost of the human post-editing step by filtering improper distractors.

In section 2, previous works are introduced. In section 3, we will introduce methods. And then, the experiment result is explained in section 4. Finally, we give a conclusion in section 5.

2. Precious Work

So far, very little work has been devoted to filering distractors. Several studies suggested implementing naïve filtering steps using a the web-based approach [8];[1]. In this approach, distracters are eliminated when sentences receive a non-zero number of hits in a web search because distracters must be incorrect. However, this approach has many problems. First, hits may come from non-native speakers' websites and contain invalid language usage. Second, even if sentence fragments cannot be located on the web, it does not necessarily imply that they are incorrect. Additionally, the number of web searches performed using Google and Bing each month is limited. Grammaticality and collocation to select distractors are used and semantically similar words are removed to avoid multiple answers [4]. However, semantic similarity is also a useful feature for generating cloze items and a mixed strategy demonstrates the best performance as described in [5]. To overcome these limitations, we suggest including a sophisticated filtering step in automatic question generation techniques to improve the acceptance rates of generated test items. If a distractor candidate is considered a possible answer to the question, it is eliminated from the candidate list. The filtering method is independent of candidate distractors selection. Thus, we can deploy more varied features and can apply a mixed strategy to generate cloze items using our filtering method. In this paper, we generated the distractors using semantic similarity for test data sets.

3. Method

3.1. Overview

We propose performing automatic cloze test generation using the following steps (Figure 1). 1) Input the sentence with a blank position 2) We select the distractor candidates using a target based on previous research such as semantic and phonetic similarity [5], synonym or related word using thesaurus extraction [9], WordNet [10], collocation and grammaticality [4] methods, or mixed strategy etc. 3) To make all distractors have the same Part-of-

speech (POS) form, we save the part-of-speech of the target word (answer word); then, we change all words to the same form as the answer word using the English Synthesizer[1]. 4) N-grams can be used to remove potential multiple answers

3.2. Filtering Distractor

N-gram scores are used for filtering the words among distractor candidates. We build a probabilistic language model using the Google Web1T N-gram Count corpus[2], which is built by a huge amount of data. One of baselines is established with Laplace smoothing the N-gram model [11] which avoids zero probability in equation (1).

$$\widehat{\mathrm{P}}_{\mathrm{Laplace}}(\mathrm{w_i}|\mathrm{w_{i-3}w_{i-2}w_{i-1}}) = \frac{\mathrm{C(w_{i-3}w_{i-2}w_{i-1}w_i)}+1}{\mathrm{C(w_{i-3}w_{i-2}w_{i-1})}+|\mathrm{V}|} \quad (1)$$

Zero probability problems are not evitable when considering 5-ngram probability. Backoff N-gram models were introduced by Katz (1987). If the N-gram that we need has zero counts, we approximate it by backing off to the (N-1)-gram, as shown in equation (2).

$$P_{katz}(w_n|\mathrm{w}_{n-N+1}^{n-1}) = \begin{cases} P^*(w_n|\mathrm{w}_{n-N+1}^{n-1}) & if\ C(\mathrm{w}_{n-N+1}^{n}) > 0 \\ \alpha(\mathrm{w}_{n-N+1}^{n-1})P_{katz}(w_n|\mathrm{w}_{n-N+2}^{n-1}) & otherwise \end{cases} \quad (2)$$

However, we recognize that the Google N-gram corpus cannot be used to build previous smoothing methods because of the frequency cut-offs, which implies that only N-grams appearing more than 40 times were kept and appear in the N-gram tables. Some methods need low-order word counts. However, all N-grams with counts lower than 40 were discarded, we cannot use most of previous smoothing methods Calculating normalized factor α in the Katz backoff is also difficult. Our purpose to use N-gram is to find the most proper word among distractor candidates in the question sentence as equation (3).

$$\mathrm{E} = \underset{x \in \{C_0, C_1, .., C_n\}}{\operatorname{argmax}} P(x|Stem) \quad (3)$$

C denotes that distractor candidates.(e.g., C = {addressing, coming, recognizing, greeting, saluting, presenting, receiving, contenting, bidding, accosting}). Stem denotes the words in the question sentence (e.g., Stem = "The manager suggested [ ] the resort guests with a culture themed show after they had finished settling into their private rooms.") (Figure 1). E denotes the potential answers from N-gram filtering model.

Previous N-gram model has low performance (the accuracy is 52%) for the sentence completion challenge (see the results in Table 4). Only half of questions in the close test could be correctly answered by the N-gram approach which implies that it is a challenging task. However, to filter the potential answer words in the distractor candidates, the accuracy of finding the correct answer by N-gram must be much higher than current state-of-art performance. Thus, we improved the performance of N-gram model on sentence completion task by considering more effective features when calculating the N-gram probability. The preceding words and following words are already given in a question sentence, so we can consider the both directions and various

1. Input (Text with Black position)
The manager suggested **[welcoming]** the resort guests with a culture themed show after they had finished settling into their private rooms .

2. Candidates distractors (Target : welcoming)
address, come, recognize, greet, salute, present, receive, content, bid, accost

3. English Synthesizer (Target: welcome, POS: VBG)
addressing, coming, recognizing, greeting, saluting, presenting, receiving, contenting, bidding, accosting

4. Removing multiple answer by N-gram
addressing, coming, recognizing, ~~greeting~~, saluting, ~~presenting~~, ~~receiving~~, ~~contenting~~, bidding, accosting

5. Final Selection
addressing, coming, recognizing, saluting, bidding, accosting

6. Output (cloze Test)
The manager suggested _____ the resort guests with a culture themed show after they had finished settling into their private rooms .
a) addressing b) coming c) accosting d) **welcoming**

Figure 1. Overall Process of Generating Cloze test

ranges. Our proposed for using N-gram model can consider both forward and backward N-gram probability and has much less data sparseness problem for the filtering model. From 5-gram to 2-gram count information are used to develop the proposed N-gram model.

A backoff proposed N-gram model is used for the N-gram filtering model as in equation (4). If the N-gram probabilities we need have all zero counts for every C, we approximate them by backing off to the (N-1)-grams.

$$\mathrm{E} = \underset{x \in \{C_0, C_1, .., C_n\}}{\operatorname{argmax}} P_{\mathrm{N}}(x|Stem) \quad (4)$$

where

$$P_{\mathrm{N}}(x|Stem) = \begin{cases} P_{\mathrm{N}}^*(x|Stem) & if\ not \sum_{i=0}^{n} P_{\mathrm{N}}(C_i|Stem) = 0 \\ P_{\mathrm{N-1}}(x|Stem) & otherwise \end{cases}$$

The model uses a variety of contexts and different sizes and positions to replace the distractor candidates' words in C, where C = { addressing, coming, recognizing, greeting, saluting, presenting, receiving, contenting, bidding, accosting } in Figure 1. We can retrieve a count for each context pattern of length-N with a filler word replacing c_i, which constitutes a single N-gram. We then retrieve a count using the Google 5-gram for sequences

[1] http://www.languagetool.org

[2] Available from the LDC as LDC2006T13.

Table 1: Proposed back off N-gram (Example: The manager suggested $[C_i]$ the resort guests with a culture themed show after they had finished settling into their private rooms.)

5-GRAM	
P_5	P(Ci | the resort guests with) + P(Ci | <s> The manager suggested) + P(suggested Ci |<s> The manager)+ P(Ci the resort | guests with)+ P(manager suggested Ci | <s> The)+ P(Ci the resort guests | with)+ P(The manager suggested Ci | <s>)+ P(suggested Ci | the resort guests)+ P(Ci the | The manager suggested)+ P(suggested Ci the | resort guests)+ P(suggested Ci the | The manager)+ P(suggested Ci the resort | guests)+ P(manager suggested Ci the | The)+ P(manager suggested Ci | the resort)+ P(Ci the resort | manager suggested)+ P(manager suggested Ci the | resort)+ P(suggested Ci the resort | manager)+ P(The manager suggested Ci | the)+ P(Ci the resort guests | suggested)
4-GRAM	
P_4	P(Ci | the resort guests)+ P(Ci | The manager suggested)+ P(Ci the | resort guests)+ P(suggested Ci | The manager)+ P(Ci the resort | guests)+ P(manager suggested Ci | The)+ P(suggested Ci | the resort)+ P(Ci the | manager suggested)+ P(suggested Ci the | resort)+ P(suggested Ci the | manager)+ P(manager suggested Ci | the)+ P(Ci the resort | suggested)
3-GRAM	
P_3	P(Ci | the resort)+ P(Ci | manager suggested)+ P(Ci the | resort)+ P(suggested Ci | manager)+ P(suggested Ci | the)+ P(Ci the | suggested)
2-GRAM	
P_2	P(Ci |the)+ P(Ci | suggested)

including N-grams of length 2 to 5. For each target word c_i, five separate 5-gram context patterns that span its range are found. We describe the notation of P_N in more detail in Table 1).

4. Experiment and Result

4.1. Data

For our proposed filtering strategies, we used Google N-gram corpus which contain 1 trillion words of running text and the counts for all 1 billion five-word sequences that appear at least 40 times. There are 13 million unique words after words that appear less than 200 times are discarded. We used WordNet-based semantic similarity to generate candidate distracters in this paper. For computing the WordNet-based semantic similarity, we employed a popular word similarity measure using the Python NLTK package [12]: Jiang and Conrath's (JCN) measure [13] to generate the distractor candidates. A total 100 questions are generated using WordNet-based semantic similarity. Each question has 10 distractor candidates. Four experts in English Education annotated every generated distractors in the JCN measure regarding whether distractors could be potential answers. Two annotators' native language is English, the others are non-native speakers. Kappa value between two native was 0.7. However, the agreement between non-native speakers and native speakers was 0.409, 0.401, 0.45, and 0.455. Between two non-native speakers, Kappa value is 0.764. Even though they are all experts in English Education, it is a challenging tasks for non-native speakers; the kappa value is about 0.4 (Table 2, Table3). It indicates that questions by semantic similarity could be good test items for identifying native speakers. Moreover, it would be good test items for non-native speakers to teach the real usage of words among the similar meanings. To investigate the performance of the filtering method that uses N-gram model, we use 500 semantic

Table 2. Multiple answer annotations, NonNS1 denotes Non-native speaker 1, NS denotes Native speaker. O means proper distractor, X means potential answer.

Question) Remember to [ ] your complete company information when filling out the tax form .

Answer) include

Distractor Candidates)

	NonNS1	NonNS2	NS1	NS2
bear	O	O	O	O
involve	X	X	O	O
carry	O	O	O	O
embroil	O	O	O	O
admit	O	O	O	O
add	O	O	X	X
hold	O	O	O	O
tangle	O	O	O	O
drag	O	O	O	O
contain	X	X	O	O

Table 3: Kappa value between annotators

	NonNS1	NonNS2	NS1	NS2
NonNS1	1			
NonNS2	0.764	1		
NS1	0.409	0.401	1	
NS2	0.455	0.455	**0.70**	1

questions from TOEIC data[1] and the MSR sentence completion challenge data[2] [14].

4.2. Filtering By N-Gram

We only explore content words (verbs and nouns) in this paper. To evaluate our filtering method, we must select the N-best candidate distractors. For the experiment, we selected the N-best semantically similar words to the target vocabulary scheduled by JCN WordNet based semantic similarity measure in Figure 1, such as *addressing, coming, recognizing, greeting, saluting, presenting, receiving, contenting, bidding,* and *accosting*. The goal of the filtering method is to remove *Greeting, presenting, receiving,* and *contenting* which labeled as multiple answers using N-gram model. Among the 100 test items, each item has 10 distractor candidates, which yields a total 1000 distractors candidates labeled as multiple answers or proper distractors. If a distractor could be an answer in the given text, we count the question as an instance of "multiple answers". The performances are quantified using precision, recall, and F-score. To explore the filtering performance for multiple answers, we deploy the proposed N-gram. Because the number of web searches performed using Google and Bing in a month is limited, it is improper to use those corpora as the baseline system. Thus, the baseline is randomly selected from distractor candidate. We consider two perspectives of the results: one is the proper distractors perspective and the other is filtering perspective.

Proper distractor perspective:

$$\text{Precision} = \frac{\#\text{ of proper distractor in final}}{\#\text{ of distractors in final}}$$

$$\text{Recall} = \frac{\#\text{ of proper distracotrs in final}}{\#\text{ of proper distractors in candidates}}$$

Filtering perspective:

$$\text{Precision} = \frac{\#\text{ of filtered improper distractors in final}}{\#\text{ of total filtered distractors}}$$

$$\text{Recall} = \frac{\#\text{ of filtered improper distractors in final}}{\#\text{ of total improper distractors in candidates}}$$

From the proper distractor perspective, eliminating the proper distractors is not a critical problem when final distractors are all proper distractors which indicate precision is important from this perspective. The reason of the low recall rate is that many proper distractors are also removed until the final number of distractors K is remains (K=the number of Distractors). Therefore, low recall is not a critical problem. However, from the filtering perspective, recall is much more important than precision. If recall is 100%, the filtering method can eliminate every improper distractor. We found that the recall rate is much higher than the baseline which implies that our method filters improper distractors properly (Table 3). We compared the performance after applying filtering model in as shown (Table 4). The 46.8% of generated questions have distractors that are improper for practical use without human editing. After applying our proposed filtering strategies, significant parts of questions are removed the improper distractors.

Table 3. Performance of Proper Distractor Selection

		Precision	Recall	F-Score
Proper Distractor Perspective	PROPOSED	**90.9**	44.64	59.90
	BASELINE	83.51	40.99	54.99
Filtering Perspective	PROPOSED	24.82	**80.45**	37.94
	BASELINE	10.99	35.63	16.80

Table 4. The portion of errors on generated Cloze Test

	Suitable Questions
Baseline	53.2
After Filter	70.2

Table 5. Performance of methods on the TOEIC test

	TOEIC data
Chance	25%
Baseline Laplace Smoothing	61
*Proposed Method	**74.0***

Table 6. MSR sentence completion (SC) performance with the proposed N-gram method

	MSR SC
Chance	20 %
(1) GT N-gram LM	39
(2) LSA- Total Similarity	49
Combination (1) + (2)	52
*Proposed Method	**87.4***

We found a 17% improvement gain after applying our filtering methods.

An essential step in the filtering process is to eliminate potential answers, so we believe testing our method's ability to find the correct answers in sample questions is a useful assessment. Therefore, for additional evaluation, we explore how effective the filtering method is in selecting potential answers, we apply our filtering method on TOEIC questions. In all, 74% of the questions are correctly answered (the fraction that would be answer by chance is 25%, and baseline is 61%) (Table 5). The TOEIC questions that we use consist of all semantically related questions. Moreover, we explore the accuracy for the sentence completion challenge [14] with the proposed N-gram model using the Google corpus. The challenge was designed as a benchmark for semantic models and consists of SAT-style sentence completion problems. Given 1,040 sentences, each of which is missing a word, the task is to select the correct word out of the candidates provided for each sentence. The best result 52% was produced by a combination of latent semantic allocation total similarity and N-gram models (Zweig & Burges, 2011). Our method achieves better performance than these previous the results from [15], as indicated in Table 6. We used the evaluation tool that MS provided. Note that the result

[1] http://www.toeflgoanywhere.org/, data cannot open because of license problem

[2] http://research.microsoft.com/en-us/projects/scc/, The evaluation tool is also available in the website.

of "*proposed N-gram" in Table 6 are only our experimental result, others results such as (1), (2), and the combination (1)+(2) are from the [15] paper. The accuracy of our proposed method is 87.4 %, which is significantly improved. The annotated distractors and experimental results are made available to the public[1].

5. Conclusions

We found that machine generated test items have many errors, such as multiple answers. To solve these problems, we proposed filtering method to remove improper words from candidate distractors. We found that proposed methods significantly reduce the generated test item error rate. Moreover, our method also performs well on sentence completion challenge. We also found that the annotation agreement between native speakers are much higher than between native speaker and non-native speakers. It indicates that questions made by semantic similarity are challenging test items for non-native speakers. We have plan to explore a real student test for evaluations on generated cloze test set.

6. Acknowledgements

Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2012-0008835). "This research was supported by the MKE (The Ministry of Knowledge Economy), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2012-C1090-1231-0009)

7. References

[1] T. Goto, T. Kojiri, T. Watanabe, T. Iwata, and T. Yamada, "Automatic Generation System of Multiple-Choice Cloze Questions and its Evaluation," *Knowledge Management & E-Learning: An International Journal (KM&EL),* vol. 2, pp. 210-224, 2010.

[2] C. Y. Chen, H. C. Liou, and J. S. Chang, "FAST: an automatic generation system for grammar tests," in *COLING-ACL '06 Proceedings of the COLING/ACL on Interactive presentation sessions*, 2006, pp. 1-4.

[3] J. Lee and S. Seneff, "Automatic generation of cloze items for prepositions," 2007.

[4] J. Pino, M. Heilman, and M. Eskenazi, "A selection strategy to improve cloze question quality," in *Proceedings of the Workshop on Intelligent Tutoring Systems for Ill-Defined Domains. 9th International Conference on Intelligent Tutoring Systems, Montreal, Canada*, 2008, pp. 22-32.

[5] R. Mitkov, L. A. Ha, A. Varga, and L. Rello, "Semantic similarity of distractors in multiple-choice tests: extrinsic evaluation," 2009, pp. 49-56.

[6] R. Mitkov and L. A. Ha, "Computer-aided generation of multiple-choice tests," 2003, pp. 17-22.

[7] R. Mitkov, L. A. Ha, and N. Karamanis, "A computer-aided environment for generating multiple-choice test items," *Natural Language Engineering,* vol. 12, pp. 177-194, 2006.

[8] E. Sumita, F. Sugaya, and S. Yamamoto, "Measuring non-native speakers' proficiency of English by using a test with automatically-generated fill-in-the-blank questions," in *Proceedings of the second workshop on Building Educational Applications Using NLP*, 2005, pp. 61-68.

[9] M. Heilman and M. Eskenazi, "Application of automatic thesaurus extraction for computer generation of vocabulary questions," in *Proceedings of the SLaTE Workshop on Speech and Language Technology in Education*, 2007, pp. 65-68.

[10] J. C. Brown, G. A. Frishkoff, and M. Eskenazi, "Automatic question generation for vocabulary assessment," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 819-826.

[11] G. J. Lidstone, "Note on the general case of the Bayes-Laplace formula for inductive or a posteriori probabilities," *Transactions of the Faculty of Actuaries,* vol. 8, p. 13, 1920.

[12] E. Loper and S. Bird, "NLTK: The natural language toolkit," 2002, pp. 63-70.

[13] J. J. Jiang and D. W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy," *Arxiv preprint cmp-lg/9709008,* 1997.

[14] G. Zweig and C. J. C. Burges, "The Microsoft Research Sentence Completion Challenge," 2011.

[15] G. Zweig, J. C. Platt, C. Meek, C. J. C. Burges, A. Yessenalina, and Q. Liu, "Computational Approaches to Sentence Completion," in *the Association for Computational Linguistics*, Jeju, Korea, 2012.

[1] https://sites.google.com/site/dataforslate/

Methodological Issues in Evaluating a Spoken CALL Game: Can Crowdsourcing Help Us Perform Controlled Experiments?

Manny Rayner, Nikos Tsourakis

University of Geneva, FTI/TIM, 40 bvd du Pont-d'Arve, CH-1211 Geneva 4, Switzerland
{Emmanuel.Rayner,Nikolaos.Tsourakis}@unige.ch

Abstract

We summarise a series of experiments we have carried out over the last three years on CALL-SLT, a speech-enabled web-based CALL game for learning and improving fluency in domain language, focussing on the methodological aspects. In particular, we argue that our previous evaluations have been systematically flawed due to the lack of a control group. We present a detailed description of our most recent evaluation, where 130 subjects, recruited using crowdsourcing methods, followed a short course in basic French over a period of one week, with 24 subjects completing the course. About a third of the subjects (half of the ones that finished) were assigned to a control group who used a version of the system with speech recognition feedback disabled; subjects in both groups demonstrated significant improvements in language skills over the duration of the experiment, but the improvements were significantly larger for the non-control subjects. We argue in conclusion that this type of experiment opens up interesting new ways to attack the difficult problem of performing controlled experiments with CALL applications.

Index Terms: CALL, speech recognition, evaluation, methodology, crowdsourcing

1. Introduction

Many people have now built CALL tools which use speech recognition and other complex technology. But how can we establish that the advanced functionalities incorporated in these tools do anything useful, in terms of actually helping students develop their language skills? In this paper, we present a case study based on a series of evaluations we have carried out on different versions of CALL-SLT [1], a web-enabled CALL app for learning and improving fluency in domain language.

In the specific instance of CALL-SLT, the central idea is to use speech recognition to provide feedback to the student about their ability to speak in a foreign language. The system prompts the student, indicating what they are supposed to say; the student responds, and the system either accepts or rejects their response. But the student also has a much simpler way to learn. If they are stuck, they have the option of asking for help and hearing a correct response, which they can imitate. Listening and imitating is almost certainly useful, so the question is whether the sophisticated speech-recognition functionality adds anything extra. Without some kind of control group, it seems difficult to make any strong claim to this effect.

In the rest of the paper, we explain how we have developed an evaluation methodology designed to address these issues. We start by describing the CALL-SLT system (§ 2) and critically examining the adequacy of our previous evaluations (§ 3). The novel results of the paper are in § 4 and § 5, where we present our most recent experiment, carried out over the Amazon Mechanical Turk (AMT) on 130 crowdsourced subjects, in which we included a control group who used a version of the system where feedback from speech recognition was disabled. We argue in the final section that this style of evaluation can potentially address the methodological difficulties associated with our earlier efforts; the question is the extent to which the new problems it creates are acceptable.

2. The CALL-SLT System

CALL-SLT [1] is an open-source speech-based translation game designed for learning and improving fluency in domain language. It is based on the "spoken translation game" idea originating with [2]; a related application is TLTCS [3]. The system is accessed via a client running on a web browser. Most processing, in particular speech recognition and linguistic analysis, is carried on the server side, with speech recorded locally and passed to the server in file form [4]. The current version, available at `http://callslt.org`, supports French, English, Japanese, German, Greek and Swedish as L2s and English, French, Japanese, Arabic and Chinese as L1s.

The system is based on two main components: a grammar-based speech recogniser and an interlingua-based machine translation (MT) system, both developed using the Regulus platform [5]. This architecture presents several advantages in the context of the web-based CALL task. The system is not related to a particular language or domain, as in [2]. The Regulus platform offers many tools to support addition of new languages and new coverage (vocabulary, grammar) for existing languages: the

recogniser's language model is extracted by specialisation from a general resource grammar in order to get an effective grammar for a specific domain, with the specialisation process driven by a small corpus of sentences. The general grammar can thus easily be extended or specialised for new exercises by changing the corpus, enabling rapid development of new content. The specialised grammar-based language models give good recognition performance on in-coverage sentences even without speaker adaptation; for example, the evaluation exercise described in [6] showed that recognition Word Error Rates for native speakers were very low, typically around 1–2%.

Each turn begins with the system giving the student a prompt, consisting of a surface realisation of an interlingua structure in a predicate-argument notation called Almost Flat Functional Semantics (AFF; [7]). In the plain version of the system, the surface realisation is a text string formulated in a telegraphic version of the L1. The student gives a spoken response; it is in general possible to respond to the prompt in more than one way. Thus, for example, in the version of the system used to teach English to French-speaking students, a simple text prompt might be:

`DEMANDER DE_MANIERE_POLIE BIÈRE`

("`ASK POLITELY BEER`"). representing the underlying interlingua representation

```
[null=[utterance_type,request],
 null=[politeness,polite]
 arg2=[drink,beer]]
```

The responses "I would like a beer", "could I have a beer", "please give me a beer", or "a beer please" would all be regarded as valid.

The system decides whether to accept or reject the response by first performing speech recognition, then translating to an interlingua representation, and finally matching this interlingua representation against the interlingua representation of the original prompt. A "help" button allows the student, at any time, to access a correct response, given in both written and spoken form. The text forms come from the initial corpus of sentences or can be created by the MT system to allow automatic generation of variant syntactic forms. The associated audio files are collected by logging examples where users registered as native speakers got correct matches while using the system. Prompts are grouped together in "lessons" unified by a defined syntactic or semantic theme.

The student thus spends most of their time in a loop where they are given a prompt, optionally listen to a spoken help example, and attempt to respond to the prompt. In the version of CALL-SLT used in the present experiment, the system gave the following minimal feedback to the response. First, it echoed back the student's recorded utterance; second, it placed a border around the text part of the prompt which was either green (accept) or red (reject). We have experimented with more complex feedback, but the difficulty of making it sufficiently reliable means that it is perceived by many students as confusing rather than helpful. At each turn, the student can repeat the current prompt, use arrow controls to move to the next or previous prompt, or switch to a different lesson. The interface used is shown in Figure 1.

Figure 1: Screenshot of "video" version. The video is saying "Quelle est ta nationalité?" (What is your nationality?); "GREEK" is the text part of the prompt, and "Je suis grec" is a help example.

In the plain version of the system, the relationship between abstract interlingual representations of prompts and their surface text realisations is defined by means of another Regulus grammar [8]; the abstract representation is converted into the prompt by running this grammar in generation mode. In the versions of CALL-SLT used here, this mechanism was extended to generate multimedia prompts. The grammar was modified slightly so that some lexical rules define elements of the form `multimedia:`⟨`MultimediaTag`⟩. In a post-processing step, the multimedia tags are removed from the string, and replaced by names of prerecorded multimedia files according to a table which defines a non-deterministic mapping from ⟨multimedia-tag, lesson⟩ pairs to files. Thus, continuing the previous example, the French multimedia version of the prompt intended to elicit a response similar to "I would like a beer" would be

`multimedia:ask-for-drink BIÈRE`

In the initial multimedia configuration we have deployed in the present experiment, multimedia prompts are con-

cretely realised by playing a recorded file corresponding to the multimedia tag and displayed the remaining text; so in the above example, the system would play a recorded video file with a question meaning "What would you like to drink?", while the text "BIÈRE" was displayed.

3. Previous experiments

We start by describing previous evaluations of CALL-SLT which did not use a control group, all of which followed the same basic pattern. A group of students were asked to use the system; we then analysed logged data, looking for evidence that the students improved their language skills between the start and the end of the period in question. Some evaluations were performed over short periods, ranging from a day to a week; others over longer spans of time, as part of a formal language course.

To take a typical example, [9] reported an experiment where ten students used the French-for-Chinese-speakers version for two sessions over one day. We argued that the results provided evidence that subjects had learned from using the system. First, students had a higher proportion of utterances accepted by the system in the later utterances than in the earlier ones, this difference being statistically significant. Second, grammar and vocabulary tests carried out before and after the experiment showed large differences; most of the students appeared to have picked up some vocabulary, and there was also reason to believe that they had consolidated their knowledge of grammar.

Looking critically at the design, we can advance various objections against the validity of our conclusions. One obvious question is whether the fact that students have more utterances accepted by the system after they have used it for a while really does mean that they have improved their generative spoken language skills. Other explanations are a priori quite possible. In particular, they may only have become more skillful at using the interface, learning to speak in a way that is better adapted to the machine, but not necessarily better in itself. The experiments described in [6], however, suggest that these criticisms are not so serious. When native speaker judges are presented with pairs of utterances chosen so that both utterances are responses by the same student to the same prompt, one of which is accepted by the system and one rejected, they tend to agree reasonably well with the recogniser about which member of the pair is better.

A more serious objection, however, is that, even if the results unambiguously show that the student has improved their language skills over a given period, it is still not clear that the improvement can be ascribed to the fact that the student has been using the system. This problem is particularly acute when use of the system is integrated into a formal language course; given that the student is also receiving other kinds of instruction, it is obviously possible that any improvement measured is independent of use of the system. Even if the student is only learning through use of the system, at least over the duration of the experiment, it is still unclear which aspects of the system are responsible for the improvement. In an application like CALL-SLT, the student spends a large part of their time listening and repeating, which may well be helpful for them. It remains to be shown that any of the more sophisticated system functionalities are useful in practice.

Considerations like those above naturally point in the direction of performing controlled experiments, where students using the system are contrasted against a suitable control group. Unfortunately, experience has shown that it is far from easy to define such a group, partly because motivation is always an important factor in language learning. For example, suppose, as in e.g. [10], that we pick subjects randomly from one class, assigning half of them to the group using the system and the other half to the control. The two groups of students will talk to each other. If the system is perceived as useful, which the authors claim in the cited study, it is reasonable to wonder whether students in the control group felt correspondingly unmotivated; it is methodologically better if no subject is aware that any version exists except the one they are using.

If, on the other hand, we take the two groups from two different classes that have no contact with each other, not mixing them, it is impossible to know whether the classes are comparable. Most teachers we have asked say their experience suggests high variability between classes. Yet another possibility is to use a crossover methodology, letting students in the same class alternate between the two groups. Some clear successes have been claimed for this methodology, in particular by the LISTEN project [11, 12, 13]; if the learning effect from using the system is large enough, as appears to be the case there, it is reasonable to hope for a clear result. There are however many known problems with crossover, since it is difficult to account correctly for the effect of using the main system and the control version in different orders. In the context of CALL, students may once again be disappointed if they like the main system and are then forced to use the inferior control, and react accordingly.

For the kinds of reasons outlined above, it has often been argued that controlled experiments are unproductive in CALL [14], and that single-case design methodologies [15] are more appropriate. Recently, however, the introduction of easily available crowdsourcing platforms like the Amazon Mechanical Turk (AMT) has opened up new possibilities. In a large, diverse online community, it is not unreasonable to hope that subjects can be chosen randomly, and in general have no contact with each other; under circumstances like these, a controlled experiment has greater chances of avoiding the known methodological pitfalls. In the next two sections, we describe an experiment of this kind carried out on CALL-SLT.

4. Controlled evaluation using crowdsourcing

The experiment we describe here was carried out in early 2013 using a multimedia-enabled Android phone version of the French CALL-SLT system. The main content consisted of four lessons, *about-me* (simple questions about the subject's age, where they live, etc); *about-my-family* (similar questions about family members); *restaurant* (ordering in a restaurant) and *time-and-day* (times and days of the week). Three additional lessons called *overview-1*, *overview-2* and *revision* will be described shortly. The course was designed for students with little or no previous knowledge of French. It covered about 80 words of vocabulary and a dozen or so basic grammatical patterns.

We created four different versions of the basic system. Three of them differed only in the way the multimedia part of the prompt was realised: in **video** it had the form of a recorded video segment of a human speaker, in **avatar** it was an animated avatar, and in **text** it was a piece of text. The fourth version, **no-rec**, was the same as **video**, except that the student was given no feedback to show whether speech recognition and subsequent processing had accepted or rejected their response.

Subjects were recruited through AMT; we requested only workers from the US. After discovering during a previous study that experiments of this kind can easily attract scammers, we required all workers to have a track record of at least 50 previously completed Human Interface Tasks (HITs), at least 80% of which had been accepted.

The experiment was carried out in two cycles, each of which had the same sequence of eight HITs. In the first HIT, the task was to check that one version of the app (we chose **no-rec**) could be successfully run on an Android phone. Subjects who gave a positive response were then randomly assigned to the four different versions of the system and given different versions of the subsequent HITs. AMT "qualifications" were used so that subjects doing one version of a HIT were unable to see that HITs for other versions existed. The seven HITs were issued at 24-hour intervals; workers were paid $1.00 for the first HIT and $2.00 for each subsequent one, reasonable pay by AMT standards. The HITs had the following content:

Pre-test: The student was asked to do *overview-1* and *overview-2*, each of which consisted of a balanced selection of examples from the other lessons. During *overview-1*, they were encouraged to use the Help function as much as they wished, so the main skill being tested was ability to imitate. In *overview-2*, Help was switched off, so the main skill tested was generative ability in spoken French.

Lessons 1–4: The student was asked to attempt each of the four lessons in turn, one lesson per HIT, with Help turned on. They were told to spend a minimum of 20 minutes practising, and speak to the system at least 25 times.

Revision: The student was warned that the next HIT would be a test (they were not told what it was), and was asked to revise by doing the *revision* lesson, which contained the union of the material from the four main lessons, for at least 20 minutes.

Post-test: The student was asked to do *overview-1* and *overview-2* again. They were told that the intent was to measure how much they had learned during the course, and were asked to do the test straightforwardly without cheating.

The purpose of the pre- and post-tests was to measure the progress the students had made during the main course of the experiment by comparing their results across the two rounds. The mode of comparison will be described shortly.

In the first cycle, we started with 100 subjects. The second column of Table 1 shows the number of students left in play after each round of HITs. At the end of the cycle, there were 17 students who had completed both the pre- and post-tests. A preliminary examination of the results suggested that students performed similarly on the three versions which gave recognition feedback, but worse on **no-rec**; there was not, however, sufficient data to be able to draw any significant conclusions.

Round	Remaining	
	Cycle 1	Cycle 2
Recruit	80	22
Pre-test	36	14
About-me	29	11
My-family	24	10
Restaurant	22	9
Time-and-day	20	8
Revision	18	8
Post-test	17	7

Table 1: Number of students left after each round in the two cycles.

We decided that the most interesting way to continue the experiment was to collect more data for **no-rec**; in the second cycle, we consequently started with 30 subjects, assigning all of them to the **no-rec** group. The third column of Table 1 shows the number left after each round. At the end of the cycle, we had adequate data for 12 subjects in **no-rec** and 12 in the union of the three groups which included recognition feedback, which we will call **rec**. The analysis in the next section thus focusses on exploring the difference between **no-rec** and **rec**.

5. Analysis of results

The main hypothesis we wish to investigate when comparing **no-rec** and **rec** is the obvious one: whether including recognition feedback in the application helps the student. Our basic strategy is equally obvious. For each of the two versions, we compare student performance in the pre- and post-tests. We wish to determine whether this difference is significantly larger in **rec** than in **no-rec**.

rec			**no-rec**		
ID	B-S-W	Signif	ID	B-S-W	Signif
1	1 7-13-8	—	13	6-18-3	—
2	4-14-2	—	14	4-13-1	—
3	9-6-1	$p < 0.05$	15	2-18-2	—
4	9-18-1	$p < 0.05$	16	7-15-6	—
5	8-19-0	$p < 0.02$	17	7-19-2	—
6	10-12-5	—	18	14-9-4	$p < 0.05$
7	6-12-1	—	19	18-7-3	$p < 0.01$
8	8-5-0	$p < 0.02$	20	4-22-1	—
9	6-15-3	—	21	5-15-6	—
10	5-14-9	—	22	10-15-2	$p < 0.05$
11	9-12-2	—	23	5-17-6	—
12	12-11-5	—	24	9-17-2	—

Table 2: Improvement between pre-test and post-test for **rec** and **no-rec** versions, broken down by student. "B-S-W" shows the number of prompts on which the student performed BETTER, SAME and WORSE. "Signif" gives the significance of the difference between BETTER and WORSE according to the McNemar test. Students who described themselves as beginners are underlined.

The pre- and post-tests are the same[1] and contain a total of 28 prompts (13 without help available, 15 with). We compare a given student's performance on each prompt by determining whether the system accepts the student's response or not. As already noted, this correlates reasonably with human judgements [6]. Students can get BETTER (not recognised in pre-, recognised in post-), WORSE (recognised in pre-, not recognised in post-), or stay the SAME (identical outcomes in both tests).

We can compare either across students or across prompts. The simplest way to compare across students is to take each student and count how many examples of BETTER/WORSE/SAME (B/W/S) they get. We can then look at the difference between BETTER and WORSE using the McNemar test to find how significant it is (Table 2); note that $B + S + W$ does not always total to 28, since students sometimes omitted a few items from one or both tests. The comparison turns up four students in the **rec** group who get a significant difference, against three in **no-rec**; in the right direction, but obviously not strong evidence that **rec** is better. Other more complex tests also failed to show a statistically significant difference when we compared across all students, though some were close. It is however worth noting that we do get a significant difference on the two-tail t-test ($t = 1.7$, $df = 11$, $p < 0.01$) when we use only the subset of students, underlined in the table, who described themselves as beginners.

[1] We wondered if it was methodologically sound to use the same items for the pre- and post-tests. Students were however going to take the two tests at least a week apart, during which they would practice many similar examples. We felt it was unlikely that they would remember the specific sentences from the pre-test, and that it was more important to give ourselves the option of performing a clear item-by-item comparison.

Prompt	BETTER-SAME-WORSE, score			
	rec		**no-rec**	
With help				
P1	**7-1-1**	**66.7**	4-7-1	25.0
P2	**2-8-0**	**20.0**	3-7-3	0.0
P3	1-6-0	14.3	**5-5-2**	**25.0**
P4	**1-7-0**	**12.5**	1-8-4	–23.1
P5	3-3-3	0.0	**4-7-2**	**15.4**
P6	**5-6-3**	**14.3**	5-7-4	6.2
P7	4-3-5	–8.3	2-7-3	–8.3
P8	**3-2-2**	**14.3**	2-5-3	–10.0
P9	**3-2-2**	14.3	**6-7-3**	**18.8**
P10	4-4-1	33.3	**5-5-1**	**36.4**
P11	**4-6-1**	**27.3**	3-5-4	–8.3
P12	**6-6-2**	**28.6**	4-7-2	15.4
P13	**3-4-0**	**42.9**	4-6-2	16.7
Without help				
P14	**3-8-0**	**27.3**	4-7-1	25.0
P15	**3-7-1**	**18.2**	1-10-1	0.0
P16	3-5-2	10.0	**4-6-0**	**40.0**
P17	4-4-3	9.1	3-6-2	9.1
P18	1-10-0	9.1	**2-9-0**	**18.2**
P19	5-4-1	40.0	**6-5-1**	**41.7**
P20	**2-8-0**	**20.0**	3-7-2	8.3
P21	**6-3-2**	**36.4**	3-6-2	9.1
P22	**4-6-2**	**16.7**	2-7-1	10.0
P23	**2-7-0**	**22.2**	1-9-1	0.0
P24	**3-7-1**	**18.2**	1-8-1	0.0
P25	2-7-1	10.0	**2-9-0**	**18.2**
P26	3-7-1	18.2	**2-8-0**	**20.0**
P27	**7-4-0**	**63.6**	7-5-0	58.3
P28	**3-6-0**	**33.3**	3-9-0	25.0

Table 3: Improvement between pre-test and post-test for **rec** and **no-rec** versions, broken down by prompt. The version with the larger improvement is marked in **bold**.

Comparing across prompts produces a convincing result even when we use all the students (Table 3). This time, we look at all the B/W/S scores for a given prompt and version, using the measure $(B - W)/(B + W + S)$. The value will be 100% if every example is BETTER,

zero if BETTER and WORSE are equal, and –100% if every example is WORSE.

We can now perform a prompt-by-prompt comparison of **rec** and **no-rec**, contrasting the scores. For example, looking at prompt P11, we have under **rec** $B = 4$, $S = 6$ and $W = 1$, giving a score of $(4-1)/(4+6+1) = 3/11 = 27\%$. Under **no-rec**, we have $B = 3$, $S = 5$ and $W = 4$, giving a score of –8.3%. Applying the Wilcoxon signed-rank test to the whole set of prompts, the comparison between **rec** and **no-rec** on the above measure yields a difference significant at $p < 0.02$.

6. Conclusion and discussion

We can reasonably argue that the experiment described above shows that speech recognition actually does help the student improve their speaking ability. Nonetheless, the modest size of the difference compared to the **no-rec** control group makes us rather thoughtful; in particular, some control students showed significant improvements. Unfortunately, the result leaves it uncertain whether several of our earlier experiments can be interpreted as showing that the student improvements we found need be due to any interesting properties of the system.

We find some aspects of the crowdsourced evaluation methodology attractive, but we are so far reluctant to make strong claims for it. On the negative side, there are several obvious weaknesses: the one which concerns us most is the fact that we have very little control over our subjects. When working with ordinary students, we have the opportunity to meet them, and we usually have some idea of their motivation for wanting to use the CALL tool. (For most of our experiments, we have used subjects who were already learning the language in question). Here, we recruit people randomly through a crowd-sourcing site, and their motivation is unclear. A fair number of the people who completed the course did appear to be interested in learning French: they left positive comments, and, more significantly, many of them logged sessions which were longer than the 20 minutes we required. But not all of them did this.

The nature of the recruitment process readily explains the fact that only a small proportion of the subjects (24 out of 130, or 18%) reached the end of the 8 HIT series. At the beginning, subjects had no clear picture of the level of involvement required in order to complete the course; they only understood this after they had completed the second HIT. It is unsurprising that many of them decided afterwards that they did not want to spend a substantial part of the next week learning to speak better French. The rate of attrition dropped sharply after the third HIT, when subjects knew what to expect. We could have recruited a larger pool, but were limited by financial constraints; the whole experiment cost about $750, a non-trivial amount in our context.

Despite the known problems, our current feeling is that crowdsourced evaluation is well worth further investigation, and opens up interesting new possibilities for carrying out controlled experiments in CALL; as pointed out by Jurčiček and his colleagues [16], whose experiences seem to be fairly similar to ours, the fact that it enables cheap, rapid recruitment of a diverse pool of users is worth a good deal. We expect to see other researchers experimenting with these techniques.

7. References

[1] M. Rayner, P. Bouillon, N. Tsourakis, J. Gerlach, M. Georgescul, Y. Nakao, and C. Baur, "A multilingual CALL game based on speech translation," in *Proceedings of LREC 2010*, Valetta, Malta, 2010.

[2] C. Wang and S. Seneff, "Automatic assessment of student translations for foreign language tutoring," in *Proceedings of NAACL/HLT 2007*, Rochester, NY, 2007.

[3] W. Johnson and A. Valente, "Tactical Language and Culture Training Systems: using AI to teach foreign languages and cultures," *AI Magazine*, vol. 30, no. 2, p. 72, 2009.

[4] M. Fuchs, N. Tsourakis, and M. Rayner, "A scalable architecture for web deployment of spoken dialogue systems," in *Proceedings of LREC 2012*, Istanbul, Turkey, 2012.

[5] M. Rayner, B. Hockey, and P. Bouillon, *Putting Linguistics into Speech Recognition: The Regulus Grammar Compiler*. Chicago: CSLI Press, 2006.

[6] M. Rayner, P. Bouillon, and J. Gerlach, "Evaluating appropriateness of system responses in a spoken CALL game," in *Proceedings of LREC 2012*, Istanbul, Turkey, 2012.

[7] M. Rayner, P. Bouillon, B. Hockey, and Y. Nakao, "Almost flat functional semantics for speech translation," in *Proceedings of COLING-2008*, Manchester, England, 2008.

[8] P. Bouillon, S. Halimi, Y. Nakao, K. Kanzaki, H. Isahara, N. Tsourakis, M. Starlander, B. Hockey, and M. Rayner, "Developing non-European translation pairs in a medium-vocabulary medical speech translation system," in *Proceedings of LREC 2008*, Marrakesh, Morocco, 2008.

[9] P. Bouillon, M. Rayner, N. Tsourakis, and Q. Zhang, "A student-centered evaluation of a web-based spoken translation game," in *Proceedings of the SLaTE Workshop*, Venice, Italy, 2011.

[10] B. Coyne, C. Schudel, M. Bitz, and J. Hirschberg, "Evaluating a text-to-scene generation system as an aid to literacy," in *Speech and Language Technology in Education*, 2011.

[11] R. Poulsen, *Tutoring Bilingual Students With an Automated Reading Tutor That Listens: Results of a Two-Month Pilot Study*. DePaul University, Chicago, IL: Masters Thesis, 2004.

[12] K. Reeder, J. Shapiro, and J. Wakefield, "The effectiveness of speech recognition technology in promoting reading proficiency and attitudes for Canadian immigrant children," in *15th European Conference on Reading*, 2007.

[13] G. Korsah, J. Mostow, M. Dias, T. Sweet, S. Belousov, M. Dias, and H. Gong, "Improving child literacy in Africa: Experiments with an automated reading tutor," *Information Technologies and International Development*, vol. 6, no. 2, pp. 1–19, 2010.

[14] J. Kulik, C.Kulik, and P. Cohen, "Effectiveness of computer-based college teaching : A meta-analysis of findings," *Review of Educational Research*, vol. 50, pp. 177–190, 1980.

[15] C. H. Kennedy, *Single-Case Designs for Educational Research*. Allyn and Bacon, 2005.

[16] F. Jurčiček, S. Keizer, M. Gašić, F. Mairesse, B. Thomson, K. Yu, and S. Young, "Real user evaluation of spoken dialogue systems using Amazon Mechanical Turk," in *Proceedings of Interspeech 2011*, Florence, Italy, 2011.

An Automatic Feedback System for English Speaking Integrating Pronunciation and Prosody Assessments

Jeesoo Bang, Sechun Kang, Gary Geunbae Lee

Department of Computer Science and Engineering
Pohang University of Science and Technology, South Korea
{jisus19, freshboy, gblee}@postech.ac.kr

Abstract

We have proposed a computer-assisted language learning (CALL) system, called Postech English Speaking Assessment and Assistant (PESAA) for non-native English learners, especially Koreans, to improve their overall language skills. PESAA is an automatic feedback system for speaking English that integrates both pronunciation and prosody assessments. The system has three error-feedback modules: the pronunciation, rhythm, and phrase break error-feedback modules. PESAA generates scores on each of three modules as well as a combined score. The pronunciation assessment gives feedback that is based on comparing canonical and actual phoneme alignment results. The rhythm and phrase break assessments give feedback that is based on comparing predictions with detected results. English learners can use PESAA to practice pronunciations, rhythm and phrase breaks by themselves. We evaluated PESAA in three different ways: accuracy, correlation of the system's assessment with human assessments, and user satisfaction from the expected learning effectiveness and user interface. The evaluation showed that PESAA could work as a CALL system well, with good accuracy and positive learning results for the users.

Index Terms: computer-assisted language learning, CALL, language assessment, error feedback

1. Introduction

Foreign language learners must practice a target language frequently and repeatedly to learn the language efficiently and rapidly. Computer-assisted language learning (CALL) systems offer an advantage in foreign language training outside of the classroom. They meet the language learners' need for practice by providing a private and stress-free environment in which to train comfortably and conveniently, within the bounds of the learners' schedules and circumstances.

CALL systems can provide instructional materials for pronunciation, grammar, vocabulary, and other areas. Pronunciation is a difficult skill to acquire alone, unlike memorizing grammar or vocabulary. Serious pronunciation problems can even hinder communication and degrade the intelligibility of speech [1]. To learn a foreign language's pronunciation accurately, language learners should receive appropriate and timely feedback. In addition to pronunciation, prosody is an important skill to acquire in learning languages. Lexical stress, intonation and rhythm help a listener to understand utterances more accurately [2, 3, 4]. Previous research has shown that training in prosodic features is more effective in improving intelligibility than teaching only in the segmental features [5, 6]. However, very few systems address prosodic features, whereas there have been a number of studies on pronunciation training systems. Moreover, there has been little focus on CALL systems that combine segmental or pronunciation features with supra-segmental or prosodic features.

A significant amount of research related to CALL systems has been conducted [7, 8, 9, 10]. The authors in [7] have developed a CALL system that specialized in English pronunciation training for Japanese English learners. The system provides general feedback for a learner through a reading session of role-playing dialogues. The system gives feedback for phonemes, which Japanese learners occasionally provide. The authors in [8] have developed a hand-held pronunciation evaluation device that uses the log-posterior probability of the expected phoneme sequence from automatic speech recognition (ASR) to assess the pronunciation accuracy and that shows the pitch contour to help correct the intonation. The author in [9] proposes a strategy of modeling the pronunciation variation at the syllable level using different subsets of context features. The authors in [10] present a system that assesses spoken English for call center agents with multiple parameters: articulation of sounds, correctness of lexical stress in words and spoken grammar proficiency.

Most of the systems described above focus mainly on pronunciation evaluation, and few of them focus on evaluating syllable stress. We have developed a system called PESAA that integrates pronunciation training with prosody training. PESAA has three assessment aspects: pronunciation, rhythm and phrase breaks. For a single user utterance, each part delivers its assessment result independently of the other parts. The result comprises detailed feedback on phonemes, words or breaks in addition to 0-to-100 scores for each part and for the overall system. PESAA also provides a specific view of each assessment modules, which are the pronunciation, rhythm and phrase break, to help learners practice English intensively.

2. System architecture

PESAA is designed as a client-server model, to allow the learners to learn English wherever they want. PESAA has two main components, the user interface and speech processing.

2.1. User interface

The user interface component records a user's utterances and shows the feedback and evaluation result of the recorded utterances. This component displays sentences to be recorded for evaluating the pronunciation and prosody; it provides prompts to be shown back to the users with error feedback, and provides assessment results of recorded utterances, by recording a user's utterances and playing back the recorded utterances. This

component also involves some amount of speech processing, to ensure that the candidate's speech is recorded at an appropriate volume level and to warn the user in case the recorded speech is too short or too long, after which the times can be modified. Users can select the sentence level and the sentence that they prefer to practice, or they even can type the sentences that they want to practice by themselves.

2.2. Speech processing

The speech processing component, which resides on the server, uses the speech recognition engine to obtain the phoneme alignments and the confidence scores. The speech processing component has three main modules: pronunciation, rhythm and phrase break assessment modules. Each module gives detailed feedback on phonemes, rhythms and phrase breaks with 0-to-100 scores as well as combined scores. The detailed module architecture and score computation are described in Section 3 and Section 4.

3. Module architecture and implementation

3.1. Pronunciation error feedback

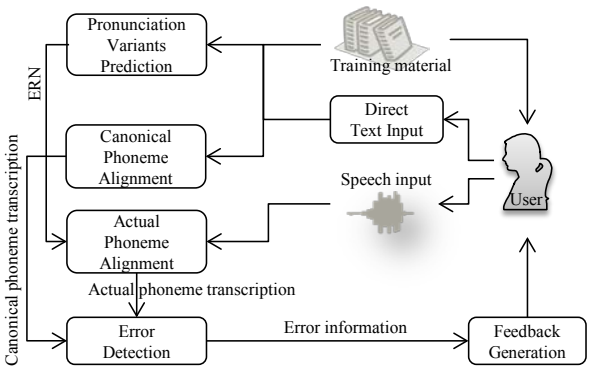


Figure 1: *Architecture of pronunciation error feedback.*

We focused on the differences between Korean and English because PESAA was developed for Korean learners of English. The Korean phoneme inventory differs largely from the English phoneme inventory. For example, Korean does not have the sounds /F/, /V/, /TH/, and /Z/ [11]. Thus, the erroneous phonemes should be pointed out, to train learners in pronunciation effectively.

We built a pronunciation error feedback module to detect the phoneme errors of the input utterances and to provide feedback to the learners. The pronunciation module has four main components: actual phoneme alignment, canonical phoneme alignment, error detection and feedback generation (Fig. 1). Actual phoneme alignment uses an extended recognition network (ERN) [12] that is generated by the pronunciation variants prediction model.

3.1.1. Data

We collected English-reading speech data spoken by Korean English speakers who had high oral proficiency levels. We distributed 6,600 text segments to 170 Korean people learning English; the text segments consist of sentences, words, and phrases. We collected 12,000 speech segments; each person recorded an average of 74 text segments, and two different people recorded each text segment. Two annotators trained in phonetics and phonology curriculum annotated phoneme-level transcriptions of the collected data. The annotators were provided with the automatic phoneme-level transcription that corresponded to the text segment read; they revised the transcription, repeatedly listening to the recording, if necessary. The phoneme-level transcription is represented in ARPAbet [1] symbols. The annotators' phoneme level agreement had a Fleiss' kappa value of 0.8685 and had an 86.90% agreement on 9,327 phonemes from 498 sentences. [13]

3.1.2. Mispronunciation detection and feedback generation

The first component, actual phoneme alignment, outputs a phoneme sequence given speech input and text. The actual phoneme decoder uses ERN instead of an unlimited phoneme loop, which means all possible phoneme sequences. The phoneme alignment on the unlimited phone loop tends to be slow because of the large search space or erroneous because of the pruning. We reduced this problem by replacing the unlimited phoneme loop with the ERN that was generated by our phoneme variant prediction method [14]. The ERN has possible variations of English phonemes that are generally used by Korean speakers of English. Using ERN can reduce the overall phoneme prediction error, and the actual phoneme recognizer can output reliable phoneme decoding results in a shorter amount of time.

The second component, the canonical phoneme decoder, outputs a phoneme sequence given speech input and the text as the actual phoneme alignment. A canonical pronunciation is a reference pronunciation that is in the pronouncing dictionary. We used the CMU pronouncing dictionary [2] to select the correct phoneme transcript for the speech input.

The third component, pronunciation error detection, detects mispronunciations using logistic regression. We used logistic regression because the regression function outputs a real-valued score between zero and one, which can be directly mapped into the decision probability; this probability can be used for CAPT applications. Additionally, the feature functions that were designed are proportional to the mispronunciation probability. Furthermore, it is easy to analyze or control the contribution of each feature using its weight.

The last component, the feedback generation, provides learners' pronunciation assessment results for input utterances. This component decides whether to provide positive feedback or negative feedback. The feedback generation component compares the confidence of the canonical and actual phoneme alignment results that were generated by canonical and actual phoneme recognizers. The detailed process of determining the feedback decision boundary is discussed in Section 4.

[1] The ARPAbet symbol list used in this work can be found in the CMU pronouncing dictionary.
Also, http://en.wikipedia.org/wiki/Arpabet provides a parallel representation of ARPAbet to IPA symbols.

[2] Available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict, version 0.7a is used in this work

3.2. Rhythm error feedback

The rhythm error detection and feedback is composed of three main parts: a prediction part, a detection part, and a feedback part (Fig. 2).

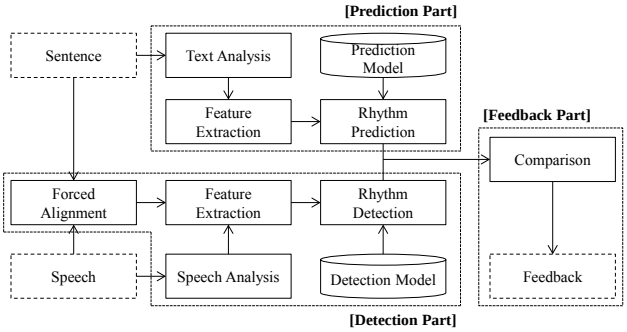


Figure 2: *Architecture of rhythm error feedback.*

From the input sentence, the prediction tool generates a rhythm pattern. The input sentence is analyzed by a part-of-speech tagger (text analysis), and the machine learning features are extracted from the tagged words. From the input speech, the detection part identifies which words are stressed. The input speech is analyzed to extract acoustic parameters (speech analysis). To provide corrective feedback to learners, a comparison is performed between the predicted and detected rhythm patterns. A word gives positive feedback (the sign "O") or negative feedback (the sign "X") depending on whether it is correctly or incorrectly uttered with rhythm, respectively. The color of the words represents the rhythmical degree; the closer to the color red, the word is full-stressed and has a full-length vowel; if a word color is black, then the word is not stressed and has reduced vowels, similar to function words.

To build rhythm prediction and detection models, we utilized two types of data: Aix-Machine Readable Spoken English Corpus (Aix-MARSEC) [15] and Korean learners' English accentuation corpus (KLEAC) [16]. We used Aix-MARSEC corpus to train the rhythm prediction model. Jassem's narrow rhythm unit (NRU) notation groups words into rhythmic phrases, and we have considered a stressed syllable that appears in each NRU for the first time to be a rhythmic word.

KLEAC is used for building the detection model, because non-native English learners have their own prosodic habits when uttering English sentences. To increase the accuracy of the rhythm detection in the learners' utterances, we adapted acoustic characteristics of non-native learners to the detection model for rhythm. The KLEAC is composed of six hours of speech with 5,500 English sentences produced by 75 native Korean speakers, including orthographic transcription, rhythmic word marks and proficiency labels. The rhythm labels were manually annotated by five phonetic experts. The inter-annotator agreement for rhythms is 87.1% in the KLEAC corpus.

While content words tend to be stressed, function words tend to be unstressed. Several rules for the prediction of rhythmic words were constructed based on the following assumptions: content words are rhythmic words, function words are not rhythmic words, negative auxiliary verbs are rhythmic, and in a sequence of verbs, only the last verb is to be rhythmic. Each rule has its own precedence, to avoid conflicts. However, these rules are not perfect, and erroneous linguistic analysis results can degrade the performance. To compensate for the limitations of the rule-based method, an approach based on machine learning is adopted.

Under the machine learning framework, the existence of rhythm notation for each word can be regarded as a label, and encoded rules with other useful information for classification can be represented into feature vectors. Rhythm labels are affected by surrounding labels, but there is no significant long-distance relationship between them. To reflect these characteristics, we adopted the linear-chain CRF model, which has been widely used in the natural language processing fields [17, 18, 19]. The machine learning features used for the prediction of rhythm are the rules that are mentioned above, part-of-speech tags, and words.

The detection model also adopts the linear-chain CRF model as the prediction model. The detection model utilizes a KLEAC corpus that is different from the prediction model, which utilizes the Aix-MARSEC corpus. This choice was made because the detection model should detect the acoustically unique characteristics of Korean speakers of English. We extracted machine learning features from the acoustic parameters on the middle of the vowel that had the primary word stress. The features are pitch, intensity, duration, and phonetic value.

The corrective feedback for learners is determined by the comparison of predicted and detected rhythm patterns, which is categorized into the following three groups: positive feedback (the sign "O"), negative feedback (the sign "X"), and no feedback (no sign). For the measure that decides the confidence in the feedback, the adjusted score was designed by adopting the output probability of the CRF classifier for each stress label. The adjusted score is calculated by the absolute difference between the probabilities of the predicted and detected rhythm.

3.3. Phrase break error feedback

The phrase break variation detection and feedback is designed to provide appropriate phrase break feedback when a learner utters a given reference sentence. To achieve this goal, the proposed system comprises prediction, detection and feedback provision parts as the rhythm error detection and feedback. If a reference sentence is given without additional information, then the prediction part determines the appropriate breaks, at which phrase breaks can appear with high probabilities. If a learner utters a reference sentence with a microphone connected to the phrase break system, then the detection part determines the positions at which the phrase breaks of the given utterance are imposed with high probabilities. Based on the determined break positions and the probabilities of the prediction and detection parts, the feedback provision part shows positive, negative or no sign at each juncture.

The overall architecture of the phrase break variation detection and feedback is the same as the rhythm variation detection and feedback. However, the data and features are slightly different from the data and features of the rhythm.

We used the Boston university radio news corpus (BURNC) [20] annotated with prosodic markers called Tones and break indices (ToBI) [21]. The ToBI framework is one of the most popular schemes for representing the annotation of prosodic events in an utterance. The ToBI break labeling uses indices between 0 and 4 to represent the disjuncture between successive words. In this paper, we handle the ToBI break indices with the coarse mapping manner, which groups the break indices 3 and 4

into the presence of breaks and the other indices into the absence of breaks.

We used the BURNC corpus for building both the prediction model and the detection model. To achieve an accurate phrase break prediction, we adopt a linear-chain CRF classifier [22]. We use syntactic and lexical features from the given reference sentence: word identity, POS tag, word class, number of syllables/vowels, and punctuation marks. Similar to in the prediction model, the detection model uses the CRF classifier to detect the phrase breaks from the learners' utterances. The acoustic features are required to distinguish phrase breaks from utterances. To achieve the high accuracy of phrase break detection, syntactic and lexical features are required in addition to acoustic features [17, 19]. The detection model utilizes the features: duration, pitch mean, and intensity mean of the last syllable of each word, silence after the word and the features that are in the prediction model.

4. Error feedback-based scoring

4.1. Feedback proficiency assessment

Feedback is an important point of the proposed system, and it is important because of its pedagogical effect. The learners should receive appropriate and understandable feedback for each of their utterances, to improve their language skills. The proposed system gives feedback that is based on comparing the predicted and detected results for the pronunciation, rhythm and phrase break. As the feedback systems for all of the three parts are the same, we represent the canonical result as the predicted result and the actual result as the detected result in this section.

The feedback system gives three types of feedback: positive, negative, and ambiguous. If the detected result that is extracted from a learner's utterance is close to the predicted result, as a reference standard, we give positive feedback. If the result is far away from the reference result, we give negative feedback. Otherwise, we give no feedback. For educational purposes, the positive feedback helps to motivate learning, and the negative feedback helps to correct the mistakes of learners. Incorrect feedback, such as false positives and false negatives, however, adversely affect the reliability of the learning system and the learning motivation, also. Therefore, if the comparison result is not trustworthy, then feedback will not be provided.

The feedback decision equation is

$$\text{Feedback} = \begin{cases} \text{Positive, if } | \pi_{\text{pre}} - \pi_{\text{det}} | < \theta_1 \\ \text{Ambiguous, if } \theta_1 \le | \pi_{\text{pre}} - \pi_{\text{det}} | \le \theta_2 \\ \text{Negative, if } | \pi_{\text{pre}} - \pi_{\text{det}} | > \theta_2 \end{cases} \quad (1),$$

where π_{pre} and π_{det} are output probabilities that are hypothesized by the prediction and detection models, respectively, and θ_1 and θ_2 are the decision boundaries for the feedback signs.

We conducted experiments in which the feedback results of PESAA were compared to human ratings to compute the correlation coefficients, because the threshold values θ are not determined and can vary from 0 to 1. Therefore, regarding the threshold values, we calculated every correlation coefficient value to determine the optimum thresholds and the best correlations in the respective parts.

We calculated correlation coefficients by using the method of the Pearson product-moment correlation coefficient. The human ratings and speech resources used in the experiments are derived from the KLEAC, which provides overall pronunciation and fluency ratings for 75 persons, as assessed by five English experts. We obtained the values $0.49(\theta_1)$ and $0.50(\theta_2)$ for pronunciation, $0.13(\theta_1)$ and $0.41(\theta_2)$ for the rhythm, and $0.2(\theta_1)$ and $0.7(\theta_2)$ for the phrase break at the best correlation.

4.2. Scoring method

We assessed the utterances on a scale of 1 to 100, using the feedback information. The scoring equation is

$$Score = \frac{Feedback\ count\ of\ "positive"\ type}{Entire\ feedback\ count\ except\ "ambiguous"\ type} \quad (2).$$

Each module gives its score, and the combined score is the average score of the three module scores.

5. Experiments and results

The proposed system is measured in three different areas: its accuracy, the correlation of the assessment results with human assessments, its user satisfaction on expected learning effectiveness and its user interface (UI). These measures evaluate the proposed system in different ways and infer the usability and the appropriateness of the system as an effective CALL system.

5.1. Prediction and detection accuracy

5.1.1. Pronunciation module accuracy

Table 1. *Accuracies, precisions, recalls and F1-scores (in percentages) of the pronunciation module*

Phoneme recognition		unlimited	simulated
SA		75.6	82.4
Correct pronunciation	P_c	92.2	91.6
	R_c	78.8	87.8
	F_c	84.9	89.7
Mispronunciation	P_e	26.3	34.3
	R_e	53.1	44.2
	F_e	35.2	38.6

There is no point in calculating the prediction accuracy, because in the pronunciation assessment, the canonical phoneme sequence is generated from the pronunciation dictionary. Thus, we did not calculate the prediction accuracy in the pronunciation module; instead, we calculated only the detection accuracy.

We used the data described in Section 3.1.a, which has 12,000 English speech segments from 170 Korean speakers. The segments consist of words, phrases, and sentences. We generated ERN to limit the phoneme search space with speech segments of 136 speakers, which are 80% of the data. We used a random split n-fold cross validation to calculate the detection accuracy. Additionally, we compared the detection accuracy with and without ERN to verify whether limiting the phoneme search

space is effective or not. We measured the performance with a scoring accuracy (SA) [23], which represents the overall correctness of the decisions. In addition, we computed the precision *P*, the recall *R*, and the F1-score *F*. We computed the precision, recall, and F1-score for each of the mispronunciation and correct pronunciation labels.

The mispronunciation detection method that uses simulated phoneme recognition achieved better results than the method that is based on the unlimited phone loop (Table 1). The accuracy is not very high compared to the F1-score of the correct pronunciation detection. The mispronunciation detection rate appears to be low; however, compared to the recent mispronunciation detection work [24], it is not a low number (precision: 32.7%, recall: 62.7%, F1-score: 42.1). Of course, we cannot directly compare the two values, because the used corpora are different and L1-speakers are also different.

5.1.2. Rhythm module accuracy

We trained the rhythm prediction and detection models with the Aix-MARSEC and KLEAC corpora, respectively. We measured the precisions, recalls and F1-scores of our models as well as the accuracies using the five-fold cross validation. The numbers indicate how exactly the proposed models can predict and detect the rhythm in given sentences.

Table 2. *Accuracies, precisions, recalls and f1-scores (in percentage) of the rhythm module*

Models	Accuracy	Precision	Recall	F1-score
Prediction	96.6	98.3	96.1	97.2
Detection	81.2	84.0	85.6	84.8

The proposed work demonstrates that the precision, recall and F1-score values in our models are sufficient to predict and detect rhythms (Table 2). The accuracies of the prediction model and detection model are 96.6% and 81.2%, respectively, with higher F1-scores: 97.2% and 84.8%. The prediction model appears to be the most accurate when it is comparing to the detection model's accuracy or F1-score. Although the values for the detection model are lower than the values for the prediction model, the numbers are not low values in the classification task when using machine learning methodologies; in the proposed method, the machine learning framework is the CRF.

5.1.3. Phrase break module accuracy

We evaluated the prediction model and the detection model of the phrase break module using five-fold cross validation. We used the BURNC corpus for both the prediction and detection model.

Table 3. *Accuracies, precisions, recalls and f1-scores (in percentage) of the phrase break module*

Models	Accuracy	Precision	Recall	F1-score
Prediction	90.2	84.0	80.1	82.0
Detection	91.6	86.1	83.7	84.9

We measured the precision, recall and F1-scores of the models as well as the accuracy (Table 4). To measure the accuracy and F1-score, we used binary classification. The accuracy, precision, recall and F1-score values are considered to be sufficient to accurately predict and detect phrase breaks in our classification tasks.

5.2. Automatic assessment

We compared the assessment results of PESAA with human assessments. We used KLEAC data to compute the correlation, because the KLEAC corpus was annotated with proficiency labels. The KLEAC corpus has three types of proficiency labels that five phonetic experts marked: pronunciation, fluency, and overall level. The experts labeled each criterion with a scale of 1 to 5, where 1 is the lowest score and 5 is the highest score. The pronunciation criterion involves accurate pronunciation; the fluency criterion involves the rhythm, break, speed and intonation; the overall criterion is how much the utterances are native-like or not.

The correlation between the pronunciation scores of the system and human assessments was 0.41. The correlation itself is not high; however, the value is not small compared to the results of a previous study [25], which was 0.41. This value represents that the pronunciation feedback by the system is not similar to a human's assessment. The previous experiments of the pronunciation error detection showed quite good results (82.4% of accuracy); thus, we can infer that the assessment criteria of the pronunciation part are not the same as a human's assessment criteria.

The correlation values between the prosody score of the system and the fluency score are 0.64 for rhythm and 0.74 for phrase break. We can show that we can judge a person's fluency in English with only the prosody features. This capability means that the prosodic features are important in language learning, especially in language speaking learning.

5.3. User satisfaction

We evaluated the user satisfaction in two ways: feedback satisfaction and UI satisfaction.

5.3.1. Feedback

We designed an experiment for eight Korean university students to utilize the system, to evaluate whether PESAA can be applied in real learning. The students utilized PESAA for an hour, and we collected questionnaires for expected learning effectiveness on a scale of 1 to 5, with positive adjectives anchoring the high end and negative adjectives anchoring the low end.

Table 4. *Learners' questionnaires for expected learning effectiveness on a scale of 1 to 5.*

Questions	Mean	s.d.
Pre-test		
Do you know about pronunciation, rhythm, and phrase break? (average)	2.79	0.40
Do you care about pronunciation, rhythm, and phrase break? (average)	3.04	0.68
Post-test		
Does the system help you to understand about pronunciation, rhythm, and phrase break? (average)	3.25	1.05
Does the system point correctly to your problems in pronunciation, rhythm, and phrase break? (average)	3.67	0.68
Does the system help you to improve English proficiency?	3.75	0.19
Are you interested in using this system to improve your English skills?	3.63	0.98

Table 4 shows that the students were not aware of pronunciation, rhythm, or phrase breaks very much (2.79). However, they cared about their English when they were speaking (3.04). These findings mean that the students attempted to speak English intelligibly, even if they did not completely understand the English pronunciation or prosody. After the use of the proposed system, they answered that they understood pronunciation, rhythm, and phrase breaks. The effectiveness of our system was deemed to be good, with many of the students answering that our system helped them to improve their English proficiency (3.63). Additionally, the students answered that they would use the system to improve their English skills. Two of them absolutely agreed (5) with the question, and all of the others were positive at using the system (3 and 4), except for only one user (2).
The students appeared to obtain knowledge about pronunciation and prosody, even when the time that they were on the system was as short as an hour. The students were satisfied with our system and reported some improvement in their English skills or English knowledge.

5.3.2. Questionnaire for user interaction satisfaction

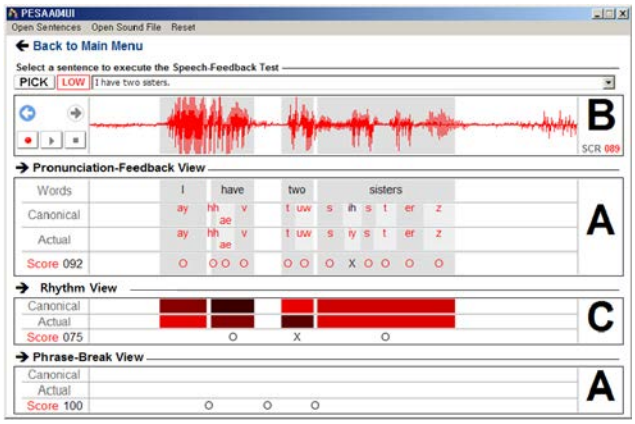


Figure 3: *A screen capture of PESAA*

As a qualitative evaluation, the subjective feelings of the testers were surveyed with a questionnaire for user interaction satisfaction (QUIS) style usability evaluation. The QUIS of [26] was created to gauge the satisfaction aspect of the software usability in a standard, reliable, and valid way. QUIS focuses on the user's perception of the usability of the interface as it is expressed in specific aspects of the interface (i.e., overall reaction to the system, screen factors, terminology and system feedback, learning factors, and system capabilities). Each of the specific interface factors and optional sections has a main component question followed by related sub-component questions. Each item is rated on a scale from 1 to 5, with positive adjectives anchoring the high end and negative adjectives anchoring the low end.

We used the short form of QUIS 5.0 to evaluate the usability of PESAA (Fig. 3). The sections of original QUIS were retained, but some items that were not appropriate were dropped. To evaluate PESAA, we designed an experiment for 18 Korean university students attending an English class who utilized the implemented CALL system as a tool for the class, to aid in learning.

Table 5. *Learners' questionnaires for user interface satisfaction on a scale of 1 to 5 for PESAA. Each major item comprises 4-6 detailed questions.*

Major items for satisfaction	Mean	s.d.
Overall system	3.72	0.51
Display content	3.54	0.51
Terminology	3.48	0.50
Easy use	3.72	0.36
Processing speed	3.55	0.56
Total	3.60	0.49

We asked the experiment participants to utilize the system for 30 minutes per a day over three weeks. After the three weeks, we collected QUIS answering sheets from the participants. According to the learners' QUIS, the overall satisfaction score on 1-5 was 3.60, with a standard deviation (s.d.) of 0.49 (Table 5). This score can be considered to mean that PESAA is meaningfully useful in real English learning and that it has an appropriate user interface that helps the learners to understand their English assessment results easily.

6. Discussion and conclusions

In this paper, we described a computer-assisted language learning system for non-native English learners, especially Koreans, to improve their overall language skill. We designed a system that can assess a learner's pronunciation and prosody. The system has three error feedback modules: pronunciation, rhythm, and phrase break error feedback modules. Each module has a prediction and detection model, and when comparing their results, the module generates appropriate assessments with regard to feedback of errors.

Through three types of evaluation, we found PESAA to be a useful CALL system. PESAA effectively predicts and detects pronunciation and prosody features. The accuracy of each module is more than 80%. Additionally, PESAA assesses a learner's utterances more similar to the way that a person would assess a learner's utterances. The correlation between human annotators of fluency and the assessment results from PESAA reached 0.58 on average. The learners that used our system

reported that the system helped them to improve their English skills and that the UI was satisfactory.

7. Acknowledgements

This work was supported by the Industrial Strategic technology development program 10035252, Development of dialog-based spontaneous speech interface technology on mobile platform funded By the Ministry of Trade, industry & Energy(MI, Korea). This work was supported by the MKE (The Ministry of Knowledge Economy), Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H0503-1201-1002).

8. References

[1] S. J. van Wijngaarden, "Intelligibility of native and non-native Dutch speech," *Speech communication,* vol. 35, pp. 103-113, 2001.

[2] A. Cutler, "Stress and accent in language production and understanding," *Intonation, accent and rhythm: studies in discourse phonology,* vol. 8, pp. 76-90, 1984.

[3] J. Anderson-Hsieh, R. Johnson, and K. Koehler, "The Relationship Between Native Speaker Judgments of Nonnative Pronunciation and Deviance in Segmentais, Prosody, and Syllable Structure," *Language learning,* vol. 42, pp. 529-555, 1992.

[4] A. Cutler and S. Butterfield, "Rhythmic cues to speech segmentation: Evidence from juncture misperception," *Journal of Memory and Language,* vol. 31, pp. 218-236, 1992.

[5] A. R. Elliott, "On the teaching and acquisition of pronunciation within a communicative approach," *Hispania,* pp. 95-108, 1997.

[6] T. M. Derwing and M. J. Rossiter, "The Effects of Pronunciation Instruction on the Accuracy, Fluency, and Complexity of L2 Accented Speech," *Applied Language Learning,* vol. 13, pp. 1-17, 2003.

[7] Y. Tsubota, M. Dantsuji, and T. Kawahara, "Practical use of autonomous English pronunciation learning system for Japanese students," in *InSTIL/ICALL Symposium 2004*, 2004.

[8] K. You, H. Kim, H. Chang, J. Lee, and W. Sung, "A handheld english pronunciation evaluation device," in *Consumer Electronics, 2005. ICCE. 2005 Digest of Technical Papers. International Conference on*, 2005, pp. 267-268.

[9] E. Fosler-Lussier, "Contextual word and syllable pronunciation models," in *Proceedings of the 1999 IEEE ASRU Workshop*, 1999.

[10] A. Chandel, A. Parate, M. Madathingal, H. Pant, N. Rajput, S. Ikbal, O. Deshmukh, and A. Verma, "Sensei: Spoken language assessment for call center agents," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*, 2007, pp. 711-716.

[11] L. Cheng, "Assessing Asian language performance," *Academic Communication Associates, Oceanside, CA,* 1991.

[12] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training," in *Proc. of the 2nd ISCA Workshop on Speech and Language Technology in Education*, 2009.

[13] H. Ryu, K. Lee, S. Kim, and M. Chung, "Improving transcription agreement of non-native English speech corpus transcribed by non-natives," in *Speech and Language Technology in Education*, 2011.

[14] J. Lee, J. Bang, M. Chung, S. Kim, and G. G. Lee, "A Pronunciation Variants Prediction Method for ASR-based Mispronunciation Detection," *Computer Speech and Language,* submitted.

[15] P. Roach, G. Knowles, T. Varadi, and S. Arnfield, "Marsec: A machine-readable spoken English corpus," *Journal of the International Phonetic Association,* vol. 23, pp. 47-53, 1993.

[16] H. Lee, "Evaluation of Korean Leraners' English Accentuation", Keynote Address, in *Proc. of the 16th National Conference of the English Phonetic Society of Japan and the Second International Congress of Phoneticians of English*, 2011.

[17] J. H. Jeon and Y. Liu, "Automatic prosodic events detection using syllable-based acoustic and syntactic features," in *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, 2009, pp. 4565-4568.

[18] Y. Qian, Z. Wu, X. Ma, and F. Soong, "Automatic prosody prediction and detection with Conditional Random Field (CRF) models," in *Chinese Spoken Language Processing (ISCSLP), 2010 7th International Symposium on*, 2010, pp. 135-138.

[19] V. Rangarajan Sridhar, S. Bangalore, and S. S. Narayanan, "Exploiting acoustic and syntactic features for automatic prosody labeling in a maximum entropy framework," *Audio, Speech, and Language Processing, IEEE Transactions on,* vol. 16, pp. 797-811, 2008.

[20] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The Boston University radio news corpus," *Linguistic Data Consortium,* 1995.

[21] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, C. Wightman, P. Price, J. Pierrehumbert, and J. Hirschberg, "ToBI: A standard for labeling English prosody," in *Proceedings of the 1992 international conference on spoken language processing*, 1992, pp. 867-870.

[22] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," 2001.

[23] S. Kanters, C. Cucchiarini, and H. Strik, "The Goodness of Pronunciation algorithm: a detailed performance study," *Proceedings of SLATE,* 2009.

[24] W.-K. Lo, S. Zhang, and H. Meng, "Automatic derivation of phonological rules for mispronunciation detection in a computer-assisted pronunciation training system," in *Proc. of Interspeech*, 2010.

[25] S.-Y. Yoon and S. Bhat, "Assessment of ESL learners' syntactic competence based on similarity measures," in *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2012, pp. 600-608.

[26] J. P. Chin, V. A. Diehl, and K. L. Norman, "Development of an instrument measuring user satisfaction of the human-computer interface," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, 1988, pp. 213-218.

Visual Approach to Speech Sounds

Haruko Miyakoda

[1] Department of English, Tsuda College, Tokyo, Japan

miyakoda@tsuda.ac.jp

Abstract

Many people often struggle to master the pronunciation of foreign languages without much success. One of the reasons why L2 learners are not successful is because teaching pronunciation in the classroom is usually marginalized. With the advent of computers, this problem may partially have been overcome, due to the fact that many different types of systems and software for autonomous learning have been developed, allowing learners to improve their pronunciation skills outside the classroom. However, there are few, if any, systems and software that can present a form of visual feedback that allows learners to actually understand what their problems are.

In this paper, we present the auditory-visual pronunciation system that we have developed. One of the key features of this system is that it employs easy-to-understand visuals of the speech organ that can be seen from different angles. In addition, the internal organs can also be presented by changing the mode to transparent. Furthermore, movement of the speech organs can freely be adjusted by the instructors so that the learner's movements (especially the deviant) can be highlighted by comparing them with those of the model samples.

Index Terms: pronunciation system, speech organs, visual feedback

1. Introduction

There is general agreement in the literature that teaching pronunciation is given the least attention in many English language classrooms [1]. One of the reasons why pronunciation is marginalized in this matter is because most instructors think that the goal of phonological instruction in the classrooms should be attaining reasonable intelligibility, rather than native-like pronunciation (e.g. [2], [3]). In other words, instructors would rather spend time dealing with other aspects of language learning such as grammar since advancement in pronunciation skills can just be a waste of time. This holds true especially for late L2 learners. For example, researchers such as Scovel have argued that the so-called critical period for the acquisition of L2 pronunciation exists, and that learners who start to learn a new language after a certain age will never be able to attain native-like pronunciation [4]. However, it is also true that pronunciation is by far the most perceptible aspect in distinguishing a non-native speaker from a native one, and major difficulties in pronunciation often results in the learners facing difficulty in finding employment [1] (or, to put it the other way around, better opportunities lie ahead for those who have acquired better pronunciation skills). This may be why pronunciation has recently regained importance with some university centers, and institutions such as the British Council are also actively promoting the teaching of pronunciation skills in their business courses [5, 6].

2. The role of visuals in language learning: focus on vocabulary learning

In developing a pronunciation system, one of the first decisions that need to be made is what kind of information to include in the system. Of course, no one can deny the use of auditory data for this type of system, but our main concern was whether in addition to auditory data, was there the need to take visual data into consideration? In order find out the role that visual data plays in language learning, we conducted an experiment that compared the effectiveness of different types of learning materials in vocabulary learning.

In the field of vocabulary learning, many different results have been obtained on which factor plays a significant role. For example, we can find many studies that support the effectiveness of visual factors, but when the learning effects of movies and still images were compared, movies were favored over the latter in some studies [7], while others claim the effectiveness of still images [8]. Since incompatible results based on different experiments have been reported in the literature, we conducted our own experiment by designing and developing four systems.

In order to find out what factor leads to effective learning, we conducted an experiment that compared vocabulary learning based on the following four methods: 1) learning the words with game-oriented activity (System 1); 2) learning the words within contexts (System 2); 3) learning the words with their pronunciations (System 3); and 4) learning the words with their image data (System 4) [9], [10]. We compared the effectiveness of the systems by conducting a vocabulary memorization experiment that consisted of three tests over 11 weeks. The first test was carried out just after the exercise, the second one after 2 to 3 weeks, and the third one after 10 to 11 weeks. 11 undergraduate students attending a university in Tokyo participated in the experiment.

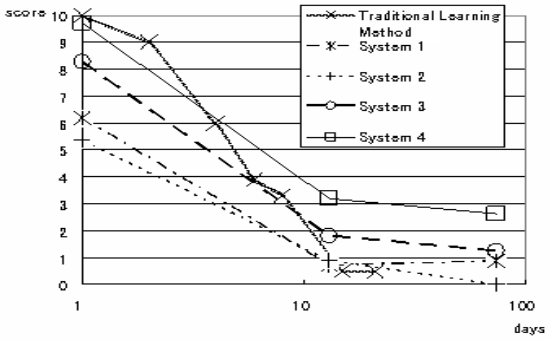


Figure 1. *Comparison of test results of different methods of learning*

The graph in Figure 1 summarizes the test results of Systems 1 through 4 depicted above together with the result

obtained for the traditional so-called "paper-and-pencil" learning method. The results demonstrate that learning by System 4 (i.e. the system using visual data) is the most effective in the long run.

We also conducted six patterns of the t-tests between the system scores of each pair in order to determine whether a significant difference can be observed. The result indicated that the learning effect is highest when both auditory and visual information are employed in the learning process. Based on this finding, we decided that employing these two types of information in our pronunciation system would also lead to better understanding of the pronunciation of foreign languages.

3. The visual element in pronunciation

The findings obtained from our experiment on vocabulary learning seem to support the viewpoint that visual data does indeed play a role in language learning. In this section, we briefly review some of the research findings concerning the use of visual element in pronunciation.

Research in the field of phonology has long been dominated by a focus on the auditory aspect. However, in actual face-to-face communication, a significant source of information about the sounds a speaker produces comes from visual cues such as lip movements [11, 12]. There are some studies that have reported that the information value of visual cues can be improved with training. In one study, hearing-impaired adults were trained in visual consonant recognition. After a total of 14 hours of training, the accuracy rate for the recognition of consonants showed dramatic improvement. For example, in the recognition of /r/, which was the most improved of all the consonants trained, the accuracy is reported to have increased from 36.1% to 88.6% after the training [13].

Although studies on the potential benefits of auditory-visual speech training for L2 learners has only recently started to gain focus, the importance of lip shapes as beneficial cues has long been recognized by language instructors in teaching English as a second language. For example, a study claiming that the degree of difficulty that lies in acquiring the phonemes of a foreign language may be due to the difference in visual cues had already been published more than 40 years ago [14]. In this study, the difficulty that Japanese learners of English face in making the distinction between /r/ and /l/ is taken up. While this difficulty is usually attributed to the fact that these two phonemes do not exist in the language, Goto claims that the difficulty is due to the fact that there is the disadvantage of not being able to "read the lips of the speaker".

Using computer based methods for visual speech in language learning is still a fairly new enterprise; however, several studies have attempted to test its effectiveness. The talking head Baldi, for example, was used to teach non-native phoneme contrasts. The improvement of both speech identification and production had been observed, but the results indicated that the viewing of the internal articulators was not an additional benefit [15]. In another study, native English speakers were tested on their pronunciation accuracy of non-native segments using Bao (the Mandarin speaking version of Baldi) and Badr (the Arabic version). The analysis showed support for the value of employing visual speech in learning a non-native segment, but here again, the outside of the face was more beneficial compared to the sagittal view illustrating the tongue, palate and velum [16].

Some studies, however, have shown that the information of vocal tract articulator movements can assist in pronunciation instruction. Furthermore, the output of visual articulatory 2D or 3D models are claimed to be correctly interpreted even by young children, thus implying the usefulness of these types of models [17].

4. Segmental vs. Suprasegmental

In addition to the type of information to include in the system, another aspect that needs to be taken into consideration is the phonological unit to be focused upon. The general trend nowadays is to lay emphasis on the communicative factor, and there are several studies suggesting that focusing on the suprasegmental element in pronunciation teaching has an impact on the comprehensibility of learners' output [e.g. 18, 19].

There is agreement among some researchers that the suprasegmental errors observed in L2 speakers have more serious effect on intelligibility than segmental ones. In one study, two groups of L2 students of English received instruction in segmental and suprasegmental features respectively. The result of this study indicated that in terms of narrative reading, only pronunciation teaching based on suprasegmental features had any effect on the comprehensibility of the learners' production [20]. But here again, we find inconsistent results in the literature. For example, Jenkins, based on data collected from six learners of English, maintained that instruction in segments should be prioritized over suprasegmentals. The learners, two Japanese, three Swiss-German, and one Swiss-French, were instructed to engage in various pair work that included social conversation, information exchange and problem solving tasks. When analyzing the interaction that took place between the receiver and the interlocutor, Jenkins found that out of the 40 cases where the receiver could not understand the intended meaning of his/her interlocutor, 27 were designated as cases of difficulty in producing segments. Based on this finding, she concluded that instruction in segments should be prioritized.

Although there is no denying the fact that both segmental and suprasegmental aspects play significant roles in pronunciation, instruction in segments may be a good starting point, especially for beginners of L2. Since pronunciation involves the physical aspect of language, it would be beneficial for learners of L2 to locate the muscles that they need to make the individual sounds of the foreign language, especially those that do not exist in their native tongue before focusing on the higher prosodic levels. In this sense, we decided that our priority in developing the pronunciation system would be on the segments

5. The system

Based on the findings of previous studies, we decided that our pronunciation system will be characterized by the following two features: 1) making good use of visual data; 2) focusing on the segments rather than the suprasegmentals.

As we have already mentioned above, the contribution of visual cues to the understanding of individual speech sounds goes back several decades and there is nothing new about the concept itself. The explanation taken from a traditional pronunciation drill book states that "you should feel your lips and tongue move and your jaw drop lower, then rise again as

you go from one sound to the other. Use a mirror to watch your mouth produce the sounds." [21]. However, the major problem to this "mirror" approach as well as the other "traditional" approaches based on visual cues is that the movement of the tongue cannot be made clear since it hides inside the mouth.

The advent of computers has allowed learners to display visual cues of sound in a more "sophisticated" way. Technology now allows learners to easily convert sound data into sophisticated digital representations. However, this type of representation can be quite useless because in most cases the users would not have the slightest idea of how to interpret them. Even if one gets formal training in how to interpret these representations, it is still hard to link them into actual physical action.

Some systems take a more user-friendly approach to pronunciation by presenting a colorful and easy-to-understand illustration of the speech organ together with detailed explanation of the phonemes used in different languages. The website of the University of Iowa is one such example where an animated articulatory diagram and video clips of the sound spoken in context are presented in a user-friendly way [22]

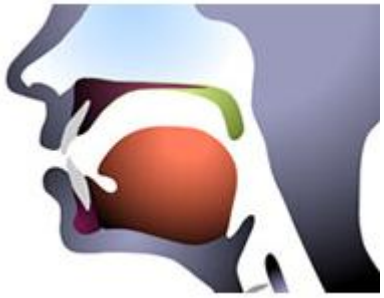


Figure 2: *Example of an interactive articulatory diagram [22]*

While systems such as the one shown in Figure 2 give excellent detailed account of how the speech organs move and the state in which they are situated in pronouncing each individual sound, one of the drawbacks is that although the learners can see how the ideal movements should be, there are no functions that allow learners to receive visual feedback on what they are pronouncing. In other words, these systems, although are very useful in presenting the ideal state, they do not make it possible for learners to actually understand what their own problems are because there is no way the learners can visualize what takes place inside their mouths. Another problem with these articulatory diagrams is that they can only be viewed from one side, i.e., the side view. This makes it extremely difficult for learners to link the visual information into actual physical action. Even using mirrors does no good here, because humans cannot observe what takes place inside the mouth from this angle.

With the pronunciation system that we have developed, the users have access to the speech organs from two different angles: i.e., the side view (cf. Figure 3) and the front view (cf. Figure 4)

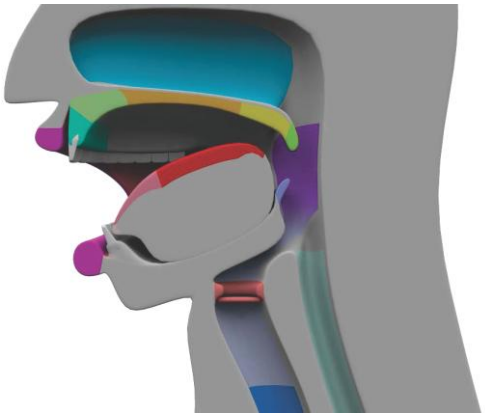


Figure 3: *The side view*

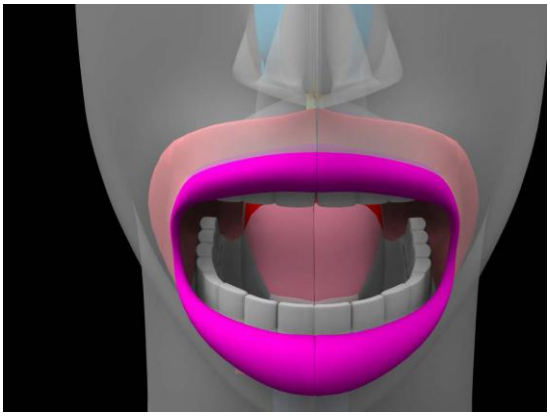


Figure 4: *The front view*

In addition to these two angles, there is the transparent mode as indicated in Figure 5:

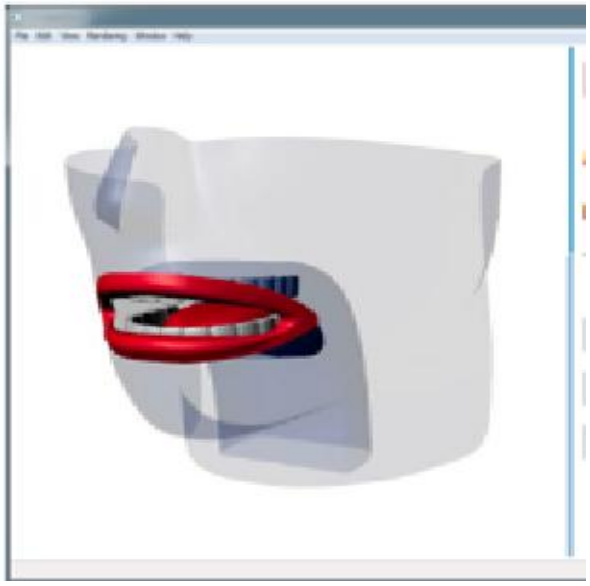


Figure 5: *The transparent mode*

The transparent mode allows users to take a look into the internal organs in actual movement. The user is able to rotate the angle at any degree that he/she likes while slowing the

speed of animation, or taking pauses or freezing the frame in cases where they would like to replicate the movements that take place. At present, we are compiling a list of English words which are claimed to be commonly difficult for Japanese to learn, but instead of just presenting the model movements for these words, we have added on controllers that allow instructors to freely move and adjust the lips, tongue, jaws so that they can point out the deviant movements made by the learners themselves.

At present, the lip, tongue and jaw movements have to be adjusted manually by the instructors, but we are planning to link the visual representations to a speech recognition system for use in autonomous learning.

In addition to learners of foreign languages, this system is also designed to contribute in training the speakers of pathological speech. The speech organs in transparent mode is expected to be especially useful for the hearing impaired population as well as others with severe aural and oral communication difficulties.

6. Conclusion

In this paper, we reported on the pronunciation system that we are developing. The two main features of the system are: making good use of visual data and focusing on the segments rather than the suprasegmentals. This is a practical and productive pronunciation software that should be enjoyable to use for foreign language learners of all ages.

Although it is not easy to prove the effectiveness of employing the internal articulatory movements for training, it may be possible to compare and monitor the articulatory changes made before and after the training using ultrasound images [23]. This we leave for future research.

7. Acknowledgements

The author would like to thank Fumiki Teratani, Tomohide Kano, Masahiro Tachibana and Akira Ishii for their contributions to this study. This study is funded by Grants in Aid for Scientific Research (C) (No. 23520594).

8. References

[1] Gilakjani, A. P. and Ahmadi, M.R., "Why is pronunciation so difficult to learn?", English Language Teaching, 4(3) 74-83, 2011.

[2] Celce-Murcia, M. Brinton, D.M. and Goodwin, J.M., Teaching pronunciation: a reference for teachers of English to speakers of other languages. Cambridge University Press, Cambridge, 1996.

[3] Pica, T. "Questions from the language classroom: research perspectives"., TESOL Quarterly 28(1), 49-79, 1994.

[4] Scovel, T. A time to speak. A psycholinguistic inquiry into the critical period for human speech. Rowley, MA, Newbury House, 1988.

[5] Bamkin S. "How not to fix a problem: misapplications of pronunciation theory", Gateway Papers. A Journal for Pedagogic Research in Higher Education 1, 163-174. 2010.

[6] British Council Singapore "Focus on Pronunciation", Online: http://www.britishcouncil.org.sg/en/course/business-english-part-time/focus-pronunciation, accessed on 20 Mar 2013.

[7] Al-Seyghayar, K. "The effect of multimedia nnotation modes on L2 vocabulary acquisition: a comparative study", Language Learning and Technology, 202-232. 2001.

[8] Yeh, Y. and Wang, C. "Effects of multimedia vocabulary annotations and learning styles on vocabulary learning", CALICO Journal, 21 (1), 131-144. 2003.

[9] Hasegawa, K., Amemiya, S., Kaneko, K., Miyakoda, H., and Tsukahara,W. Promoting autonomous learning: a multilinguistic word learning system based on iPod", Proceedings of the 2007 International Conference on ESL/ EFL, 70-83. 2007.

[10] Miyakoda, H., Hasegawa, K., Ishikawa, M., Shinagawa, N., Kaneko, K., and Fukaya, K. "Designing an autonomous learning environment for vocabulary acquisition", Computer and Education 24, 90-95. 2008.

[11] McGurk, H. and MacDonald, J. "Hearing lips and seeing voices", Nature 264, 746-748. 1976.

[12] Hardison, D.M. "The visual element in phonological perception and learning", in M. C. Pennington [Ed], Phonology in Context, 135-158, Palgrave , 2007.

[13] Walden, B. E., Prosek, R.A., Montgomery, A.A., Scherr, C.K., and Jones,C.J. "Effects of training on the visual recognition of consonants", Journal of Speech and Hearing Research 20, 130-145. 1977.

[14] Goto, H. Auditory perception by normal Japanese adults of the sounds "l" and "r". Neuropsychologia 9, 317-323. 1971.

[15] Massaro, D. W. and Light, J. Read my tongue movements: bimodal learning to perceive and produce non-native speech /r/ and /l/. In Proceedings of the 8th European Conference on Speech Communication and Technology (CD-ROM). Geneve, Switzerland, 2003.

[16] Massaro, D.W., Bigler, S. Chen, T. Perlman, M and S. Ouni Pronunciation training: the role of the eye and ear. In Proceedings of Interspeech 2008, 2623-2626. 2008.

[17] Kroeger, B.J., Graf-Borttscheller, V. and Lowit, A. two-and three-dimensional visual articulatory models for pronunciation training and for treatment of speech disorders. In Proceedings of Interspeech 2008, 2639-2642. 2008.

[18] Nakashima, T. "Intelligibility, suprasegmentals, and L2 pronunciation instruction for EFL Japanese learners" Fukuoka Kyouikudaigaku Kiyou 55 (1), 27-42, 2006.

[19] Elliot, A.R. "Foreign language phonology: field independence, attitude, and the success of formal instruction in Spanish pronunciation", The Modern Language Journal 79 (iv), 530-542

[20] Derwing, T.M., Munro, M.J., and Wiebe G. "Pronunciation instruction "fossilized" learners: Can it help?" Applied Language Learning 8, 185-203, 1998.

[21] Orion, G.F. Pronouncing American English. Heinle and Heinle, 1987.

[22] The university of Iowa, "The sounds of spoken language", Online: http://www.uiowa.edu/~acadtech/phonetics/index.html #pluginCheck, accessed on 2 Apr 2013.

[23] Engwall, O. "Can audio-visual instructions help learners improve their articulation?—an ultrasound study of short term changes", In Proceedings of Interspeech 2008, 2631-2634. 2008.

OJAD: a Free Online Accent and Intonation Dictionary for Teachers and Learners of Japanese.

Hiroko Hirano‡, Ibuki Nakamura†, Nobuaki Minematsu†, Masayuki Suzuki†, Chieko Nakagawa∗
Noriko Nakamura⋆, Yukinori Tagawa∗, Keikichi Hirose†, Hiroya Hashimoto†

‡ Northeast Normal University, Jilin, China † The University of Tokyo, Tokyo, Japan
∗ Waseda University, Tokyo, Japan ⋆ Tokyo University of Foreign Studies, Tokyo, Japan

Abstract

We developed the very first online and free framework for teaching and learning Japanese prosody including word accent and phrase intonation. This framework is called OJAD (Online Japanese Accent Dictionary) [1], which provides three functions. Subjective assessment by teachers shows very high pedagogical effectiveness of the framework.

Index Terms: language education, Japanese prosody, accent sandhi, OJAD, TTS synthesizer, assessment experiment

1. The three functions of OJAD

1.1. Comprehensive illustration of accent changes

Japanese is a pitch accent language and along with the conjugation of verbs and adjectives, their accent patterns also change regularly and systematically. If a learner desires to speak sounding not foreign accented, he or she will need to follow the accent rules. However, as existing word dictionaries merely list Dictionary Form before conjugation of a verb/adjective, and even accent dictionaries just describe the accet rules of conjugation with a few samples at the end, learners don't have accesible resourse at present. Therefore, we realized a system that can show the accent changes due to conjugation of these words. Users type verbs and/or adjectives of interest to know their accent changes. Here, twelve kinds of fundamental conjugation were adopted and their accents are displayed in a table. Fig. 1 shows an example. Seven widely-used textbooks were selected and all the verbs and adjectives found in them were manually extracted.

Figure 1: Illustration of the accent patterns of conjugated forms

1.2. Illustration of the accent of long verbal expressions

The first function only shows the accent patterns of the twelve fundamental conjugated forms of verbs and adjectives. Since Japanese is an agglutinative language, a verb can be combined with multiple postpositional and auxiliary words.

So, we developed another system as a second function to show the accent pattern of a given long verbal expression. Fig. 2 shows examples of two accent groups found in Japanese verbs and the accent of the user's input in the right red rectangle.

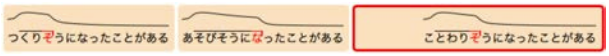


Figure 2: Illustration of the accent patterns of long expressions

1.3. Illustration of the pitch pattern of any input sentence

The first and second functions only focus upon verbs and adjectives. Word accent changes are not only found in these words but also in other words such as nouns. So, as a third function, we developed a prosodic reading tutor to support learners by presenting the pitch pattern of an any given sentence.

This function is realized easily by using several internal modules developed for TTS synthesizers. Three analyses of morphological analysis, accent phrase boundary detection, and accent nucleus location are run for each phrase. An example is shown in Fig. 3.

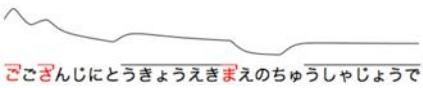


Figure 3: Illustration of the pitch pattern of the part of a sentence

2. Subjective assessment

We asked teachers of Japanese to join a subjective assessment test after learning how to use OJAD. Eighty teachers joined the test, two thirds of whom were teaching Japanese outside Japan. Although the subjective assessment was composed of a series of questionnaire items, we show in Tab. 1 the results of only two fundamental questions: a) How useful do you think the system is for learners? and b) Do you want to use the system in your class? Considering that teaching Japanese prosody is just only one aspect of Japanese language education, we consider that the eighty teachers of Japanese recognize very high pedagogical effectiveness of the proposed framework.

Table 1: Assessment of the three proposed systems (%)

a) How useful do you think the system is for learners?

	1st system	2nd system	3rd system
Very useful	71.0	54.8	62.7
Rather useful	29.0	45.2	28.8
Not so useful	0.0	0.0	8.5
Not useful at all	0.0	0.0	0.0

b) Do you want to use the system in your class?

	1st system	2nd system	3rd system
Yes, definitely	38.7	29.0	42.6
Yes, if needed	59.7	64.5	50.0
No	1.6	6.5	7.4

3. References

[1] OJAD: http://www.gavo.t.u-tokyo.ac.jp/ojad/

SPARSAR: a System for Poetry Automatic Rhythm and Style AnalyzeR

Rodolfo Delmonte, Ciprian Bacalu

Department of Language Studies and Comparative Cultures
Ca' Foscari University of Venice – DD. 1075 30123 Venice - (Italy)
delmont@unive.it

Abstract

Any poem can be characterized by its rhythm which is also revealing of the poet's peculiar style. In turn, the poem's rhythm is based mainly on two elements: meter, that is distribution of stressed and unstressed syllables in the verse, presence of rhyming and other poetic devices like alliteration, assonance, consonance, enjambements, etc. which contribute to poetic form at stanza level.

Traditionally, poetic meter is visualized by a sequence of signs, typically a straight line is used to indicate vowels of stressed syllabes and a half circle is positioned on vowels of unstressed ones. The sequence of these sings makes up the foot and depending on number of feet one can speak of iambic, trochaic, anapestic, dactylic, etc. poetic style.

English poetry has been for centuries characterized by iambic pentameter, that is a sequence of five feet made of a couple of unstressed + stressed syllables. Modern English poetry on the contrary – after G.M.Hopkins – has adopted a variety of stanza schemes.

A poetic foot can be marked by a numerical sequence as for instance in [4] [5] who uses "0" for unstressed and "1" for stressed syllables to feed a connectionist model of poetic meter from a manually transcribed corpus. There he also tries to state the view that poets are characterized by their typical meter and rhythm, which work as their fingerprint.

We also agree with this view, however, we would like to be more specific on the notion of rhythm that we intend to purport. We do that in two ways: by considering stanzas as structural units in which rhyming – if exixtent - plays an essential role. Secondly and foremost, in our view, a prosodic acoustic view needs to be implemented as well, if any precise definition of rhythm and style is the goal. Syllables are not just any combination of sounds, and their internal structure is fundamental to the nature of the poetic rhythm that will ensue. This is partly amenable to the use and exploitation of poetic devices, which we also intend to highlight in our system. But what is paramount in our description of rhythm, is the use of the acoustic parameter of duration. The use of duration will allow our system to produce a model of a poetry reader that we intend to implement in the future by speech synthesis. In our demo we will show how poems can be characterized by the use of rhythmic and stylistic features in a highly revelatory manner, by comparing metrically similar poems of the same poet and of different poets.

To this aim we assume that syllable acoustic identity changes as a function of three parameters:

- internal structure in terms of onset and rhyme which is characterized by number consonants, consonant clusters, vowel or diphthong
- position in the word, whether beginning, end or middle
- primary stress, secondary stress or unstressed

These data have been collected in a database called VESD (Venice English Syllable Database) to be used in the Prosodic Module of SLIM, a system for prosodic self-learning activities. Syllables have been collected from WSJCAM, the Cambridge version of the continuous speech recognition corpus produced from the Wall Street Journal, distributed by the Linguistic Data Consortium (LDC). We worked on a subset of 4165 sentences, with 70,694 words which constitute half of the total number of words in the corpus amounting to 133,080. We ended up with 113,282 syllables and 287,734 phones. The final typology is made up of 44 phones, 4393 syllable types and 11,712 word types. From word-level and phoneme-level transcriptions we produced syllables automatically by means of a syllable parser. The result was then checked manually. This work has been presented elsewhere [1][2].

The analysis in SPARSAR starts by translating every poem into its phonetic form: we used the CMU Pronouncing Dictionary for North American English to translate words into phoneme sequences, augmented with words derived from work above – see also [6]. In a second pass we try to build syllables starting from longest possible phone sequences to shortest one. This is done heuristically trying to match pseudo syllables with our syllable list. Matching may fail and will then result in a new syllable which has not been previously met. We assume that any syllable inventory will be deficient, and will never be sufficient to cover the whole spectrum of syllables available in the English language.

For this reason, we introduced a number of phonological rules to account for any new syllable that may appear. Duration values are derived by comparison with phonologically closest ones – for this we use place, manner of articulation as parameters. We assign mean duration values in msecs to all syllables considering position and stress. We also take advantage of syntactic information computed separately to highlight chunks' heads as produced by our bottomup parser. In that case, stressed syllables takes maximum duration value.

In our demo we will show how poems can be chacterized by the use of rhythmic and stylistic features in a very revelatory manner, by comparing similar poems of the same poet and of different poets. The importance of metrical structure and of poetic rhyming devices is evaluated and also compared.

References

[1] Bacalu C., Delmonte R. (1999a), Prosodic Modeling for Syllable Structures from the VESD - Venice English Syllable Database, in Atti 9° Convegno GFS-AIA, Venezia.

[2] Bacalu C., Delmonte R. (1999b), Prosodic Modeling for Speech Recognition, in Atti del Workshop AI*IA - "Elaborazione del Linguaggio e Riconoscimento del Parlato", IRST Trento, pp.45-55.

[3] Delmonte R. (1999), A Prosodic Module for Self-Learning Activities, Proc.MATISSE, London, 129-132.

[4] Hayward, M. (1991). A connectionist model of poetic meter. Poetics, 20, 303-317.

[5] Hayward, M. (1996). Application of a connectionist model of poetic meter to problems in generative metrics. Research in Humanities Computing 4. (pp. 185-192). Oxford: Clarendon P.

[6] Kaplan, D., & Blei, D. (2007). A computational approach to style in american poetry. In *IEEE Conference on Data Mining.*

The digital instructor for literacy learning

Catia Cucchiarini[1], *Ineke van de Craats*[2] , *Jan Deutekom*[3], *Helmer Strik*[1,2]

[1]Centre for Language and Speech Technology, Radboud University, Nijmegen, the Netherlands
[2]Department of Linguistics, Radboud University, Nijmegen, the Netherlands
[3]Department of Innovation, Friesland College, Leeuwarden, the Netherlands

c.cucchiarini@let.ru.nl, i.v.d.craats@let.ru.nl, J.Deutekom@fcroc.nl, w.strik@let.ru.nl

Abstract

The DigLIn project aims at providing concrete solutions for adult literacy students by developing and testing L2 literacy acquisition material in four different languages and by employing Automatic Speech Recognition (ASR) to analyse the learner's read speech output and provide feedback. We develop the technology, design sample exercises for different languages (Dutch, English, German and Finnish) and test them in literacy classes in adult education centres with adult L2 learners.
Existing language learning material for low-educated second language learners is augmented with an ASR module capable of recognizing what the learners say, of diagnosing possible errors in reading aloud or pronunciation and of providing practice and feedback in learning to read aloud in the L2.

Index Terms*: adult literacy learning, language and speech technology, second language acquisition*

1. Introduction

Europe has many immigrant and refugee adults with a low level of education, who lack basic skills such as reading and writing in both their native language and second language. However, the Common European Framework of Reference (CEFR) for Languages [5] departs from the basic level of primary school and implicitly assumes that adults are readers and writers. This does not correspond to reality. In addition to a high number of low-literate native-born adult residents, Europe counts many non-literate adults who need to learn to read and write for the first time, in a language other than their mother tongue. The numbers of non-literates and low-literates (unable to read and write well enough to use these skills in their daily lives) differ from country to country, but are between 10–15% of the population. Part of them are nonnatives who participate in (integration) courses with a literacy component. For instance, in Germany there were 65.000 of these immigrants between 2005 and 2012. 37,2 % of them are primarily illiterates. The DigLIn project aims to support this group of immigrant learners.

Being able to read and write is a prerequisite for active participation in society and employability. Poor oral and written proficiency in the second language (L2) leads to social exclusion [3], and prohibits social and economic integration [8]. Literacy – the ability to use reading and writing– is clearly a key factor in the integration and participation of immigrants in the society in which they live. Helping people acquire basic skills such as reading and writing is a crucial step in supporting social inclusion and citizen participation. Many European countries have programmes and initiatives aimed at promoting literacy acquisition and look for innovative methods that can boost the efficiency of literacy teaching by making it more flexible and more individualized.

In spite of these considerations, all kinds of financial cuts threaten adult education across the EU; in the UK for example, immigrant adults wishing to take ESOL classes exceed availability [4], and government funding continues to be reduced across the ESOL sector. This means that low-literate and non-literate adults are increasingly expected to be responsible for the costs of their individual learning. This implies that, more than ever before, speed of learning has become an essential factor for the least skilled L2 learners.

A Dutch study [14] aimed at assessing the learning load (in hours of instruction) for learning to read and write revealed two important findings. The time allotted to computer work – in which individuals had to work actively – correlated positively with reading scores, but the time allotted to whole group work correlated negatively. Moreover, there were substantial individual differences in pace across learners. The fast learners had additional opportunities to progress faster when there were facilities and materials to serve them. Facilities in this case means computer facilities and materials means ICT, CALL (Computer Assisted Language Learning) and multi-media.

Unfortunately, little attention is devoted to this group of learners. Publishers of course materials have little or no interest in this group because it is too small for gaining a reasonable profit and the development of computerized materials and multimedia adapted to this target group is expensive

2. Research background

Language and literacy development by first time (adult) readers in a second language is a new and underdeveloped domain of research [16] [18] [20]. Most research on reading concerns children who learn to read in their mother tongue. Adults who learn to read in an L2 with quite a different phonological system than that of their mother tongue face additional problems in mastering the phoneme-grapheme correspondences of the L2. These problems tend to mean that, in the Netherlands for example, many immigrants and refugee adults do not attain the level for reading and writing now required for integration and naturalisation [12] [13].

In basic reading instruction, two main approaches can be distinguished: the sight-word approach and the phonics approach. Finnish with its excellent correspondence between graphemes and phonemes lends itself to phonics instruction, and

due to its regular syllable structure, the syllable is most often focused on as a unit. For Dutch and German with a relatively transparent orthography, phonics instruction is preferred in both L1 and L2 literacy programs. For English with its deep orthography, a sight-word approach is possible, yet the alphabetic code still has to be cracked for decoding the many regularly spelt words. Models of beginning reading development agree on a first stage of direct-word recognition using basic visual cues, a second stage of indirectly mediated word recognition through graphic cues (grapheme-phoneme correspondence) and a third stage of direct word recognition based on automatization [11].

The question arises as to why immigrant adults are ultimately less successful than children. Important reasons are that L2 adults receive fewer hours of reading instruction, the course material is of a lower quality and the ICT applications now readily available to children are not appropriate for adults. In a class, children often decode graphemes aloud and synthesize them into a word. In adult classes, this is often seen as childish and is therefore restricted to an absolute minimum. Moreover, in a class of 10 or 15 adults, individual differences are often too large to make it a useful activity. The same holds for reading texts aloud: only one student reads, the others succeed in passively following to various degrees.

More active practice in which literacy students can produce the sounds or words while a computer tells them whether they are correct is a much needed improvement. This becomes possible through the application of ASR technology because the computer recognizes the word uttered and can provide feedback to the learner on whether the word was correctly read or not. There is a long history of experimentation on children using ASR [7] [15] [17] [21], but this technique has not yet been applied in adult literacy education. In addition, most of the studies on reading support through ASR concern systems that can follow the learner while reading aloud, but which are not aimed at identifying errors at the phoneme level to diagnose grapheme-phoneme connections. For this latter kind of application, more advanced technology is required.

.

3. CALL for literacy development

Computer Assisted Language Learning (CALL) applications offer enormous advantages compared to teacher-fronted classes: learners can practice as much as they want at their own pace in a stress-free environment and can receive individualized, adaptive feedback from the computer. This is particularly important for adult language learners who lack basic skills such as literacy and who can use materials when they are able to take time off from family and other responsibilities. Becoming literate in a second language can be particularly challenging and requires much practice and patience. A reader not only identifies letters (graphemes) and words (the analyzing part of the reading process), s/he also makes a correspondence with the sound (phoneme) represented by the grapheme and the sounds that together form a word (synthesis). Perception is only one side of the reading process; the learning reader also has to translate graphemes into sounds, combine them into words and produce them. Feedback is traditionally given by the teacher or another proficient reader, but could be provided individually for a large group of learners, by the computer as proposed here.

4. Digital Literacy Instructor (DigLIn)

The Lifelong Learning Program (LLP) project 'Digital Literacy Instructor' aims at providing concrete solutions for adult literacy students by developing L2 literacy acquisition material in four different languages and by employing Automatic Speech Recognition (ASR) to analyse the learner's read speech output and provide feedback. We develop the technology and design sample exercises for different languages (Dutch, English, German and Finnish) and test them in literacy classes in adult education centres with adult L2 learners.

Existing language learning material for non-literate and low-literate L2 learners developed at Friesland College (the digital sources of the FC Sprint^2 [6]) is augmented with an ASR module capable of recognizing what the learners say, of diagnosing possible errors in reading aloud or pronunciation and of providing practice and feedback in learning to read aloud in the L2.

This is a considerable improvement in comparison to existing systems in which learners can listen to audio recordings and carry out receptive exercises of the sound-to-grapheme type. In our system learners have the possibility of engaging in production exercises to learn and practice grapheme-to-sound or graphemes- to-word correspondences in the L2, reading a sound, a word, or a sentence out loud and receiving corrective feedback from the computer.

In addition to the already mentioned advantages of ASR-based CALL, it is important to underline the importance of a private, stress-free environment in L2 beginning reading and speech production, because low-literate language learners often feel ashamed of their weak skills and then refrain from practicing in the presence of teachers and other students. After a short introductory period, the system we develop can be used at home so that learners can feel comfortable and can practice anytime for as long as they want.

5. The pedagogical approach in DigLIn

The pedagogical approach in FC-Sprint^2

As explained above, in this project we depart from a common framework (digital sources of FC-Sprint²), and develop content and exercises in keeping with the specific features and requirements of the language and the teachers in question.
The concept of FC-Sprint² [6] is based on two pillars:

a. A different approach to students by teachers: from control by the teacher to autonomy for the students.

Students have to work with their resources, the teacher is the last resort.

b. Providing students with resources so that they can become more autonomous learners.

We try to build small programs so that student can find out themselves instead of being told by a teacher how it all works.

The principles underlying FC-Sprint[2] [6] can be summarized as:

- Start with high expectations as teachers who expect more get students who perform better.
- Students should carry responsibility (prevent passive behaviour).
- Learning efficiency grows if the student carries responsibility.
- Learning is doing what you cannot do yet.
- Students need to make mistakes in order to learn.
- Learning is more effective when students feel the need to learn.
- Students should first employ their own resources and ask for help when they need it.

When a student has been struggling with a certain subject the effect of instruction likely is much stronger than when a topic is completely new. So first we try to let students work with their own resources before a teacher explains.

- Talent is always an observation afterwards.

There actually is quite a lot of evidence that "talent" is at least a highly overrated concept and that achievement takes a lot of time and effort. Relying on talent can slow down development. (see [2] [9] [10])

- A student can learn everything until (s)he proves otherwise.
- The student is addicted to learning efficiency.
- Motivation is the result of a process.

Teachers can have a lot of influence on a student's motivation (negative and positive).

The DigLIn approach to literacy instruction

The pedagogical approach of FC-Sprint2 is translated in DigLIn by giving non-literate learners the materials for cracking the alphabetical code and providing them all necessary feedback. The teacher makes clear that (s)he is confident that learners will manage to read these words in a few days and will be able to show that to the whole class. As soon as learners have found out what the system can do for them, they will have the feeling of success and will be more and more motivated to continue.

The underlying method for a system like FC-Sprint[2] [6] and the one to be used in DigLIn is in fact a phonics-based method: the structure method. The primary aim of the structure method is grasping the structure of the spelling system or associating specific sounds (phonemes) with specific letters (graphemes). This is done on the basis of a whole word which is visually and auditorily structured in smaller units (analysis). In this way the student learns to consider a written word as a composite unit of separate elements and to make use of the systematic nature of letter-sound associations for autonomously decoding new words.

The basis of this method is a restricted number of concrete basic words the meaning of which is clear. In classes of 6- and 7-year-old children, those words are presented in a context of a story or a picture story and learnt by heart. In DigLIn those words can be made clear by pressing a button. Basic words should have a 'one-on-one grapheme-phoneme correspondence', that is to say that the sounds are not influenced in their pronunciation by preceding or following sounds or by the fact that they are in word-final or syllable-final position, as is the case in Dutch. We use the label "pure sound".

Examples for:

- English: dad, map, mop, jump, bin, big, yes
- Dutch: mat, kap, kip, boom
- German: Rat, Hut, Oma
- Finnish: eno, iso, akka

Ideally, there is a one-to-one relationship between phoneme- and grapheme Many languages have too few graphemes for the repertoire of phonemes, which is the case for Dutch, but more particularly for English with one and the same grapheme representing different phonemes.

As soon as a couple of basic words are recognized, the analysis and synthesis exercises can start. The spoken word is analyzed in sounds, the written word in letters. Next, the sounds are blended to a spoken word. Many analysis and blending exercises are needed for establishing a tight association between sound and letter. Software can help to automatize this phase of the reading process. For this stage, FC-Sprint[2] has found many challenging exercises with feedback (e.g., a letter dragged to an incorrect position, does not stay, but jumps away, back to its original position). An example of such a drag-and-drop task for a Danish version of the system is given in Figure 1:

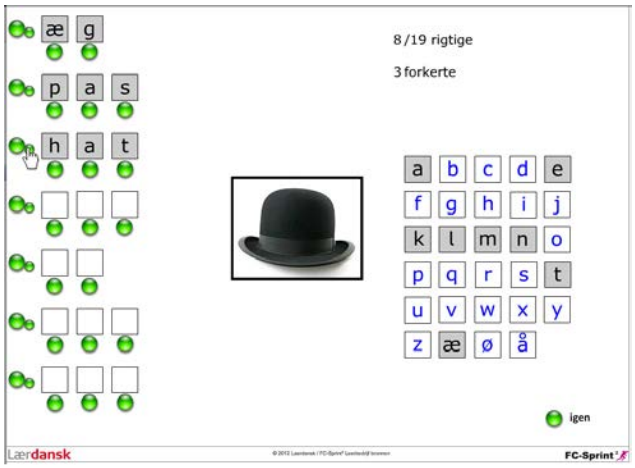


Figure 1. From letters to words, a drag-and-drop exercise for Danish

In this example a student can drag letters to the right square. On the left hand side students can hear the complete word by clicking on the big green button. By hovering over the little green button a student can see what the word means. By clicking on the button below the square he/she can hear the individual sounds of the letters. If a student drags a wrong letter to the square the letter jumps back to its original position and a "mistake" sound is heard.

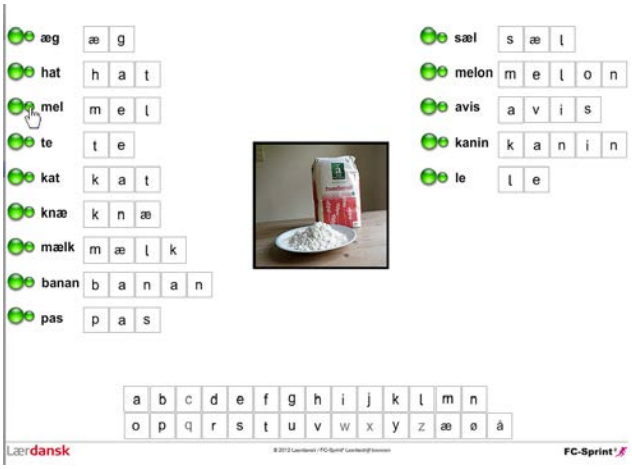


Figure 2. Presentation of the words for Danish with sound bar

In Figure 2 students have a "sound bar" with all phonemes at their disposal at the bottom of the screen. By clicking on a letter in this sound bar students can hear the individual sounds of the complete alphabet. By clicking on the big green buttons students can hear the sound of the word. Hovering over the little green buttons gives students a photo displaying the word. By clicking on the letter squares behind the words students can hear the individual sounds corresponding to the letters. (See for the working examples: http://diglin.eu/?page_id=222.)

In DigLIn we provide feedback on reading aloud the blended words. The step to reading new words and the transition from spelling words to a more automatized stage is supported first by exercises in which either the onset or the rime is kept constant, as illustrated in Table 1, and later, by a mix of these words.

Table 1. Example of building up complexity of the word structure for Dutch

CVC with same rime	CCVC with same onset	CCVC with same onset
p-ak	st-ok	sch-ool
z-ak	st-ak	sch-aap
b-ak	st-op	sch-ep
t-ak	st-ip	sch-ip
l-ak	st-ik	sch-uur
v-ak	st-ap	sch-oen

FC-Sprint[2] materials offer possibilities for analyzing and blending and for training automatization. Feedback on reading aloud the words has to be built in by applying ASR.
In addition to the autonomy stimulating approach adopted in FC-Sprint[2], we provide a limited set of video-taped instructions, so as to allow students to work quite alone with the program, if they wish to do so, without the support of classmates and/or teachers.

5.1.1. Criteria for selecting words

Words are selected taking into account an order of increasing complexity. Words also target adult LESLLA (Low-Educated adult Second Language and Literacy) readers who are still at an early stage of reading, e.g., the glance and guess stage. Because preferably photographs, but also pictures are used for explaining the meaning, words should be concrete content words so that pictures can be attached. Frequency is also a selection criterion, but is applied with caution since many frequent words are function words that cannot be represented by images (and are more difficult to understand). Frequency lists are also of less importance because they are not based on what low-literate immigrants are likely to encounter. Systematic stress variation in polysyllabic words is also taken into account. Together with the criteria for usefulness for literacy instruction we come to the following criteria for building up complexity.

5.1.2. Criteria for building up complexity

Start with:
1. CV(C) words
2. "Pure sound'' words
3. Maximal difference: first cardinal vowels: /i/, /u/, /a/ occurring in most languages of the target group of learners (so, not /y/). Followed by consonants that are maximally different on the basis of other features.
4. No minimal consonant pairs in one word or series of words for reasons of auditory similarity (not: **p**ak and **b**ak) or visual similarity (not: **d**ak and **b**ak).

And proceed with:
1. Vowels and consonants from maximally different (/a/-/u/-/i/) to minimally different (/i/-/I/ or /u/-/Y/) and from very common in other languages to language-specific sounds (e.g., for Dutch ui in huis ('house').
2. From CVC to CCVC or CVCC and more extensive consonant clusters
3. From monosyllabic to disyllabic words then polysyllabic words
4. From concrete to abstract words
5. From noun to adjective and verb
6. From pure sound to spelling conventions (e.g. in Dutch for open and closed syllables: raam-ramen)
7. From word to sentence

6. Language and speech technology

Innovative in the DigLIn project is that within the CALL system for literacy training use is made of language and speech technology' (LST), and especially 'Automatic Speech Recognition' (ASR). As is well known, developing ASR-based applications for L2 learners implies having to deal with non-native speech which, for many reasons, is more challenging than native speech [1] [19].

Therefore, exercises are developed such that the possible answers by the users are restricted (see e.g. the screen shot of FC-Sprint[2]). For every item, a list of correct and incorrect responses is used to limit the recognition task. This can be achieved in different ways: by using confidence measures to identify an utterance in the list of possible responses, or by using the list of responses to train constrained language models. In

doing so, care is taken to also include a number of possible meta-responses, such as the equivalents of "I don't understand" or "what?".

The DigLIn system is intended to be web-based, and should run in different browsers. Since practical, technical details can be important for a good performance, we carefully look at issues such as head-sets, audio recording settings (for different browsers), audio file formats, signal-to-noise ratio (SNR), and noise cancelling (techniques).

When an error has been identified in the learner's response, feedback is provided to signal this to the learner. With respect to the spoken responses, feedback is provided on two levels: (1) on the utterance level, and (2) on the error level. Regarding the former, the speech recognition module determines which utterance was spoken, and before proceeding to error identification the learner is given feedback on the recognized utterance. After all, it would be highly confusing if the learner gets feedback on (parts of) an utterance that was not spoken at all by the learner. It is also more confusing if the system signals an error while the response was correct (false alarm), than v.v. (false accept). Therefore, in tuning the system we try to keep the number of false alarms smaller than the number of false accepts.

Feedback is gradual in the sense that it indicates the degree of correctness. A student can repeat again and again and a slider indicates in real time whether there is any improvement so that the student can try again immediately and see whether the new attempt is better or worse. The feedback should be simple, intuitive, and easy to interpret, such as a score presented visually (e.g. a bar, possibly with colors).

While for many languages databases of native speech are available, corresponding databases of non-native speech are in general lacking, especially non-native speech for the target groups of the application. This makes it even more challenging to develop ASR technology for this application. In DigLIn, we cope with this issue in the following way. We start with an ASR trained on native material, using native resources (lexica, speech corpora, etc.). Later we study whether using extra information can improve the system's performance. Possibilities are to use non-native resources (lexica, speech corpora, etc.), and to use information on errors made by the target group (annotations of errors). Available non-native audio recordings and error annotations are first used, while interactions of users with (initial versions of) the system, and annotations of (part of) these recordings will be employed at a later stage.

Learners can also listen to correct examples in stored audio recordings. Students can repeat the speech they listen to in the program as often as they want. We carefully considered criteria for these audio recordings, such as normal speed, careful speech (no or limited amount of reduction), sounds natural, limited amount of silence, whether or not carrier sentences should be used, good selection of speakers (male and female, amount of dialect, etc.), recording environment and conditions (studio, 'silent office'), technical specifications (e.g. file format (wav/mp3), signal-to-noise ratio (SNR), etc.). The reason for presenting the speech in the program at normal speed is to provide a "jump" from the slow speech usually spoken by teachers to real world speech.

At SLaTE 2013 we intend to show a preliminary version of the system, illustrating the feasibility of the exercises, the type of practice the learners receive and the corrective feedback provided by the system. Possible additional features and their pedagogical relevance are discussed.

7. Conclusions

ASR seems to constitute a valuable add-on to current computer-based adult literacy programs for various reasons. The nature of the language tasks involved is such that constrained ASR tasks can be designed, which in turn guarantees adequate ASR performance. For the first time, this makes it possible for learners to receive automatic, immediate feedback on their reading performance, without learners having to make comparisons themselves between what they heard and they produced themselves. This is an important improvement for L2 reading instruction, which paves the way to more autonomous learning conditions.

8. Acknowledgements

This project has been funded with support from the European Commission under project number 527536-LLP-1-2012-NL-GRUNDTVIG-GMP. This publication reflects the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

We are indebted to the other members of the DigLIn team for their contributions, in alphabetical order: Marta Dawidowicz, Vanja de Lint, Maisa Martin, Jan-Willem Overal, Karen Schramm, Taina Tammelin-Laine, Joost van Doremalen and Martha Young-Scholten.

9. References

[1] Benzeghiba, M., R. D. Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens, Automatic speech recognition and speech variability: a review, Speech Communication, vol. 49, no. 10-11, pp. 763-786, 2007.

[2] Bloom, B. (1985) Developing talent in young People, Ballantine Books.

[3] Bynner, J. (2001). *Outline of the research, exploratory analysis and summary of the main results*. London: Centre for Longitudinal Studies.

[4] Cooke, M. (2010) ESOL in the United Kingdom. Paper at the inaugural EU-Speak workshop, Newcastle, 6 November.

[5] Council of Europe (2001). *A Common European Framwork of Reference for Languages: Learning, teaching,, assessment.*. Cambridge: Cambridge University Press.

[6] Deutekom, J. FC-Sprint², Grenzeloos Leren, Boom 2008.

[7] Duchateau J. Kong, Y. Cleuren, L. Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquière, P., Verhelst, W., Van hamme, H. (2009) Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules. Speech Communication 51(10): 985-994.

[8] Dustmann, C. and F. Fabbri (2003). Language proficiency and labour market performance of immigrants in the UK. *The Economic Journal*, 113, 695-717.

[9] Dweck, C. (2006) Mindset: The New Psychology of Success, Random House Publishing Group,

[10] Ericsson, A. (2006) The Cambridge Handbook of Expertise and Expert Performance, Cambridge University Press.

[11] Kurvers, J. (2007). Development of word recognition skills of adult L2 beginning readers. In N. Faux (ed.) Low-Educated Second Language and Literacy Acquisition. Richmond, Virginia: The Literacy Institute at Virginia Commonwealth University, 23-43.

[12] Kurvers, J. and I. van de Craats (2009). Het Haalbaarheidsonderzoek van De Voortwijzer [Feasability Assessment]. Tilburg: Universiteit van Tilburg.
[13] Kurvers, J., W.Stockmann, & I. van de Craats (2010). Predictors of success in adult L2 literacy acquisition. In Th. Wall & M. Leong (eds.). Low-educated Adult Second Language and Literacy Acquisition. Banff: Bow Valley College: 64-79.
[14] Kurvers, J. and W. Stockmann (2009). Alfabetisering in beeld. Leerlast en succesfactoren. [Literacy in the picture. Learning load and success factors.] Tilburg: Universiteit van Tilburg.
[15] Mostow, J., Roth, S., Hauptmann, A.G., Kane, M., A Prototype Reading Coach that Listens. Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-1994), pp. 785-792, 1994.
[16] Roll, H., Schramm, K. (Hrsg.), Alphabetisierung in der Zweitsprache Deutsch, Osnabrücker Beiträge zur Sprachtheorie, 77, Duisburg: Gilles & Francke, pp. 5-10, 2010.
[17] Russell, M., D'Arcy, S. (2007) Challenges for computer recognition of children²s speech, In proc. Of SLaTE-2007, pp. 108-111.
[18] Tarone, E., M. Bigelow, & K. Hansen (2009). Literacy and Second Language Oracy. Oxford: Oxford University Press.
[19] van Doremalen, J., Cucchiarini, C. and Strik, H., Optimizing automatic speech recognition for low-proficient non-native speakers. EURASIP Journal on Audio, Speech, and Music Processing, Article ID 973954, 13 pages, 2010.
[20] Wagner, D.A. (2004). Literacy (ies), culture(s), and development(s): The ethnographic challenge. *Reading Research Quarterly,* 39, 234-241.
[21] Li, Y., and Mostow, J. (2012). Evaluating and improving real-time tracking of children's oral reading. In Proceedings of the 25th Florida Artificial Intelligence Research Society Conference (FLAIRS-25), Marco Island, Florida.

Off-line mobile-assisted vocabulary training for the developing world

Nic J. de Vries[1], *Febe de Wet*[1,2]

[1]Human Language Technologies Research Group, Meraka Institute, CSIR, Pretoria, South Africa
[2]Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa
ndevries@csir.co.za, fdwet@csir.co.za

Abstract

Mobile-assisted language learning applications (MALL) has significant potential in the developing world where access to teachers and classrooms are real barriers to learning.

This demonstration of an Android-based mobile language learning application is used to teach vocabulary and employs off-line Automatic Speech Recognition (ASR) and Text-to-speech (TTS) technologies with custom-built language and acoustic models, incorporating the key design criteria outlined in the article.

Index Terms: Computer-assisted language learning (CALL), Mobile-assisted language learning (MALL), Automatic Speech Recognition (ASR), Text-to-speech (TTS) synthesis, Android, on-device ASR and TTS, game-based learning, developing world.

1. Introduction

Mobile-assisted language learning applications (MALL) could potentially cause a shift from teacher-centric learning where student participation is externally encouraged, to student-centric interactive learning with increased participation from each student [1]. This would be a positive shift especially in the process of acquiring a second or further language.

This article focusses on the the conceptual design aspects of a mobile vocabulary training application, and not on the need for such applications.

2. Approach

Various approaches could be used in designing a mobile application focussed on language learning. Some of the design principles that we have used in this application is outlined below.

2.1. Design for primary goals

Keeping the primary learning goals central throughout the design process, is key to the success of a language learning application. The specific learning goals that we aim to achieve in this application is four-fold:

Firstly, we aim to *elicit actual spoken utterances* from the student, as this step is crucial in gaining the necessary confidence in speaking a language as opposed to 'theoretically' speaking a language. Audible practice makes all the difference.

Secondly, allowing the student to audibly hear a *target pronunciation* of a specific word or phrase, encourages the verbalisation of the actual utterance when doubt or ill-confidence exist. Thirdly, displaying the *graphemic representation* of the word to be pronounced, connects the representation with the verbal pronunciation that will be spoken. Lastly, by requiring the association between a picture representing a word to be matched with the specific word, the *semantic meaning* of the word is linked to the graphemic and audible representation [2].

By purposefully aiming at these primary goals throughout the design and construction stages, various other inevitable design decisions are allowed much more flexibility, which aids in meeting all design criteria, and still ensures the overall outcome of the application.

2.2. Design for content independence

Developing effective mobile language learning applications is no trivial matter. With each application comes new challenges and new pitfalls—besides the cost of such development. By designing this application in such a way as to make the actual contents that needs to be mastered independent of the specific learning method, this mobile application could easily be deployed to teach a completely different language or a different curriculum without redesign.

2.3. Design for pedagogical support

The order in which material is to be mastered forms the heart of any pedagogical approach. As part of the design of a vocabulary training curriculum, a specific sequence of words, potentially with targeted phonemic content, will need to be mastered prior to progressing to a new set of words.

In order to accommodate such a specific work flow, the concept of a lesson is enforced within the application, while maintaining a certain degree of freedom within each lesson to stimulate dynamic learning and provide flexibility for performing tasks in a slightly different or-

der. The choice of first locating a word on the screen and subsequently looking for the matching picture, versus seeing a picture and finding the word that is associated with that picture, should be up to the personal preference of the student—in the same way as many would approach building puzzles differently.

Also, a process of elimination in matching the remaining words with the pictures is totally acceptable to aid the confidence of the student, as long as the overall milestones are reached prior to commencing to the next lesson.

2.4. Design for fun learning

Learning can be lots of fun. By not being overly prescriptive in *how* a task is completed, and by introducing subtle competitive metrics typically used in games, learning can be very stimulating. Using a time-based metric of completing each lesson combined with a quality metric that is loosely coupled to the pronunciation accuracy, the student can experience the excitement of playing a game whilst mastering certain aspects of a language at the same time.

In this way each student could approach the game in a different way, similar to how Angry Bird players approaches the game in different ways, focusing on accuracy or speed of completion, yet still achieving all the primary goals of the task.

2.5. Design for the developing world

Internet connectivity—at least during the key speech intensive stages of a mobile language learning game—is something that cannot be assumed for large parts of the developing world [3, 4], where the need is arguably the greatest for mastering English or other more localised languages.

With this application, the two major technologies, namely the Automatic Speech Recognition (ASR) and Text-to-speech (TTS) components, are running on the mobile device itself and does not depend on real-time Internet connectivity of any kind. For any back-end services such as performance tracking via a Learning Management System (LMS), a dependence on the Internet does exist, but such connectivity could be asynchronous [5, 6] and does not impact on the real-time learning experience of the student.

3. Demonstration

The demonstration of this mobile language learning application, employing off-line ASR and TTS technologies, will seek to exhibit some of these design aspects on an Android smartphone or tablet device.

4. References

[1] L.-H. Wong and C.-K. Looi, "Vocabulary learning by mobile-assisted authentic content creation and social meaning-making: two case studies," *Computer Assisted Learning*, vol. 26, pp. 421–433, 2010.

[2] A. Kumar, P. Reddy, A. Tewari, R. Agrawal, and M. Kam, "Improving Literacy in Developing Countries Using Speech Recognition-Supported Games on Mobile Devices," in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems (CHI)*, Austin, Texas, USA, May 2012, pp. 1149–1158.

[3] N. J. De Vries, M. H. Davel, J. Badenhorst, W. D. Basson, F. De Wet, E. Barnard, and A. De Waal, "A smartphone-based ASR data collection tool for under-resourced languages," *Submitted to Speech Communication*, 2013.

[4] E. Brewer, M. Demmer, B. Du, M. Ho, M. Kam, S. Nedevschi, J. Pal, R. Patra, S. Surana, and K. Fall, "The case for technology in developing regions," *Computer*, vol. 38, no. 6, pp. 25–38, May 2005.

[5] A. Pentland, R. Fletcher, and A. Hasson, "DakNet: Rethinking connectivity in developing nations," *Computer*, vol. 37, no. 1, pp. 78–83, January 2004.

[6] M. Khabbaz, C. Assi, and W. Fawaz, "Disruption-Tolerant Networking: A Comprehensive Survey on Recent Developments and Persisting Challenges," *Communications Surveys Tutorials, IEEE*, vol. 14, no. 2, pp. 607–640, 2012.

NTU Chinese 2.0: A Personalized Recursive Dialogue Game for Computer-Assisted Learning of Mandarin Chinese

Pei-hao Su[#1], Tien-han Yu[#], Ya-Yunn Su[*], and Lin-shan Lee[#2]

[#]Graduate Institute of Communication Engineering, National Taiwan University
[*]Graduate Institute of Computer Science and Information Engineering, National Taiwan University
Taipei, Taiwan, R.O.C

[1]r00942135@ntu.edu.tw, [2]lslee@gate.sinica.edu.tw

Abstract

We present and demonstrate a cloud-based personalized dialogue game for computer-assisted learning of Mandarin Chinese. A sequence of tree-structured sub-dialogues in restaurant scenario are linked recursively and used as the script for the game. Based on NTU Chinese, a Mandarin Chinese pronunciation evaluation software (http://chinese.ntu.edu.tw/), the user can get immediate evaluation on pronunciation, pitch, timing and emphasis and corresponding corrective feedback on each syllable as well as on sentence level for each utterance produced. The system policy is optimized to offer personalized dialogue path planning for each individual learner such that more practice opportunities are given along the dialogue path to poorly produced pronunciation units. When using the system, the learner can practice the sub-dialogues in either sequential or random order; at each dialogue turn, the learner also can choose to pronounce an arbitrary candidate sentence or following the recommended sentences by the system policy. Following the system recommendation along the sub-dialogues sequentially offers the fastest learning though. The above evaluation and learning records are displayed and stored in personal profile.

The system framework is modeled as a Markov Decision Process (MDP) with high-dimensional continuous state space considering the learning status of the learner. The dialogue policy is trained using a huge number of simulated learners generated from a corpus recorded by 278 real Mandarin Chinese learners from 36 countries with various mother tongues. The detailed principles of this system are presented in a companion paper also submitted to SLaTe 2013 [1]. This is a joint work with the International Chinese Language Program of National Taiwan University.

Reference

[1] P.-H. Su, T.-H. Yu, Y.-Y. Su, and L.-S. Lee, “A cloud-based personalized recursive dialogue game system for computer-assisted language learning,” submitted to *SLaTe*, 2013.

COMPASS III: Teaching L2 grammar graphically on a tablet computer

Karin Harbusch[1], *Johannes Härtel*[1], *Christel-Joy Cameran*[1]

[1]Department of Computer Science, University of Koblenz-Landau, Koblenz, Germany
{harbusch|johanneshaertel|cameran}@uni-koblenz.de

We demonstrate a prototype of the tablet-based L2 *grammar teaching system* COMPASS III. COMPASS stands for COMbinatorial and Paraphrastic Assembly of Sentence Structure; for a description of the underlying computational-linguistic software, see [1]. COMPASS invites the student to construct sentences by composing syntactic trees out of lexically anchored "treelets" via the graphical drag&drop user interface provided by tablet and touchscreen. After each move (i.e. each attempt to combine two treelets, or to reorder a branch), the system's natural-language generator computes all possible grammatically well-formed sentences entailed by the attempted tree. COMPASS provides positive feedback if the student-composed tree belongs to the well-formed set, and negative feedback otherwise. In the latter case, COMPASS may propose alternatives based on a comparison between the student-composed tree and its own well-formed trees (informative feedback on demand). As system feedback may explicitly refer to grammar rules, the learner needs to have elementary syntactic knowledge. COMPASS III targets L2 learners of German with high-school level understanding of word classes and grammatical functions. The user interface allows the student to select words and to move (parts of) trees around through finger or stylus gestures. No typing is required. COMPASS III focuses on word order and case morphology—difficult topics in L2 German.

The grammar formalism underlying COMPASS is *Performance Grammar* [2], which assumes separate rules for the hierarchical *structures* of a sentence and the *linear order* of its constituents. This split allows the student to break sentence construction exercises into relatively small parts. For instance, the learner can select a word and inflect it according to the intended grammatical function without having to worry about the linear position of the constituent in the sentence under construction. At any time during this "scaffolded" sentence construction process, the tree built so far remains visible on the screen, ready to be expanded by attaching additional words/treelets; any earlier decision can be undone and corrected.

Fig. 1 shows a student action which is rejected by the generator due to incompatibility of accusative case of the Direct Object (DOBJ) and dative case of the personal pronoun *ihm* 'him'. The red circles denote NPs; the arrow indicates that the student attempts to attach (unify, merge) the NP dominating *ihm* with the NP that fulfills the function of Direct Object. The example presupposes that the student has already successfully constructed the hierarchical structure for *Anja baut* 'Anja builds'. COMPASS does not allow the attempted attachment, though, and the two red circles do not fuse. In response to a request, the system may suggest *ihn* as the pronouns with correct accusative case. The purple shapes serve as receptacles for linearly ordered constituents. The student can drag circles into these squares (which, in response, may expand to rectangles) and place them there in any order.

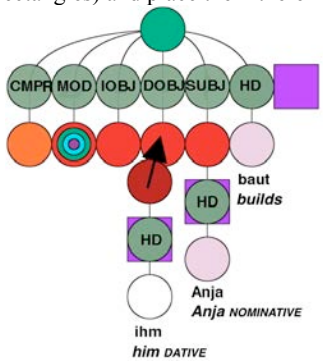


Figure 1. *Failed attempt to combine two treelets, due to missing morphological case agreement. Circles without text stand for syntactic categories (clauses, phrases, word classes).*

Fig. 2 depicts the system's reaction to the incorrect linear order of Direct and Indirect Object in *Anja baut eine Rakete ihm* 'Anja builds a rocket for him'. The student has moved four grammatical function nodes into the purple rectangle associated with the Head Verb. Although the order of DOBJ and IOBJ is wrong, COMPASS does not reject it outright but indicates an error by changing the color to yellow; this allows the student to continue with the hierarchical structure, and to return to the linear order at a later time. Due to space limitations, we cannot describe here how COMPASS "teaches" even complicated linear order rules.

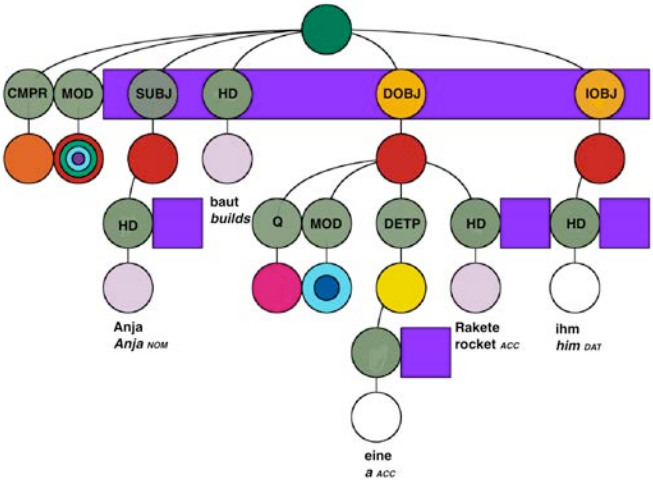


Figure 2. *Incorrect linear error in '*Anja baut eine Rakete ihm'.

Index Terms: e-learning, grammar teaching, German as L2

[1] Harbusch, K. and Kempen, G. "Automatic online writing support for L2 learners of German through output monitoring by a natural-language paraphrase generator ", in M. Levy, Blin, F., Bradin Siskin, C. and Takeuchi, O. [Eds], *WORLDCALL*, New York: Routledge, 2011, pp. 128-143.

[2] Kempen, G. and Harbusch, K. "Performance Grammar: A declarative definition", in M. Theune, Nijholt, A. and Hondorp, H. [Eds], *Computational Linguistics in the Netherlands 2001*, Amsterdam: Rodopi, 2002.

A Suite of Mobile Applications to Assist Speaking at Right Speed

Imran Ahmed, Meghna Pandharipande, Sunil Kumar Kopparapu

TCS Innovation Labs - Mumbai, Yantra Park, Thane (West), Maharashtra, INDIA

{ahmed.imran, meghna.pandharipande, sunilkumar.kopparapu}@tcs.com

Abstract

One of the prominent reason for ineffective communication in call center telephone conversations is primarily due to the *manner* in which the voice agents speak and not necessarily due to *what* they speak. Speaking rate, a non-linguistic aspect of speech, is a critical factor affecting intelligibility and comprehension of speech in general and specifically in call center telephone conversations. There have been attempts to monitor speaking rate of agents in a call center setup. However, there has been a paradigm shift to the conventional call centers with companies opting for home-agents. In this model the agents, unlike in the current call center setup, can operate virtually from anywhere using their mobile phones. In this paper, we present *SpeakRite* - a suite of Android mobile applications that can assist home-agents. While one of the components of *SpeakRite* analyzes the speaking rate during an ongoing telephone conversation and provides a real time feedback to assist the speaker modify his speaking rate, there are several other components that allow the home-agent to assess their speaking rate and learn to speak right themselves.

Index Terms: speaking rate, speech rate, mobile application, home-agents

1. Introduction

In voice based call centers the effectiveness of telephone conversations between the agent and the customer is determined more by the manner in which the agent speaks and is less dependent on what the agent speaks [1]. Ineffective communication in telephone conversations is especially prominent when people from different geographies or cultures converse in a language common to them like in the case of a voice based call center, where the call center voice agent and the customer can be from two different geographies. Speaking rate, a non-linguistic feature of speech is a critical factor affecting intelligibility and comprehension of speech [2, 3, 4, 5]. It is well known that speaking rate varies across native and non-native speakers of a language [4, 5, 6]. This variation of speaking rate additionally affects the comprehension of speech in native and non-native listeners. This observation seems to suggest that it is important that speakers converse at an optimal speaking rate for an effective conversation. Good communicators continuously monitor their speaking rate. They consciously adjust their conversational pace to get the message across effectively and efficiently. However monitoring and maintaining the speaking rate at the desired levels, may be hard for an average person who is not conscious of the rate at which he is speaking or in a very practical setup is in an emotional state that does not allow him to concentrate on the speaking rate.

Automatic monitoring of speaking rate by analyzing speech in real time can help speakers speak at the right speed and make conversations effective. In [7] a server based speaking rate monitoring tool was proposed which assists call center voice agents, in real time, to maintain an optimal speaking rate. As discussed in [7] the signal processing of speech happens on a server making it real time. However, driven by a number of business and socioeconomic parameters there has been a paradigm shift in what we know as call centers today in the form of virtual call centers or home-agents [8, 9]. A home-agent, unlike a call center voice agent, works solely from home, with no office space at a company facility. Voice agents are required to take calls on their mobile phone to converse with the customers to answer their queries. While the concept of home-agents has been practiced for more than a decade, the pace of adoption has accelerated in the past few years, however this has been restricted to non-voice agents. It is predicted [10] that more than three hundred thousand home-based agents will be working in the United States by 2013. However, the most obvious challenge with home-based agents is that a supervisor cannot simply walk over to an agent to supervise [8]. The ability to act directly and spontaneously is not possible in the changed scenario of home-agent call center paradigm requiring remote monitoring and remote coaching, preferably through an automatic process.

In this paper, we present a suite of mobile phone applications which assists home-agents go about performing their tasks efficiently. The suite of applications contain a real time speaking rate monitoring tool, which can monitor the speaking rate of the mobile phone user, first presented in [11] and discussed here for the sake of completeness, during an ongoing conversation. The suite of applications contain some off-line tools which assist the home-agent to practice to speak at the right speed and a tool that can evaluate the performance of the home-agent after a conversation, thereby allowing the home-agent to identify scope for improvement.

The rest of the paper is organized as follows: In Section 2 we briefly discuss automatic speaking rate monitoring. We present the suite of applications which assist the home-agent perform efficiently in Section 3 and conclude in Section 4.

2. Automatic Speaking Rate Monitoring

In [7] a tool that computes speaking rate in real time was proposed. The tool assists the call center agent to maintain an optimal speaking rate in real time by providing a just in time feedback to the agent. They used the count of the syllables detected in speech as a measure to compute the speaking rate. The algorithm described in [12], modified to work for real time, was used to detect syllable nuclei in spoken speech. The syllables in spoken speech were detected as:

Step 1. Identify all the intensity peaks that are preceded and followed by an intensity dip.

Step 2. Of all the intensity peaks detected in Step 1, retain only those intensity peaks that are above a certain intensity threshold and mark them.

Step 3. Discard the intensity peaks, after Step 2., that are unvoiced. The voiced intensity peaks are the syllable nuclei.

Once the speaking rate is computed in terms of number of syllables per second (sps), we can compute the speaking rate in words per minute (WPM) using a multiplication factor, namely, $\mathcal{S}_{WPM} = \gamma \times \mathcal{S}_{sps} \times 60$, where $\mathcal{S}_{WPM}$ is the speaking rate in words per minute, $\mathcal{S}_{sps}$ is the number of syllables per second and γ is a constant that captures the average number of syllables per word; this depends on the language being spoken [13]. For English language $\gamma = 1.5$ as suggested in [14]. It must be noted that since the measure is based on count of syllables in speech, same technique of computing speaking rate can be applied across different languages. Thus the speaking rate measurement tool can be used across call centres serving in different languages. The same syllable detection procedure when implemented on a Android mobile phone with a 650 MHz processor imposes enormous computational load and fails to operate in real time. In order to enable real time operation on a mobile phone, the syllable detection algorithm was modified as discussed in [11].

3. SpeakRite Mobile Application Suite

SpeakRite is a suite of mobile applications useful for a voice home-agent. The tools available for the home agent are:

1. Tool to compute the speaking rate in real time even while the speaker is in a live conversation. Figure 1 shows a screen-shot of the tool monitoring the speaking rate of the user, while a call is active.
2. *SpeakRite* tool set can also provide a complete analysis of the speaking rate for the entire duration of a recorded call. Figure 2 shows a screen-shot of the *SpeakRite* showing the variations in speaking rate for the entire call duration.
3. *SpeakRite* allows the user to manually define a desired speaking rate as a reference. This reference can be used to give a feedback to the user in terms of how he is speaking relative to the set speaking rate. Optionally, the reference value can be remotely set by the call center or the user can record any speech and set its average speaking rate as the reference.

Figure 1: Monitoring Speaking Rate in real time.

Figure 2: Graph of Speaking Rate Variations.

Figure 3: User's speech time-scaled to modify Speaking Rate

4. *SpeakRite* gives an option to modify user's speaking rate. In this option, the user is asked to read a pre scripted text displayed on the mobile phone screen. As the user speaks the speaking rate is analyzed. If the average speaking rate of this recording is different from the pre-defined reference then the speech recording is time-scaled, using WSOLA time-scale modification technique [15], such that the resulting speech has an average speaking rate equal to the reference speaking rate. This time-scaled speech is played back to the user to give him a feedback of his speech at the desired reference speaking rate. The WSOLA time-scale modification is implemented on Android using the SoundTouch Audio Processing Library [16]. Figure 3 shows a screen-shot of the *SpeakRite* application depicting the user recording before and after time-scale modification.

5. *SpeakRite* allows the user to self train. In this mode, a moving ticker text is displayed on the user's mobile phone screen and the user is asked to read out the text as it is scrolling. The speed of the scroll is varied to allow the user to practice speaking at different speaking rates. This facilitates providing the user an instant feedback of his current speaking rate by movement of the speedometer needle on the mobile phone screen. Figure 4 shows a screen-shot of the *SpeakRite* application useful for training.

6. *SpeakRite* application also facilitates the users to check their progress as they train to speak at a particular rate. Figure 5 shows a screen-shot of the *SpeakRite* application depicting the distribution of the speaking rate of the user before and after training. This enables users to monitor their own progress.

Figure 4: Training the users Speaking Rate with Ticker Text moving at a changing speed

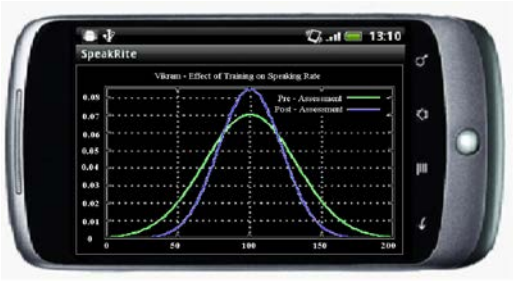


Figure 5: Distribution of Speaking Rate before training and after training

4. Conclusion

For effective communication during a telephone conversation it is important that the speakers converse at an optimal speaking rate and learn to adjust their conversational pace to make themselves intelligible and comprehensible. With the advent of voice based home-agents it is necessary to provide the home-agents with a suite of tools that will allow them the necessary support to enable them perform their task efficiently. SpeakRite is a suite of mobile phone applications on an Android platform that provides this support. The suite consists of both real time and off-line applications. The real time application allows monitoring the speaking rate of the speaker and also gives an instant feedback of the current speaking rate. This real time automatic monitoring of Speaking Rate can sensitize speakers to their current speaking rate thereby allowing them to speak at the desired speaking rate. The suite has applications that provide off-line analysis of several aspects of speech which help the home-agent to practice, train and monitor his progress in achieving to speak at the desired speaking rate.

5. References

[1] "TCS e-Serve BPO," personal communication, 2011.

[2] J. C. Krause and L. D. Braida, "Investigating alternative forms of clear speech: The effects of speaking rate and speaking mode on intelligibility," *The Journal of the Acoustical Society of America*, vol. 112, no. 5, pp. 2165–2172, 2002. [Online]. Available: http://link.aip.org/link/?JAS/112/2165/1

[3] G. WEINSTEIN-SHR and R. GRIFFITHS, "Speech rate and listening comprehension: Further evidence of the relationship," *TESOL Quarterly*, vol. 26, no. 2, pp. 385–390, 1992. [Online]. Available: http://dx.doi.org/10.2307/3587015

[4] J. Murray and M. Tracey, "The effects of speaking rate on listener evaluations of native and foreign-accented speech," *Language Learning*, vol. 48, pp. 159–182, June 1998.

[5] A. Bradlow and D. Pisoni, "Recognition of spoken words by native and non-native listeners: talker-, listener-, and item-related factors." *The Journal of the Acoustic Society of America*, pp. 2074–85, Oct. 1999.

[6] A. Janet and K. Kenneth, "The effect of foreign accent and speaking rate on native speaker comprehension," *Language Learning*, vol. 38, pp. 561–613, December 1988.

[7] M. Pandharipande and S. Kopparapu, "Real time speaking rate monitoring system," in *International Conference on Signal Processing, Communications and Computing (ICSPCC)*, Sept. 2011, pp. 1–4.

[8] "Home-based agents and workforce optimization," verint Systems White Paper by Bill Durr. [Online]. Available: http://telus.com/en_CA/content/pdf/products/Call_Centres/HomeBasedAgents_TELUS_0609.pdf

[9] "Go green with home agents." [Online]. Available: http://www.avaya.com/uk/resource/assets/whitepapers/ggha_pwp.pdf

[10] S. Loynd, *U.S. Home-Based Agent 2009-2013 Forecast: The Enigma of Arrival*. IDC, 2009.

[11] I. Ahmed, M. Pandharipande, and S. K. Kopparapu, "Speakrite: Monitoring speaking rate in real time on a mobile phone," *International Journal of Mobile Human Computer Interaction*, vol. 5(1), pp. 62–69, Jan-Mar 2013.

[12] N. H. D. Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically," *Behavior Research Methods*, vol. 41, pp. 385–390, 2009.

[13] P. Francois, C. Christophe, and M. Egidio, "Across-language perspective on speech information rate," *Language*, pp. 539–558, Sept. 2011.

[14] J. S. Yaruss, "Converting between word and syllable counts in children's conversational speech samples," *Journal of Fluency Disorders*, vol. 25, no. 4, pp. 305 – 316, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0094730X00000887

[15] M. R. W. Verhelst, "An overlap-add technique based on waveform similarity (wsola) for high quality time-scale modification of speech," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1993, pp. 554 – 557.

[16] "Soundtouch audio processing library." [Online]. Available: http://www.surina.net/soundtouch

Automatic Detection of the Words that will Become Unintelligible through Japanese Accented Pronunciation of English

Teeraphon Pongkittiphan[1], *Nobuaki Minematsu*[1], *Takehiko Makino*[2], *Keikichi Hirose*[1]

[1]The University of Tokyo, Tokyo, Japan
[2]Chuo University, Tokyo, Japan

{teeraphon,mine,hirose}@gavo.t.u-tokyo.ac.jp, mackinaw@tamacc.chuo-u.ac.jp

Abstract

This study examines automatic detection of the words that will be unintelligible if they are spoken by Japanese speakers of English. In our previous study [1], 800 English utterances spoken by Japanese speakers, which contained 6,063 words, were presented to 173 American listeners and correct perception rate was obtained for each spoken word. By using the results, in this study, we define the words that are *very unintelligible* through Japanese accented English pronunciation and also define the words that are *rather unintelligible*. Then, by using Classification And Regression Tree (CART) with linguistic features and lexical features only, we examine automatic detection of these words. After that, we introduce an additional feature derived by considering phonological and phonotactic differences between Japanese and English. This additional feature is found to be very effective and our proposed method can detect *very unintelligible* words and *rather unintelligible* words automatically with F1-scores of 65.44 and 70.45 [%], respectively.

Index Terms: speech intelligibility, second language learning, foreign accent, ERJ database, CART

1. Introduction

English is the only one common language for international communication. Statistics show that there are about 15,000 millions of users of English but only a quarter of them are native speakers, while the rest of them are speaking English with foreign accent [2]. This clearly indicates that foreign accented English is more globally spoken and heard than American or British English. Even in the case of native speakers, their English sometimes becomes unintelligible to non-native listeners because speech intelligibility depends on various factors including the nature of listeners [3].

Although it has been a controversial issue which of native-sounding pronunciation and intelligible enough pronunciation should be the target of English pronunciation learning. Recently, the concept of World Englishes [4] is more and more widely accepted by teachers, where it is claimed that, instead of mastering native-like pronunciation, foreign accented pronunciation is acceptable if it is intelligible enough. However, the pronunciation intelligibility is difficult to define because it depends on various factors e.g. the language background of listeners, the speaking context and the speaking proficiency of a speaker [5] [6].

It is known that Japanese learners tend to have poorer speaking skill of English than learners in other Asian countries. One possible reason is there are big differences in the phonological and phonotactic system between Japanese and English. Therefore, when Japanese learners have to repeat after the English teacher, many of them don't know well how to repeat. In other words, it is difficult for learners to know what kind of mispronunciations are more fatal to the perception of listeners.

Saz et al. [7] proposed a Basic Identification of Confusable Contexts (BICC) technique to detect the minimal-pairs-based confusable context in a sentence, which might lead to a miscommunication. The subjective evaluation was done by letting subjects read the sentences modified by altering minimal pairs and rate how confusable each sentence is. However, this reflects a lexical and textual confusion perceived by reading sentences not by hearing spoken utterances.

To end this, in this study, by using the results of intelligibility listening tests [1], for given English sentences, we propose a method of automatically detecting the words that will be unintelligible to American listeners if those words are spoken with Japanese accent.

2. ERJ intelligibility database

Minematsu *et al*. [1] conducted a large listening test, where 800 English utterances spoken by Japanese (JE) were presented to 173 American listeners. Those utterances were carefully selected from the ERJ (English Read by Japanese) speech database [8]. The American listeners were those who had no experience talking with Japanese and asked to listen to the selected utterances and immediately repeat what they just heard. Then, their responses were transcribed word by word manually by experimenters. Each utterance was heard by 21 listeners on average and a total of 17,416 transcriptions were obtained. In addition to JE utterances, 100 English utterances spoken by speakers of general American English (AE) were used and their repetitions were transcribed in the same way.

Following that work, in this study, an expert phonetician, the third author, annotated all the JE and AE utterances with IPA symbols. The IPA transcription shows what is phonetically happening in each of the JE and AE utterances. It would be very interesting to observe the phonetic differences between a JE utterance and an AE one of the same sentence and analyze the IPA transcriptions of the JE utterances, which shows misperceptions, based on the phonetic differences. However, it is a pity that the sentences in the JE 800 utterances and those in the AE 100 ones are not overlapped well. So, the above analysis is currently difficult to realize, but the IPA transcriptions of the 900 utterances and the 17,416 word-by-word transcriptions, i.e. misperceptions, will be included in the next release of the ERJ.

Then in this paper, by using the results of the listening test, we firstly define the words in the read sentences that became *very unintelligible* or *rather unintelligible* due to Japanese accent. Next, we investigate automatic detection of those words only by using their lexical and linguistic features, that can be extracted without referring to actual utterances. If detection suc-

ceeds, the proposed method is able to show which words of a presentation manuscript Japanese learners should be very careful of to make their English oral presentation more intelligible.

3. Detection of "will-be-unintelligible" words

3.1. Definition of "will-be-unintelligible" words

The ERJ contains the pronunciation proficiency score (1.0 to 5.0) for each speaker, which was rated by five American teachers of English. To focus on the listening test results of only typical Japanese speakers, we removed the data of too poor speakers (<2.5) and those of too good speakers (>4.0). The resulting data had 756 utterances and 5,754 words in total.

As described in Section 2, each spoken word was heard by 21 American listeners on average and the correct perception rate was obtained for each. In this study, to describe the word perception qualitatively, the words whose perception rate is less than 0.1 are defined as *very unintelligible* due to Japanese accent and the words whose rate is less than 0.3 are defined as *rather unintelligible*. The occupancies of very unintelligible and rather unintelligible words were 18.9% and 34.2%, respectively. The aim of this study is automatic detection of these words by using only lexical and linguistic features.

3.2. Preparation of features for automatic detection

From preliminary experiments, we found two things. 1) Since we wanted a binary (intelligible/unintelligible) classifier of input data, we firstly trained CART as binary classifier but results were not good. Then, we trained CART as predictor of perception rate of each word and, comparing the output to a threshold, binary classification was made possible. We found this strategy to be effective. 2) Since we wanted to train CART distinctively between intelligible words and unintelligible words, we intentionally removed words of intermediate level (0.4 to 0.6) of perception rate only from training data. This removal was effective although those data were actually included in testing data.

The features used for CART-based detection were prepared by using the CMU pronunciation dictionary and the n-gram language models trained with 15 millions words from the OANC text corpus [9]. Table 1 shows these features that are categorized into 3 groups; lexical, linguistic and other features.

The feature [C], which is the maximum number of consecutive consonants in the word, is derived by considering Japanese pronunciation habits of English that is caused by phonological and phonotactic differences between the two languages. The smallest unit of speech production in Japanese is called mora, which has the form of either CV or V. However, consecutive consonants, with the form of CCV or CCCV, are very common in English. Japanese speakers sometimes insert an additional vowel after a consonant, which increases the number of syllables in that word and is expected to decrease the intelligibility of that word easily, e.g. the word 'sky' (S-K-AY) is often pronounced as (S-UH-K-AY), where additional UH vowel is added.

3.3. Experimental results

We have three kinds of features; [A], [B], and [C] and have two levels of "will-be-unintelligible" words; very unintelligible and rather unintelligible. Table 2 shows the results of precisions, recalls, and F1-scores of 10 cross-validation experiments.

By using only either lexical [A] or linguistic [B] features, each method has low F1-scores, while combination of [A] and [B] can increase the F1-score significantly to 53.39% and 63.92% for *very* and *rather unintelligible* words, respectively. An interesting finding is that, when adding the last feature, the maximum number of consecutive consonants, the F1-score is improved significantly again from 53.39% to 65.44% and from 63.92% to 70.45% for each case.

Table 1: *The features prepared for CART*

[A] lexical features for a word
#phonemes in the word
#consonants in the word
#vowels (=#syllables) in the word
forward position of primary stress in the word
backward position of primary stress in the word
forward position of secondary stress in the word
backward position of secondary stress in the word
word itself (word ID)
[B] linguistic features for a word in a sentence
part of speech
forward position of the word in the sentence
backward position of the word in the sentence
the total number of words in the sentence
1-gram, 2-gram and 3-gram score of the word
[C] phonological and phonotactic feature for a word
the maximum number of consecutive consonants

Table 2: *Precisions, recalls, and F1-scores[%]*

		[A]	[B]	[A][B]	[A][B][C]
very unintelligible	P	44.19	42.42	60.67	74.01
	R	3.71	22.70	47.68	58.64
	F1	6.85	29.58	53.39	65.44
rather unintelligible	P	57.04	57.08	70.21	73.72
	R	11.02	45.12	58.66	67.46
	F1	18.48	50.49	63.92	70.45

The precisions in the table claim that almost 75% of the words that were identified as *very* or *rather* unintelligible are correctly detected. As described in Section 3.1, the occupancies of *very* and *rather* unintelligible words were 18.9% and 34.2%, which correspond to the precisions when detecting unintelligible words randomly. Considering these facts, although no acoustic observation is used in our proposed method, it can detect "will-be-unintelligible" words very effectively.

However, we do not claim at all that acoustic observation is ineffective. We're very interested in improving the detection performance by using acoustic and phonetic information. For that, we now continue annotating additional utterances to get IPA transcriptions of the sentence utterances of AE to get complete overlap between JE and AE. With these transcriptions, we can add IPA-based phonetic features to improve the detection performance. We're also interested in replacing manual IPA-based features with features obtained automatically by ASR.

4. Conclusions

We investigated the possibility of automatic detection of unintelligible words that would be misrecognized by native listeners due to Japanese accent. The proposed method can automatically and effectively detect unintelligible words even using only the information extracted from text, not using any acoustic information. In the future, acoustic and phonetic information will be used for performance improvement.

5. References

[1] N. Minematsu et al., “Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) Database”, Proc. Interspeech, pp. 1481-1484, 2011.

[2] Y. Yasukata., “English as an International Language: Its past, present, and future”, Tokyo: Hitsujishobo, pp. 205-227, 2008.

[3] J. Flege., “Factors affecting the pronunciation of a second language”, Keynote of PLMA, 2002.

[4] J. Jenkins., “The phonology of English as an international language”, Oxford University Press, 2000.

[5] D. Crystal, “English as a global language”, Cambridge University Press, New York, 1995.

[6] J. Bernstein., “Objective measurement of intelligibility”, Proc.ICPhS, pp. 1581-1584, 2003.

[7] O. Saz and M. Eskenazi., “Identifying confusable contexts for automatic generation of activities in second language pronunciation training”, Proc. SLaTE, 2011.

[8] N. Minematsu et al., “Development of English speech database read by Japanese to support CALL research”, Proc. Int. Conf. Acoustics, pp. 557-560, 2004.

[9] The Open American Nation Corpus (OANC), http://www.anc.org/data/oanc/.

Fusing Eye-gaze and Speech Recognition for Tracking in an Automatic Reading Tutor – A Step in the Right Direction?

Morten Højfeldt Rasmussen [1], *Zheng-Hua Tan*[2]

[1]SpeechOp ApS, Aalborg, Denmark
[2] Department of Electronic Systems, Aalborg University, Aalborg, Denmark
mr@speechop.com, zt@es.aau.dk

Abstract

In this paper we present a novel approach for automatically tracking the reading progress using a combination of eye-gaze tracking and speech recognition. The two are fused by first generating word probabilities based on eye-gaze information and then using these probabilities to augment the language model probabilities during speech recognition. Experimental results on a small dataset show that the tracking error rate of the system using only speech recognition is 34.9% whereas the tracking error rate for the system that incorporates eye-gaze tracking into the speech recognizer is 31.2% – a relative improvement of 10.6%.

Index Terms: automatic reading tutor, eye-gaze tracking, speech recognition

1. Introduction

The tasks of automatic reading tutors (ART) are many. It might provide live feedback to the reader if he or she pauses during a reading session, which could indicate that the reader has trouble reading a word. It might also provide corrective feedback if the reader misreads a word. The feedback provided could be in the form of a picture (if possible), reading the word in question out loud, or a number of other ways [1]. It could also be a platform for reading assessment [2] or provide a speech driven interface to the reader [3].

In order to automatically provide feedback or assess reading proficiency the reading tutor needs to detect some level of reading activity. At least three modalities can be used to this end: eye-gaze (or gaze), speech, and manual feedback requests. For example in [4] the authors use an eye-gaze tracker to provide assistance when the reader looks at a word for more than 360 msec and in [1], [2], and [3] the authors use automatic speech recognizers (ASR) in three different reading tutors. Manual requests for feedback can be done using the mouse; however, systems that provide this as the only way to get feedback fall in the category of interactive books rather than ART.

In this paper we focus on the domain of adult dyslexic read speech. Tracking the reading progress of people with dyslexia is challenging as they produce more miscues (misread words and other disfluencies [5]) than people without this developmental reading disorder.

Reading usually occurs in a progressive way. However, sometimes the reader returns to previously read words in order to revise or remember what was read. This is especially true for people with dyslexia who struggle with reading [1]. This means that an ART should be able to determine which word the reader is supposed to read next, in order to provide assistive feedback if necessary. In [6] we detailed a speech recognition based tracking system. In this paper we extend that work on tracking by using an eye-gaze tracker and fusing gaze and speech.

To the authors' knowledge, no prior research has been done with regards to fusing gaze information and speech recognition for tracking reading. In [7] N-Best lists generated from a speech recognizer are rescored based on gaze points for a visual-based goal-driven task. In the experiment the participant describes a geographical map with landmarks to another person. The landmarks are placed relatively far from each other. The N-Best lists were used as a substitute for implementing the rescoring in the Viterbi decoding.

The rest of the paper is organized as follows: In Section 2 we present our tracking methods, in Section 3 we describe the data collection and transcription, in Section 4 we show and discuss the results, and in Section 5 we conclude the work.

2. Tracking

In this section we'll explain the relationship between gaze and reading and how we track the reading progress using an eye-gaze tracker, an ASR and a fusion of gaze and speech.

2.1. Eye-gaze tracking and reading

During reading the eyes move in a sequence of fixations and saccades [4]. Usually the saccades move from left to right but sometimes they do the opposite – for example when the reader revises what was previously read.

2.1.1. Gaze events

Gaze events can be ordered into a number of categories. The authors of [8] list the following: fixation, glissade, saccade, smooth pursuit, and blink. For this paper we will focus on fixations as they can be thought of as the anchor points of the gaze events.

A fixation can be defined in different ways. One way is to define it as a given period where the gaze points generated from an eye-gaze tracker falls within a region. Another way is to define it as gaze points that are not classified as saccades, glissades or noise as in [8]. For this paper we use the Tobii Fixation Filter [9], which finds fixations by grouping (or clustering) gaze points using the method described in [10]. An example of a saccade and two fixation events mapped onto the text being read is shown in Figure 1.

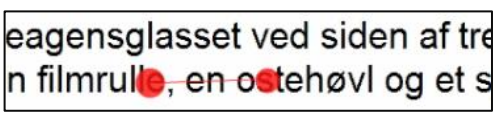


Figure 1: *Three gaze events: two fixations (the red discs) and one saccade (the red line connecting the two discs). The reader is reading the upper line, but the eye-gaze tracker offsets the gaze to the lower line.*

2.1.2. Sources of errors

Ideally the fixations will map onto the lines being read but tracking errors add noise to the gaze points. According to [11] there are at least three sources of errors when determining what is being focused on using an eye-gaze tracker: the quality of the setup, drift of the calibration, and the biological characteristics of the eye. The quality of the setup depends on the eye-gaze tracker used and how successful the session calibration is. But even if the calibration is successful the accuracy will usually degrade during a session due to drifting errors that can occur e.g. when changes in head-position are incorrectly compensated for. The biological characteristics of an eye give a visual field of focus of around 1°.

The effect of the described errors is an offset plus noise on the gaze points. An example of offset error is shown in Figure 1. The reader is reading the text in the upper line, but the eye-gaze tracker maps the gaze onto the lower line.

2.1.3. The word being focused on

The task of mapping the gaze points to text words is not just a matter of finding the nearest word given a gaze point due to the errors described in Section 2.1.2. The authors of [11] introduce the notion of "sticky" and "magnetic" lines, where sticky lines keeps the focus to the (assumed) correct line and magnetic lines sets the focus to the next line when a gaze point jump from the end of a line to the beginning of the next is detected. Sticky and magnetic lines alleviate some of the errors in the vertical direction. Horizontal errors are less pronounced – which is fortunate since they would be harder to detect.

2.1.4. Tracking the reading using eye-gaze tracking

The magnetic lines method described above was used as inspiration to our eye-gaze tracking algorithm. However, we only use the information of which line is being focused on to estimate the per-line offset error in the vertical direction. We calculate the offset for each line, since the magnitude of the offset varies as a function of the y-position.

We define "line-index" as the index of the line we believe that the reader is focusing on. We assume that the reader starts reading from the first word and onwards and sets line-index to 1. A "next-line event" is detected, when the center of the i'th fixation cluster (the collection of fixation points belonging to one fixation) is near the end of a line and the i+1'th fixation cluster is near the beginning of the next line. The procedure for determining the word being focused on can be described like this:

1. *Get the next fixation cluster from the eye-gaze tracker.*
2. *Update the per-line offset estimate for the y-axis.*
3. *Calculate the center point (or mean) of the fixation points belonging to that cluster.*
4. *Subtract the y-axis offset from the center point.*
5. *Find the text word closest to the value calculated in 4.*
6. *If a next-line event is detected: increment the line-index.*
7. *GoTo 1.*

2.2. Speech recognition and tracking

Our previous work [6] within the area of tracking in ARTs involved using speech as the only modality for tracking. In that paper we showed that a language model that models the expected reading behavior of children – allowing for jumping back and forth in the text – works better than changing the task from trying to follow the child to forcing a strict left-to-right reading policy.

In this paper we also use a language model that allows for jumping back and forth in the text. This model is further relaxed to a word-loop; essentially giving word transitions equal probability. This relaxation has been chosen in order to accentuate the effect of including gaze information in tracking. Each word was given a unique ID in order to differentiate words in the same text segment with the same spelling.

2.3. Fusing gaze and speech for tracking

This section describes how we handle synchronization issues between gaze and speech, calculate word probabilities from gaze points and apply them in the speech recognizer.

2.3.1. Synchronizing gaze and speech

No matter how accurately the gaze points and the audio stream are synchronized, there will always be a non-deterministic delay from the time the reader focused on a word to when it was uttered. This delay is at least the time it takes for the reader to see, interpret, and utter the word. The authors of [7] report this delay to be typically between 430 and 902 msec. In this paper we don't explicitly try to synchronize gaze and speech. Instead we delay the gaze points by 430 msec because we want to avoid overshooting the reference changing points due to the way we calculate the word probabilities from the gaze points.

2.3.2. Word probability from gaze points

Similar to [11] we want to update the language model based on the gaze information. Instead of making a hard decision on which word is being focused on as in Section 2.1.4 (bullet point *5.* in the list) we choose to assign a probability to all the text words at a given time. Given a fixation cluster, the probability for each text word is calculated from a bivariate normal distribution with the probability density function given as:

$$f(\boldsymbol{x}_n, \boldsymbol{\mu}_t) = \frac{1}{2\pi} \det(\boldsymbol{\Sigma})^{-\frac{1}{2}} \, e^{\left(-\frac{1}{2}(\boldsymbol{x}_n - \boldsymbol{\mu}_t)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x}_n - \boldsymbol{\mu}_t)\right)}, \quad (1)$$

where $\boldsymbol{x}_n$ is the coordinate for the center-point of the *n*'th word we want the probability for, $\boldsymbol{\mu}$ is the mean vector (center point) consisting of the mean of the x- and y-coordinates of the gaze points belonging to the fixation cluster. $\boldsymbol{\Sigma}$ is the covariance matrix of all fixation clusters and is estimated as a running average over an entire session. An illustration of the normal distribution can be seen in Figure 2. Since, however, the speech lags behind the gaze we estimate the probability of word *n* as:

$$P_t^*(w_n) = f(\boldsymbol{x}_n, \boldsymbol{\mu}_t) + f(\boldsymbol{x}_n, \mathbf{c}_{n-1}) \quad (2)$$

where P_t^* is the un-normalized word probability for word *n* at time *t* calculated from the gaze information and $\mathbf{c}_{n-1}$ is the center point of word *n*-1. $\mathbf{c}_{n-1}$ is estimated as the word index of the word closest to $\boldsymbol{\mu}_t$ minus one. The word probabilities are then normalized in order to ensure that their sum equals 1.

Linear interpolation of the word probabilities is used for periods with no fixations.

Figure 2: *Illustration of the estimated bivariate normal distribution.*

2.3.3. Integrating gaze information in the ASR

With the word probabilities calculated based on the gaze information the only thing left is to apply them during recognition. To that end, we have modified Sphinx-4 slightly so that whenever it encounters a word in the recognition lattice, it multiplies (which becomes an addition in the log domain) the gaze probability (P_t^*) of that word at the given time by the ASR probability of the search path that ends in the word.

3. Data collection and transcription

In this section we will describe the data collection and transcription.

3.1. Data collection and experiment setup

The setup for the data collection can be seen in Figure 3. The participant was equipped with a headset and was told to read the displayed text out loud. The eye-gaze tracker was calibrated before each session which took roughly half a minute.

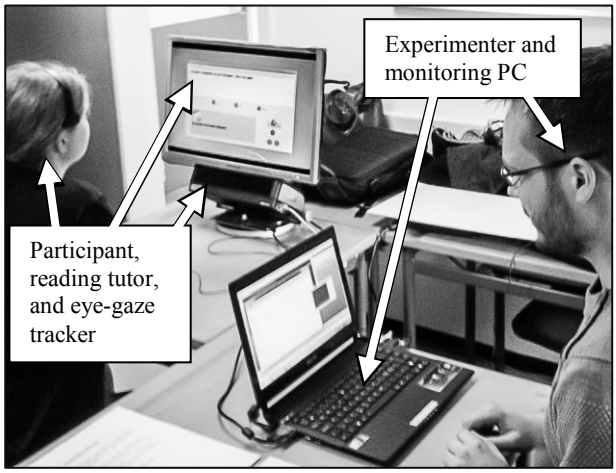


Figure 3: *The experiment setup.*

After the calibration the participant was given control of the mouse and began the reading session. Each participant read the same 23 short text segments of varying difficulty [1]. The experimenter was monitoring a live view of the eye-gaze tracking during each session and would tell the participant to adjust his or her position if the head had moved too far away from the optimal zone. Some dropped gaze points were observed, which was expected.

The data was collected for four adults with dyslexia.

3.2. Transcribing the data

Each session was time-segmented into words and miscues. Each speech event was assigned a target word index, indicating the word position in the prompt (or target) text. These word indices were then used to generate the tracking reference. Reference changing points were placed just after correctly read words were uttered. Miscues were filtered out.

4. Results and discussions

In this section we describe the evaluation method and present and discuss the results.

4.1. Evaluation method

The evaluation method used in this paper builds on [6] but moves away from the notion of speech events (segments with either a word or miscue) to word index changes – that is whenever the focus moves from one word to another. Since the tracking reference was generated by a human who applies judgment when placing changing points, a tolerance interval is introduced similar to that in [12]. Another reason for using a tolerance interval is that even though the reading tutor might be slightly off in detecting a changing point this has no practical significance in most settings.

Given a tolerance interval, the tracking error rate is calculated as the total number of changing points in the reading tutor's output minus correct changing points that fall within ±tolerance of the reference point. In this work a tolerance of ±250 msec was applied.

An illustration of the tolerance interval is shown in Figure 4; where *T* is the tolerance value, the two blue discs are the reference changing points, and the three red diamonds are the changing points detected by the system (hypothesized changing points). The two rightmost hypothesized changing points are tracking errors in the example. The leftmost hypothesized changing point falls within the tolerance interval of a reference changing point with the same word ID and is therefore marked as being correct. No hypothesized changing point falls within the tolerance interval of the leftmost reference point, which results in a tracking error. The tracking error rate for the example is 3/2 = 1.5.

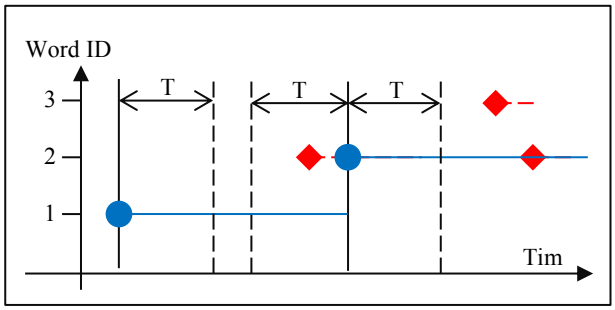


Figure 4: *Illustration of the tolerance interval. T is the tolerance value. The two blue discs are the reference changing points; the three red diamonds are the changing points detected by the system.*

4.2. Tracking error rate

The test data contains 1099 changing points in 92 utterances. The tracking error rates (TER) can be seen in Table 1. Note that the relative improvement of 10.6% of the system using gaze information and speech recognizer compared to the system that only uses the speech recognizer is significant, as the P-value of the matched-pairs test described in [13] is 0.042.

Table 1: *Tracking error rates.*

Tolerance	ASR	Gaze+ASR	P-value
±250 msec	0.349	0.312	0.042

Furthermore, a matched-pairs test of the distribution of errors (missed changing points vs. wrong and inserted changing points) was conducted and gave a P-value of 0.02 indicating that the distribution of errors is different as well.

The approach to tracking the reading position using only gaze information as presented in 2.1.4 performs poorly. The TER for this approach is 1.26. This was expected, as the reference is based on the words being read out loud and the readers will not always utter the words they look at.

The experiment has been conducted offline but the algorithms that track the reading and calculate word probabilities from gaze points (Sections 2.1.4 and 2.3.2) are causal and would perform in the same way in a live setting. Moving from an offline ASR setup to one where partial hypotheses would be used, however, would most probably result in performance degradations similar to those documented in [14].

5. Conclusions

In this paper we presented our work on fusing gaze information and speech for tracking the reading progress of people with dyslexia. The proposed fusion method achieved a 10.6% decrease in tracking error rate over the baseline ASR-only method.

In the course of doing the experiment we found that there is a multitude of parameters to tweak and equally many design options to consider – and since this was the first step in fusing eye-gaze information and speech recognition for tracking reading progress – we are confident that there'll be ways to fuse the two that improves the performance more significantly than what we have presented here.

6. Acknowledgements

The authors would like to thank master student Julia Alexandra Vigo for her help in collecting the data and Anders Olesen Sigh and Karina Fuhr Pedersen at VUC Nordjylland for facilitating the contact to the test participants.

7. References

[1] Pedersen, J. S., "User Centred Design Of a Multimodal Reading Training System for Dyslexics", Ph.D. thesis, Aalborg University, 2009.

[2] Duchateau, J., Kong, Y. O., Cleuren, L., Latacz, L., Roelens, J., Samir, A., Demuynck, K., Ghesquière, P., Verhelst, W. and hamme, H. V., "Developing a reading tutor: Design and evaluation of dedicated speech recognition and synthesis modules", Speech Communication, Volume 51, Issue 10, pp 985–994, 2009.

[3] Mostow, J., "Why and How Our Automated Reading Tutor Listens", International Symposium on Automatic Detection of Errors in Pronunciation Training (ISADEPT), KTH, Stockholm, Sweden, pp 43–52, 2012.

[4] Sibert, J. L. and Gokturk, M., "The Reading Assistant: Eye Gaze Triggered Auditory Prompting for Reading Remediation", Procedings of ACM Symposium on User Interface Software and Technology, pp 101–107, 2000.

[5] Rasmussen, M. H., Lindberg, B., Tan, Z.-H., "Combining Acoustic and Language Model Miscue Detection Methods for Adult Dyslexic Read Speech", International Speech Communication Association Special Interest Group, Workshop on Speech and Language Technology in Education, Venice, Italy, 2011.

[6] Rasmussen, M. H., Mostow, J., Tan, Z.-H., Lindberg, B., & Li, Y., "Evaluating Tracking Accuracy of an Automatic Reading Tutor", International Speech Communication Association Special Interest Group, Workshop on Speech and Language Technology in Education, Venice, Italy, 2011.

[7] Cooke, N. J., "Gaze-contingent automatic speech recognition", Ph.D. thesis, University of Birmingham, 2006.

[8] Nyström, M. and Holmqvist, K., "An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data", Behavior Research Methods, Volume 42, Issue 1, pp 188–204, 2010.

[9] User Manual, "Tobii Studio Version 3.2", Online: http://www.tobii.com/en/eye-tracking-research/global/library/manuals/, accessed on 8 April 2013.

[10] Olsson, P., "Real-time and offline filters for eye tracking", Msc. thesis, KTH Royal Institute of Technology, 2007.

[11] Hyrskykari, A., "Utilizing eye movements: Overcoming inaccuracy while tracking the focus of attention during reading", Computers in Human Behavior, Volume 22, Issue 4, pp 657–671, 2006.

[12] Koh, C.-W. E., "Speaker Diarization of News Broadcasts and Meeting Recordings", Msc. Thesis, Nanyang Technological University, 2009.

[13] Gillick, L., Cox, S. J., "Some statistical issues in the comparison of speech recognition algorithms", International Conference on Acoustics, Speech, and Signal Processing, pp 532–535, 1989.

[14] Li, Y., and Mostow, J., "Evaluating and improving real-time tracking of children's oral reading", Florida Artificial Intelligence Research Society Conference, pp 488–491, 2012.

Automatic Pronunciation Feedback for Phonemic Aspiration

Vaishali Patil, Preeti Rao

Department of Electrical Engineering, Indian Institute of Technology Bombay, India.

{vvpatil, prao}@ee.iitb.ac.in

Abstract

The computer-assisted learning of spoken language is closely tied to automatic speech recognition (ASR) technology which, as is well known, is challenging with non-native speech. By focusing on specific phonological differences between the target and source languages of non-native speakers, pronunciation assessment can be made more reliable. Aspiration, an important phonemic attribute in plosives of Indo-Aryan languages such as Hindi, Marathi and Gujarati, is rarely found in the world"s languages. The improper production of the aspiration contrast is thus often the most important cue to non-native accents of spoken Hindi. A system for the detection of phonemic aspiration in unvoiced and voiced stops based on discriminative acoustic features is shown to be effective for rating non-native accents and providing reliable phoneme-level feedback.

Index Terms: computer-assisted language learning, pronunciation scoring, non-native accent, phonemic aspiration

1. Introduction

The computer-assisted learning of spoken language is closely tied to automatic speech recognition (ASR) technology. The automatic assessment of a non-native learner based on carefully designed speaking tests coupled with focused phone-level feedback would go a long way into expanding the reach of the language education industry. While intelligibility is a prime requirement, the absence of non-native accents, as indicated by segmental (phone articulation) and suprasegmental (prosody) differences from native speech, are desirable. A key manifestation of foreign accent is the improper production of the target language (L2) phones. This is especially true when the phones in question do not belong to the phonology of the learner"s native language (L1).

The challenges of automation are linked to the known deficiencies of state-of-the-art ASR systems where phone recognition accuracies are relatively low and an acceptable performance in practical tasks is achieved only through the constraints of a powerful language model. In an application such as pronunciation assessment, however, language models would obscure genuine pronunciation errors by the non-native learner. Further, for better raw phone recognition accuracy, the acoustic models need to be trained on actual non-native speech. Such a speech database may not be easily available.

In recent past work, widely used direct measures of pronunciation quality from hidden Markov model (HMM) log-likelihoods in state-of-the-art mel-frequency cepstral coefficient (MFCC) feature based systems have not been found completely effective [1]. The MFCC features are a generic representation of the spectral envelope of the signal. More accurate judgement as well as meaningful feedback may be obtained via acoustic features that can be mapped to specific phonological attributes. In the present work, we investigate this approach to the automatic assessment of pronunciation of stop consonants of Hindi, belonging to the Indo-Aryan group, among the few language groups of the world where aspiration is a phonemic attribute. The improper production of the aspiration distinction is an important cue of non-native accent, in addition to vowel quality and intonation [2]. Limiting the scoring to the relevant aspects only would improve the reliability of the system by ignoring other natural variabilities of speech, and facilitate the use of specific discriminatory acoustic features for these aspects.

The 4-way contrast of Hindi stops where voicing and aspiration are independent for each place-of-articulation (PoA) are typically challenging for a learner from a different native language group. In the present study, we consider speakers of Tamil (as L1), a Dravidian language whose phonology is devoid of phonemic aspiration. Hindi is the native tongue of 200 million people in India and Tamil, that of 70 million. Hindi is the national language of India and together with English serves as a link language across the multilingual country. With widespread internal migration, the need for spoken language acquisition of the common languages is high.

The detection of phonetic differences involving aspiration has been previously attempted for Korean unvoiced stops, where a spectral tilt feature was added to vowel onset time (VOT), a traditionally used acoustic measure to distinguish aspirated stops from unaspirated (tense) [3]. While this achieved a good discrimination between aspirated and lax stops, it was less effective for the aspirated-tense case. Based on a phonological observation that aspiration is marked by breathy voice in the following vowel, Clements and Khatiwada [4] investigated the acoustic distinction between aspirated and unaspirated Nepali affricates on a small set of speakers to find that the acoustic measures of breathiness were not reliable across speakers. Voice quality features have been shown to enhance detection accuracies for phonemic aspiration in unvoiced Marathi stops [5], [6]. Voiced aspirated plosives, due to their rare occurrence in the world"s languages, have been studied minimally [7], [8]. Our recent work proposed and evaluated multiple acoustic features, extracted from the consonant and the following vowel regions, for the reliable detection of aspiration in word-initial voiced stops [9].

A goal of the present study is to develop and evaluate a speaker-independent automatic system for the robust detection of aspiration in Hindi voiced and unvoiced stops that can be used in a pronunciation scoring task for non-native speech. An objective measure of intelligibility is proposed based on maximum likelihood classification that is further validated by human listener ratings. The performance of the proposed system is compared with that of a baseline MFCC-HMM ASR system in the context of rating non-native pronunciation and providing corrective feedback.

2. Database and baseline system

Hindi and Tamil belong to distinct language groups that differ prominently in the plosive system. While both languages contain oral stops of 4 places of articulation, voicing and aspiration are used distinctively only in Hindi as depicted in

Table 1. Tamil does not distinguish aspiration or even voicing; stops are voiceless and weakly aspirated in initial position, and voiced after nasals [10]. Since our work is targeted towards a pronunciation assessment task, we collect data from native and non-native Hindi speakers in the form of read-out words containing the target phones in word-initial position across vowel contexts. For training the acoustic models, we use an already available database of Marathi spoken words by 20 native Marathi speakers. Marathi and Hindi are both Indo-Aryan languages and share the stop series of Table 1. In both, the unvoiced, aspirated labial is rarely used and therefore omitted.

Table 1. *IPA chart showing stops of Hindi and Tamil languages.*

Language	PoA of unvoiced and voiced stops			
	Labial	Dental	Retroflex	Velar
Hindi	p	t̪ t̪^{h}	ʈ ʈ^h	k k^h
	b b^h	d̪ d̪^{h}	ɖ ɖ^h	g g^h
Tamil	p (b)	t̪ (d̪)	ʈ (ɖ)	k (g)

2.1 Training and testing datasets

The training database comprises Marathi spoken words sampled at 16 kHz. Two distinct meaningful words with word-initial stops corresponding to each Hindi phone in Table 1 and each of the 8 vowels of the language (/ə/, /a/, /i/, /I/, /u/, /U/, /e/, and /o/) are formed and each word uttered in two carrier sentence contexts by 20 native speakers (equal male and female). The total number of words in each stop category appears in Table 2. The utterances were manually transcribed at phone level to use for acoustic model training.

Table 2. *Count of stop consonant-vowel (CV) pairs from train and test datasets.*

Stop category \ Data sets		Marathi train (20)	Hindi native (20)	Hindi non-native (10)
Unvoiced	Unaspirated	2560	1280	640
	Aspirated	1920	960	480
Voiced	Unaspirated	2560	1280	640
	Aspirated	2560	1280	640

For the pronunciation assessment evaluation, testing datasets were recorded by 20 native Hindi speakers and 10 speakers of Tamil L1. All were college-going adults. The non-native speakers had been exposed to Hindi reading and writing during their school years but had had limited exposure to the spoken language. They were fluent in Tamil, and used Hindi to varying extents as they currently lived outside their home state. The test dataset involved one meaningful word of Hindi corresponding to each consonant and vowel context embedded in 2 carrier phrases read out by each speaker. The speakers were presented the list of written words in Hindi script along with its English meaning. Table 3 shows examples of the words (that also happen to be minimal pairs) along with their typical pronunciations by native Hindi and Tamil speakers. Each dataset has an equal number of male and female speakers. A native Hindi listener was able to correctly identify every one of the speakers as native or not by listening to a small set of utterances (less than 20 words) by the speaker. It was observed by the listener that phonemic aspiration was the main discriminating attribute. Voicing was always realised correctly even though voicing is allophonic in Tamil stops.

Table 3. *Examples of stops in word initial as articulated by native and non-native speakers.*

Stop	Word	Meaning	Native pronunciation	Non-native pronunciation
t̪	ताली	Clap	t̪alI	t̪alI
t̪	थाली	Plate	t̪^{h}alI	t̪alI
b	बाग	Orchard	bag	bag
b^h	भाग	Section	b^{h}ag	bag

2.2 System frame work and baseline

Acoustic features computed from the segment of interest are used in a statistical classifier, previously trained on native speech (the Marathi database in this case), to derive a measure of correctness of pronunciation in terms of the aspiration attribute of oral stops across places of articulation. Since the test speech comprises of known utterances, an alignment with the word‟s phonetic transcription is first achieved using manner broad class models in an available state-of-the-art MFCC-HMM ASR system [11]. Such broad phonetic class based alignment can provide better robustness to speaker and language variability expected in the context of non-native speech since most confusions in a phone recognizer tend to be within the same manner class [12]. The broad classes are: vowels, sonorant consonants, unvoiced fricatives, unvoiced affricates, unvoiced stops, voiced affricates, voiced stops, silence and voice bar. The acoustic models are context independent, 3-state HMM with 8 mixtures, diagonal covariance and flat-start initialization. The standard 39 dim MFCC, delta and acceleration feature vector was computed at 10 ms intervals.

The aligned segments corresponding to voiced and unvoiced stops are processed for the extraction of acoustic-phonetic (AP) features as described in the next section. A Gaussian Mixture Model (GMM) classifier (6 mixtures, full covariance) is trained on the feature vectors of each class: unvoiced unaspirated (UU), unvoiced aspirated (UA), voiced unaspirated (VU), voiced aspirated (VA). Two-way classification is carried out on test unvoiced stop segments, and similarly on test voiced segments. For comparison, we also have a baseline system, implemented by extending the broad class MFCC-HMM system by separating the unvoiced and voiced stops further into 2 classes each to get UU, UA, VU and VA classes.

3. Acoustic-phonetic feature extraction

Aspiration is perceived as a release of breath following the stop burst. The aspiration feature has traditionally been associated with the timing of voicing onset [13]. Aspiration is also accompanied by an increased glottal opening in many languages including Hindi and the presence of aspiration noise during the following vowel [14]. Acoustic correlates of aspiration thus include VOT, aperiodicity of the vowel waveform and spectral shape attributes: H1-H2 (amplitude of the first harmonic relative to the second, reflecting the glottal open quotient) and spectral tilt. The latter two have been extensively studied as acoustic correlates of breathy voice quality in vowels where spectral tilt has been measured in various ways including H1-A3 and A1-A3 where An is the

highest amplitude in the region of the n^{th} formant [15], [16]. Thus phonemic aspiration is clearly multidimensional in terms of articulation, and trade-offs can be expected in both the production and perception of a specific realization. Therefore multiple acoustic features have been considered for reliable detection. The implementation of feature extraction is presented next.

3.1. Acoustic landmark detection

The extraction of the acoustic-phonetic features needs the precise temporal locations of landmarks corresponding to burst onset and vowel onset in the CV region of each utterance. The segmentation achieved by the broad-class HMM recognizer of Sec. 2.2 is coarse and must be refined as presented here.

The release burst onset is detected by the largest peak in the rate-of-rise (ROR) of the smoothened energy in 3500-8000 Hz within a 40 ms vicinity of the coarse boundary [17], [18]. This achieves burst localization to acceptable precision. However, cues to vowel onset are dependent on the nature of the consonant and especially difficult for aspirated and voiced stops. We employ different methods for vowel onset detection in the case of unvoiced stops and voiced stops. The rise of periodicity is a prominent cue to vowel onset after an unvoiced stop. Periodicity measured by the autocorrelation function peak computed from sliding 25 ms windows at 1 ms intervals, throughout a region of 40 ms around the initial boundary, is input to a previously trained decision tree to detect the vowel onset. The decision tree is trained on the manually labeled vowel onsets of the Marathi database. In the case of voiced stops, we use the rapid rise in the signal amplitude envelope in the low frequency band (50 Hz – 600 Hz) to detect the precise vowel onset in the vicinity of the initial coarse boundary [19]. While a median localization error of 5 ms is observed with respect to manually detected onsets, experimentally measured acoustic parameters extracted based on the automatically labeled onsets are seen to correspond well with those extracted using manual labels indicating the efficacy of the landmark detection methods.

3.2. Feature implementation

The features used in the pronunciation scoring task are from previously published (or to appear) work on acoustic-phonetic features for aspiration detection [5], [9]. VOT has been widely used to discriminate unvoiced aspirated stops from unaspirated stops in English where the former appear in word-initial context as allophones for voiced stops [13]. It was shown that including spectral tilt (A1-A3) and noise (signal-to-noise ratio – SNR) features improves the classification performance for unvoiced stops further [5]. In the case of voiced stops, performance with VOT alone is barely above chance. Including the A1-A3 and SNR improves it greatly. It was later demonstrated that a performance more comparable to that on unvoiced stops was obtained only after including all the further features listed in Table 4 for voiced stops [9].

Table 4. *Features used for aspiration detection in the AP-GMM system.*

Class of stops	Features in AP-GMM system
Unvoiced	VOT, H1-H2, A1-A3, SNR
Voiced	VOT, H1-H2, A1-A3, SNR, F1F3-sync, Low-band-slope, B3-band energy

The feature implementation available in [5], [9] is briefly reviewed here. VOT is the duration between burst onset and vowel onset. H1-H2 and spectral tilt measurements are obtained in the vowel region from magnitude spectra from 25 ms Hamming-windowed DFTs computed at 1 ms hop and averaged over a selected 5 ms duration. Low-band slope and B3-band energy provide further descriptions of spectral roll-off from the second and third formant regions respectively. The SNR provides the ratio of harmonic energy to aspiration noise energy. It is computed using a 25 ms analysis window placed at a selected time instant beyond the vowel onset. Signal power is obtained from the DFT spectrum but aspiration noise power is estimated using cepstral liftering [20]. Cepstral liftering separates the source from the vocal tract shaping and helps make the SNR less sensitive to formant influences. Since aspiration noise dominates the higher frequency region where formants are weak, an independent method to estimate the noise strength is to measure the uncorrelatedness of the signal components in two different frequency regions. "F1-F3 sync" is such a feature proposed by Ishi [16], computed using F1 and F3 bands of width 600 Hz around the automatically detected formant values corresponding to that token. The index represents correlation of the amplitude envelopes of the two band-pass filtered signals over a 25 ms region centered at a specific time instant beyond the vowel onset. The time-instants for spectral shape and noise features have been experimentally shown to be most discriminative at 13 ms and 23 ms respectively after the detected vowel onset [9].

4. Experiments and results

Table 5 shows the performances of the AP and MFCC features on 2-way classification (aspirated, unaspirated classes) of voiced and unvoiced stops. Table 4 lists the specific features used by the AP-GMM system. A 20-fold cross-validation (leave-one-speaker-out) experiment was carried out on the Marathi dataset. We observe that the MFCC features achieve an accuracy comparable to the AP features (the voiced stops performance is a bit lower) on the Marathi dataset. Next, both systems were trained on the full 20 speaker Marathi dataset, and tested on the 20 native speaker Hindi dataset. As seen in Table 5, the AP features show comparable performances on both language datasets whereas the MFCC features' performance decreases significantly. The AP features are clearly more robust to the cross-language transfer, as might be expected from their phonological basis.

Table 5. *Recognition accuracies on stops of Marathi and native Hindi datasets.*

Class	% accuracies in AP-GMM		% accuracies in MFCC-HMM	
	Marathi	Hindi native	Marathi	Hindi native
Unvoiced stops	90.5	90.2	90.3	76.4
Voiced stops	85.1	84.9	80.8	77.8

We next present an evaluation of the acoustic-phonetic system for pronunciation assessment and compare it with the baseline MFCC-HMM system on the same tasks. The tasks are designed to demonstrate the suitability of the systems for overall rating of the pronunciation quality of phonemic aspiration of a non-native learner and the accuracy of phone level feedback. The test database is as described in Sec. 2.1,

where each of the 20 native and 10 non-native speakers read out 240 words each embedded in a carrier phrase. The automatic systems are evaluated on this dataset for the (i) detection of non-native accent (with respect to ground-truth about the speaker‟s L1), (ii) rating of non-native accent via ranking of speakers (with respect to correct recognition as the intended target phone by native listeners), and (iii) accuracy of phone-level feedback with respect to human perception.

4.1 Detection of non-native accent

Each test word is automatically segmented and the classifier makes a two-way forced choice between unaspirated and aspirated classes for each test CV segment. For each speaker, we compute the percentage of instances that the target is correctly achieved (i.e. the classifier output matches the intended target phone) as an objective measure of speaker "intelligibility", separately for the unvoiced stops and voiced stops. Figures 1 and 2 show the obtained %correct for each speaker for the unvoiced and voiced stops respectively for each of the two different classifier systems. We see that the measured intelligibility varies across speakers with the non-native speakers‟ group doing worse overall. Observations of the individual scores of the 10 non-native speakers showed that their relative positions matched across the voiced and unvoiced stops, indicating that the phonemic aspiration contrast is acquired by Tamil-L1 learners similarly across both voicing classes.

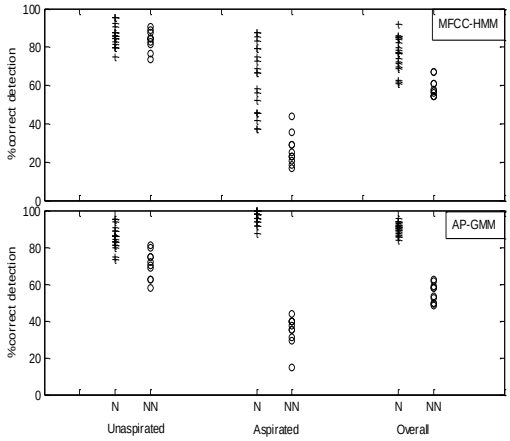


Figure 1: *Percentage correct achieved target unvoiced stops in native (N,+) and non-native (NN,o) datasets.*

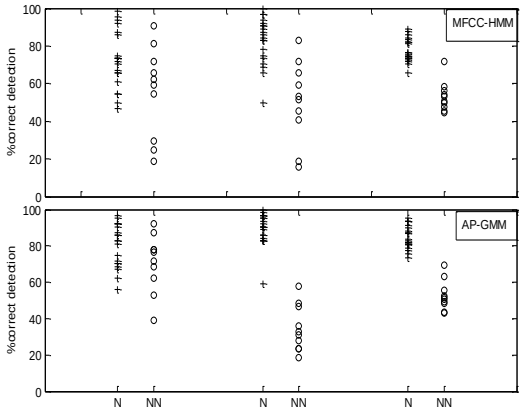


Figure 2: *Percentage correct achieved target voiced stops in native (N,+) and non-native (NN,o) datasets*

We note that the overall intelligibilities of the native (N) and non-native (NN) speakers are better separated by the AP system relative to separation achieved by the baseline MFCC system. While the non-native speakers show poor realization of aspirated targets, the AP system also indicates compromised unaspirated targets by the non-native speakers. This is not surprising in view of the allophonic usage of aspiration in Tamil word-initial stops, leading to the incorrect introduction of aspiration in the target Hindi word-initial unaspirated stops. For instance, „tall‟ of Table 3 was sometimes pronounced „t^hall‟.

4.2. Ranking non-native accent

In Figures 1 and 2 we observe an overlap in the overall intelligibility scores especially in the case of the MFCC system. That is, some native speakers are rated *lower* than the best ranked non-native speaker. We use this observation to choose a smaller set of speakers for the subjective validation of the ranking trends by human perception tests. The speakers used in the perception test data include the worst-rated native speaker from the baseline system, and 3 non-native speakers with various automatic intelligibility ratings, separately for the unvoiced and voiced stops.

Three judges, all fluent speakers of Marathi and Hindi, only one of whom is a trained speech scientist, labeled every voiced and unvoiced stop segment of each chosen speaker with one of 3 categories: unaspirated, aspirated, unsure. So as to not bias the judges, the *isolated* stop segments extracted from the word were presented for listening in random order. Each listener classified 512 and 448 segments each of the voiced and unvoiced stop CVs respectively, presented in randomized order over 4 sessions of approximate duration15 minutes each. The recognition task was chosen for the perception experiment rather than a "quality" evaluation in order to reduce subjectivity.

Table 5. *Percentage correct production of unvoiced stops as detected by listeners and by the automatic systems.*

Speaker ID	Perceptual results (%)			Classifier results (%)	
	Subject 1	Subject 2	Subject 3	MFCC-HMM	AP-GMM
N-8	94.6	93.8	93.8	60.7	89.3
NN-6	67.9	71.4	69.6	67.0	62.5
NN-1	58.9	58.0	59.8	54.5	58.9
NN-9	43.8	46.4	50.0	56.3	53.6

Table 6. *Percentage correct production of voiced stops as detected by listeners and by the automatic systems.*

Speaker ID	Perceptual results (%)			Classifier results (%)	
	Subject 1	Subject 2	Subject 3	MFCC-HMM	AP-GMM
N-20	92.2	89.8	89.8	65.6	82.0
NN-4	75.0	74.2	68.0	71.9	69.5
NN-9	57.0	57.0	50.8	53.9	51.6
NN-5	53.1	53.1	50.0	45.3	52.3

The results appear in Tables 5 and 6. A target is considered correctly achieved only if its perceived value (aspirated/unaspirated) matches the target. The maximum

number of instances rated "unsure" by any judge, were below 3% of the total targets for the native speaker, and less than 4% for the 3 non-native speakers. The unsure cases were ignored in the analysis of this section. Tables 5 and 6 show the %correct target achieved according to each of the judges as well as each automatic system, arranged in decreasing intelligibility as per the 3 judges (whose speaker rankings turned out to be identical). We observe that the rank ordering of the AP system matches the subjective ranking. This is not the case with the baseline system which ranks N-20 lower than NN-4.

4.3 Accuracy of phone-level feedback

A pronunciation assessment system that provides focused feedback in terms of flagging poorly articulated phones can be very useful in computer-aided language learning. In the classifier framework, the normalized likelihood of the target model, given the observation, provides a measure of the match between the test utterance and the native-trained model [21]. We use the log of ratio of likelihoods of the target and the opposite models as an estimate of the "goodness of pronunciation" of an uttered phone.

$$d(x) = \log\left(\frac{L(x|\wedge 1)}{L(x|\wedge 2)}\right) \qquad \ldots\ldots\ldots \quad (1)$$

where $L(x|\wedge 1)$ is the likelihood of an arbitrary point x in the feature space for model of class 1 (likewise $L(x|\wedge 2)$ for class 2). Class 1 represents the target class while class 2 the opposite class.

A ratio much greater than 1.0 would indicate native-like articulation of the target while a ratio much less than 1.0 would indicate wrong articulation. This is illustrated by Fig. 3 which shows the distribution of the log likelihood ratios over the native dataset for voiced stops for each of the AP and MFCC systems. As expected, the native utterances lie mostly to the right of the zero log-likelihood point. We choose a region around log likelihood ratio = 0 of width given by a fixed fraction (0.1) of the standard deviation of the native distribution to indicate "unsure" in the 3-way classification (correct/wrong/unsure) of the non-native utterances.

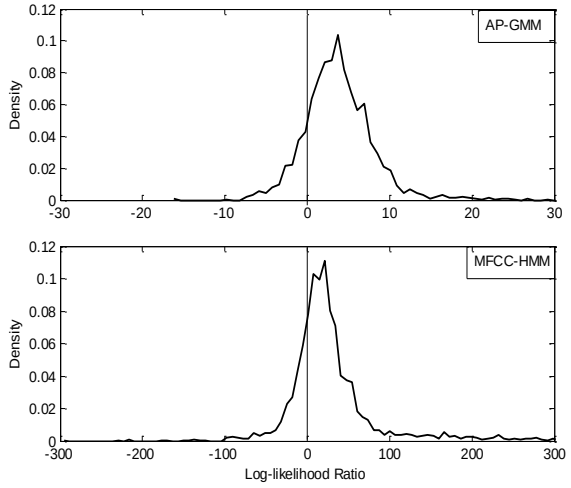


Figure 3: *Distribution of log-likelihood-ratio from AP-GMM and MFCC-HMM systems over native data set for voiced stops.*

The same speakers‟ data (extracted consonant segment only) ratings by the human judges, as already available from Task (ii) as presented in Sec. 4.2, were assigned the same numerical values viz. +1, 0, -1 for target correctly achieved, unsure and opposite achieved respectively. Next the correlation between corresponding ratings was computed between judges, and between each judge and each automatic system output. The inter-judge correlation is found to be 0.70.

Table 7. *Correlation between subjective and objective phone-level ratings for unvoiced and voiced stops.*

Class	Average correlation ratings	
	AP-GMM	MFCC-HMM
Unvoiced stops	0.53	0.39
Voiced stops	0.52	0.13

Table 7 shows the average of the correlation coefficients between each of the subject‟s ratings and the corresponding objective rating. We observe that the AP features provide phone-level feedback that is closer to subjective ratings when compared with that of the MFCC-HMM system which is especially poor for voiced stops.

5. Conclusion

The non-native accent that appears in a language learner‟s speech is, at the segmental level, related to the phonological differences between the speaker‟s L1 and the target language. Exploiting such relevant distinctions with suitable discriminating acoustic features can lead to reliable automatic assessment of degree of nativeness as well as detection of phoneme-level pronunciation errors. In this work, we presented the design and evaluation of a pronunciation scoring system for spoken Hindi where the learners‟ L1 is Tamil. The incorrect production of the aspiration distinction in voiced and unvoiced oral stops of Hindi is a prominent characteristic of non-native Indian speakers whose L1 does not belong to the Indo-Aryan language group. A statistical classifier using acoustic-phonetic features for aspiration detection was proposed based on the acoustic characteristics of voiced and unvoiced stops of Marathi. A number of methods are presented to evaluate the performance of the system in a pronunciation assessment context.

The AP features based system was shown to provide a measure of intelligibility that separates native and non-native speakers well. The acoustic-phonetic features outperformed a standard MFCC-HMM system on overall speaker intelligibility scoring as well as phoneme-level error detection. A discriminative classifier is expected to enhance the performance of the AP system further. Future work will extend the system to include other salient phonological attributes of spoken Hindi and larger scale evaluations on non-native speakers. A drawback of the AP features approach is that specific features are needed for specific phonological distinctions. Finding ways to select features automatically from suitable labeled training data would extend the scope of such work (similar to the suggestion by Strik et al [1]).

6. Acknowledgements

This work was supported in part by Bharti Centre for Communication at IIT Bombay.

7. References

[1] Strik, H., Truong, K., Wet, F. and Cucchiarini, C., "Comparing different approaches for automatic pronunciation error detection", Speech Communication, 51(10), pp. 845-852, Oct. 2009.

[2] Wiltshire, C. R. and Harnsberger, J. D. "The influence of Gujarati and Tamil L1s on Indian English: a preliminary study", World Englishes, 25(1), pp. 91-104, 2006.

[3] Cho, T., Jun, S., and Ladefoged, P., "Acoustic and aerodynamic correlates of Korean stops and fricatives", Journal of Phonetics, 30, pp. 193-228, 2002.

[4] Clements, G., N. and Khatiwada, R., "Phonetic realization of contrastively aspirated affricates in Nepali", In: Proc. ICPhS XVI 2007, Saarbrucken, Germany, pp. 629-632, Aug. 2007.

[5] Patil, V. and Rao, P., "Acoustic features for detection of aspirated stops", In: Proc. of National Conf. on Communication 2011, Bangalore, India, pp. 1-5, Jan. 2011.

[6] Patil, V. and Rao P., "Automatic pronunciation assessment for language learners with acoustic-phonetic features", In: Proc. of SLP-TED Workshop at COLING-2012, Mumbai, India, pp. 17-23, Dec. 2012.

[7] Rami, M. K., Kalinowski, J., Stuart, A. and Rastatter, M. P., "Voice onset times and burst frequencies of four velar stop consonants in Gujarati", J. Acoust. Soc. Am. 106(6), pp. 3736-3738, Dec. 1999.

[8] Miller A. L., "Guttural vowels and guttural co-articulation in Juhoansi", Journal of Phonetics, 35, pp. 56-84, 2007.

[9] Patil, V. and Rao P., "Acoustic features for detection of phonemic aspiration in voiced plosives", to appear In: Proc. of Interspeech 2013, Lyon, France, Aug. 2013.

[10] Balasubramanian, T., "Aspiration of voiceless stops in Tamil and English: an instrumental investigation", CIEFL Newsletter, pp. 14-18, 1975.

[11] Young S., et al., "The HTK Book v3.4", Cambridge University, 2006.

[12] Scanlon, P., Ellis, D. P. W. and Reilly, R. B., "Using broad phonetic group experts for improved speech recognition", IEEE Trans. Audio, Speech and Lang. Process, 15(3), pp. 803–812. Mar. 2007.

[13] Lisker, L. and Abramson, A., "Cross-language study of voicing in initial stops: Acoustical measurements", Word, 20(3), pp. 384-422, Dec. 1964.

[14] Ridouane, R., Clements, G., N. and Khatiwada, R., "Language-independent bases of distinctive features", Tones and features: Phonetic and Phonological Perspectives, pp. 264-291, 2011.

[15] Hanson, H. M., "Glottal characteristics of female speakers: Acoustic correlates", J. Acoust. Soc. Am., 101(1), pp. 466-481, Jan. 1997.

[16] Ishi, C. T., "A new acoustic measure for aspiration noise detection", In: Proc. ICSLP 2004, Jeju Island, Korea, pp. 629-632, Oct. 2004.

[17] Liu, S. A., "Landmark detection for distinctive feature-based speech recognition," J. Acoust. Soc. Am., 100(5), pp. 3417-3430, Nov. 1996.

[18] Patil, V., Joshi, S. and Rao, P., "Improving the robustness of phonetic segmentation to accent and style variation with a two-staged approach", In: Proc. of Interspeech 2009, Brighton, U.K., pp. 2543-2546, Sep. 2009.

[19] Prasanna, S. and Yegnanarayana, B., "Detection of vowel onset point events using excitation information", In: Proc. of Interspeech 2005, Lisbon, Portugal, pp. 1133-1136, Sep. 2005.

[20] Murphy, P. J., and Akande, O. O., "Noise estimation in voice signals using short-term cepstral analysis", J. Acoust. Soc. Am. 121(3), pp. 1679-1690, March 2007.

[21] Witt, S. and Young, S., "Language learning based on non-native speech recognition", In Proc. of Eurospeech 1997, Rhodes, Greece, pp. 633-636, Sep. 1997.

Pronunciation Assessment via a Comparison-based System

Ann Lee, James Glass

MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, Massachusetts 02139, USA
{annlee, glass}@mit.edu

Abstract

In this paper, we present preliminary results on applying a comparison-based framework to the task of pronunciation scoring. The comparison-based system works by aligning a student's utterance with a teacher's utterance via dynamic time warping (DTW). Features that describe the degree of mis-alignment are extracted from the aligned path and the distance matrix. We focus on a dataset in Levantine Arabic, a low-resource language for which there is not enough automatic speech recognition (ASR) capability available. Three different speech representations are investigated: MFCCs, Gaussian posteriorgrams, and English phoneme state posteriorgrams decoded on Levantine data. Experimental results show that the system can improve both correlation and mean squared error between machine predicted scores and human ratings compared to a template-based system.

Index Terms: pronunciation scoring, dynamic time warping, posteriorgrams

1. Introduction

The use of speech in computer-aided language learning (CALL) systems has enabled students to not only acquire vocabulary and grammatical concepts through reading but also practice pronunciation through speaking. More specifically, computer-assisted pronunciation training (CAPT) systems focus on the tasks of individual error detection and pronunciation assessment in nonnative speech [1], with the former aimed at detecting word or subword level pronunciation errors, and the latter targeted at scoring the overall fluency of an utterance. While these tasks can be further divided into processing read speech or spontaneous speech, their basic goal is the same, which is to compare a student's speech with that of a reference model.

In this paper, we focus on the task of pronunciation scoring on read speech. In early work, the reference models were stored as templates, and the student's speech was scored based on the percentage of the matching bits with that of templates [2, 3]. Later on, as automatic speech recognition (ASR) technologies improved, hidden Markov models (HMMs) were also applied to CAPT systems to model the reference speech statistically. Many of the fundamental features were based on HMM likelihood measures and posterior probability scores [4, 5, 6]. Timing scores such as phone segment duration, rate of speech and length of pauses, were also found to be highly correlated with human ratings [7, 8]. Some high-level features like recognition accuracy, confidence measures [9] and the ranking order of the correct phonemes [10] were also investigated. Another approach to model the reference speech was to build phonetic structures and use the distortion between two structures to estimate pronunciation proficiency [11, 12].

While ASR technology has its strengths, the process of building a recognizer requires a significant amount of annotated data and expertise. In addition, a new recognizer has to be built every time we want to build a CAPT system for a new target language. To address this issue, in our prior work [13], a comparison-based system was proposed for the task of mispronunciation detection in nonnative English. The system first aligns a student's utterance with a teacher's utterance via dynamic time warping (DTW). Features that describe the degree of mis-alignment are extracted from the aligned path and the distance matrix, and are then used for classifier training. The advantage of this framework is that it is language independent, and the speech representations that DTW compares can be obtained either in a fully unsupervised manner, such as Mel-frequency cepstral coefficients (MFCCs) or Gaussian posteriorgrams (GPs), or in a semi-supervised or fully supervised manner [14], such as phoneme posteriorgrams, depending on how much labeled data is available.

In this paper, we further explore this comparison-based framework in three aspects. First, we investigate the use of alignment-based features on the task of pronunciation scoring by training regressors instead of binary classifiers. Secondly, as there is no assumption about the target language for the framework, we turn our focus from nonnative English to Levantine Arabic, a low-resource language in which we do not have recognition capability. Lastly, besides MFCCs and GPs, we also explore using English phoneme state posteriorgrams decoded on Levantine data to examine the possibility of building a CAPT system for a low-resource language by taking advantage of a language with extensive resources.

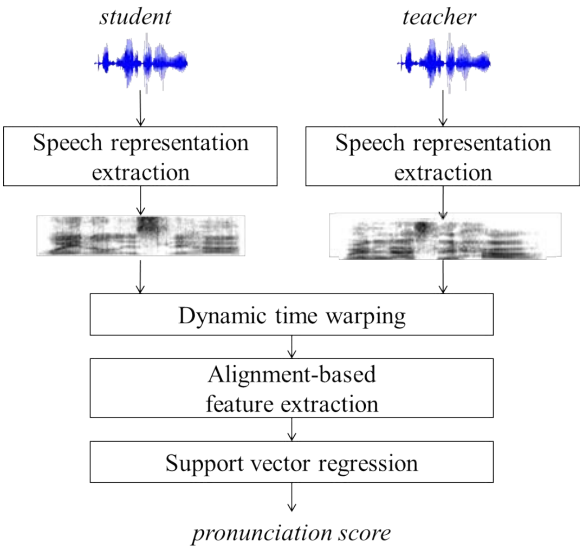


Figure 1: *System diagram. After transforming waveforms into speech representations, the system aligns the two utterances via DTW, and then extracts alignment-based features from the aligned path and the distance matrix. A support vector regressor is used for predicting an overall pronunciation score.*

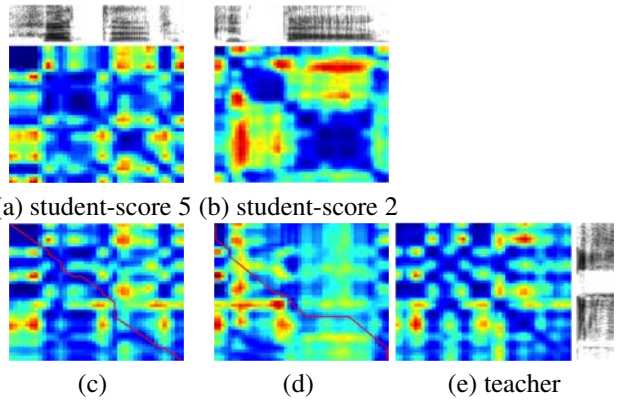
2. Corpus

The Levantine Arabic dataset consists of 21 nonnative speakers (students), including 11 males and 10 females, and 4 native speakers (teachers), including 2 males and 2 females. All students are native English speakers. Each speaker was asked to read the same 100 scripts, whose content varies from common phrases such as "Good morning" and "Thank you" to longer and more complicated sentences. Students listened to the reference audio first and then did the recording, and could repeat a recording until they were satisfied with the pronunciation. For every nonnative utterance, we have one score on a 1-5 scale for its intelligibility as decided by an expert. The scoring criterion was: 1 = many errors/unintelligible, 2 = heavy accent/difficult to understand, 3 = accented but mostly intelligible, 4 = slightly accented/intelligible, 5 = native accent/fully intelligible. There are no other human annotations on the data. After removing problematic recordings, we are left with 2064 nonnative utterances.

3. System Design

3.1. Dynamic time warping (DTW)

Fig. 1 illustrates the flowchart of the system. The first stage of the system aligns the student's utterance with a teacher's utterance through DTW. A DTW algorithm finds the optimal match between two sequences which may vary in speed. Given a teacher's utterance $T = (f_{t_1}, f_{t_2}, ..., f_{t_n})$ with n frames, and a student's utterance $S = (f_{s_1}, f_{s_2}, ..., f_{s_m})$ with m frames, an $n \times m$ distance matrix Φ_{ts} can be computed as $\Phi_{ts}(i, j) = D(f_{t_i}, f_{s_j})$, where $D(\cdot)$ denotes the distortion measure, or the distance, between two frames. DTW works by finding the path starting from $\Phi_{ts}(1, 1)$ and ending at $\Phi_{ts}(n, m)$ with the minimum accumulated distance.



(a) student-score 5 (b) student-score 2

(c) (d) (e) teacher

Figure 2: *(a) and (b) are the SSMs of two students' utterances with different scores, together with the spectrograms, and (e) is the SSM of a teacher saying the same sentence. (c) shows the alignment between (a) and the teacher, and (d) shows the alignment between (b) and the teacher. The red lines indicate the aligned paths.*

Note that the input to the DTW algorithm, i.e. f_t's or f_s's, can be of various speech representations, as long as an appropriate distortion measure can be defined. In early work, filter bank output or linear predictive features were often used [15]. More recently, posterior features have been successfully applied to facilitate not only speech recognition but also spoken keyword detection [16, 17]. The definition of a posteriorgram is as follows:

$$p_f = (P(v_1|f), P(v_2|f), ..., P(v_D|f)), \qquad (1)$$

where v_i's are the D possible models that the speech frame f might be originated from. For example, each v_i can be a single mixture in a D-component Gaussian mixture model (GMM), in which case p_f would be a Gaussian posteriorgram (GP), or each v_i can be a GMM for one single phoneme, in which case p_f would be a phoneme posteriorgram.

3.2. Alignment-based feature extraction

Fig. 2 illustrates two examples of alignments, one between a teacher's utterance and a student's utterance with a score of 5, and the other one between the same teacher's utterance and a student's utterance with a score of 2, as well as the self-similarity matrices (SSMs) of the three utterances and the corresponding spectrograms. An SSM can be obtained by aligning a sequence to itself, and thus it is symmetric on the diagonal.

We can see that a well pronounced utterance and a badly pronounced utterance have different characteristics in their alignment with the teacher. For example, for an utterance with a lower score, the aligned path would tend to be more off-diagonal, as there would be some high distortion regions along the diagonal. Also, its SSM would

be less similar to the SSM of the teacher's utterance. These observations are similar to what we had when analyzing the alignment between a reference word and a correctly pronounced word or a mispronounced word. Therefore, we can take advantage of the alignment-based features that we have designed previously. Table 1 provides an overview of each feature. More details can be found in [13].

All of the features can be extracted either on an utterance level or on a finer segmental level. In our system, we adopt an unsupervised phoneme segmentor to segment each reference utterance into smaller phoneme-like units [13]. Each distance matrix can be segmented into smaller blocks according to the segment boundaries and the aligned path. Features are extracted within each smaller unit, and we compute both the average and the standard deviation of each dimension across all the segments to form a single feature vector for an aligned pair, including the features extracted on the utterance-level.

After the alignment-based features are extracted, different regression approaches can be adopted for modeling the relationship between the features and the human ratings. In our system, we take advantage of a support vector regressor with an RBF kernel [18]. If there is more than one reference utterance for a script, we view pairs of teacher and student alignments as different instances during training, and take the average of the regressor's output for each pair during testing.

4. Experiments

4.1. Input speech representations

We explore the use of three different speech representations as inputs to our system. The first one is MFCC, for which the distance measure is defined as the Euclidean distance between two MFCC frames. The second representation is GP decoded from a 50-mixture GMM trained on all the native data (about 31 mins in total). The distance measure between two frames of GPs, p and q, can be defined as $-\log(p \cdot q)$ [16, 17].

The last representation is based on a monophone DBN-HMM English phoneme recognizer trained on the TIMIT training set to decode a set of English phoneme state posteriorgrams on the Levantine Arabic data. The DBN has 2 hidden layers (2048×2048) and a softmax layer of 183 units (3 states for each of the 61 phonemes), and takes 39-dimensional MFCCs stacked with 10 neighboring frames as input. As a result, each frame of the English phoneme state posteriorgrams is a 183-dimensional vector, and the distance measure can be also defined as the inner product distance.

Note that the first two speech representations can be obtained in a fully unsupervised manner. Though the last speech representation requires a carefully transcribed corpus in English, it does not require any phonetic labels in Levantine Arabic, a language with relatively few resources available.

Table 1: *The alignment-based features*

Aligned path & diagonal	
acc_path	accumulated distance along the aligned path
avg_path	*acc_path* normalized by path length
std_path	standard deviation of the distance along the aligned path
acc_diag	accumulated distance along the diagonal
avg_diag	*acc_diag* normalized by diagonal length
std_diag	standard deviation of the distance along the diagonal
diff_acc_p_d	*acc_path* − *acc_diag*
diff_avg_p_d	*avg_path* − *avg_diag*
ratio_avg_p_d	*avg_path* / *avg_diag*
max_seg_ratio	the length of the longest horizontal or vertical segment / path length
Distance matrix ($disMat$)	
avg_block	average distance within the block
std_block	standard deviation of the distance within the block
Duration	
dur_ratio	ratio between the length of the two sequences
diff_rel_dur	difference between the length of the two sequences that are normalized by the length of each full utterance
ratio_rel_dur	ratio between the length of the two sequences that are normalized by the length of each full utterance
Comparison with the reference	
diff_avg_block	*avg_block* − the average of the corresponding block in $SSM_{teacher}$
diff_avg_p_t	*avg_path* − the aligned path in the corresponding block in $SSM_{teacher}$
diff_avg_d_t	*avg_diag* − the aligned path in the corresponding block in $SSM_{teacher}$
diff_mat_t	element-wise difference between the rewarped $disMat$ and $SSM_{teacher}$
diff_s_t	element-wise difference between $SSM_{student}$ and $SSM_{teacher}$
hog_diff_mat_t	difference between the histograms of oriented gradients of the rewarped $disMat$ and $SSM_{teacher}$ [19, 20]
hog_diff_s_t	difference between the histograms of oriented gradients of $SSM_{student}$ and $SSM_{teacher}$

4.2. Experimental setup

We take advantage of the same English phoneme recognizer to first remove the silences at the beginning and the end of each utterance. Then, all waveforms are trans-

formed into 39-dimensional MFCCs every 10-ms, including first and second order derivatives, for the following GPs or phoneme state posteriorgrams decoding.

As there is no phonetic transcription for the data and thus we do not have recognition capability in Levantine Arabic, the baseline simulates a template-based system that scores an utterance based only on *acc_path*, *avg_path* and *std_path*. For evaluation, we run 100 iterations of 5-fold speaker-level cross validation using data from all 21 speakers. Only alignments between speakers with the same gender are considered. We compute both Pearson's correlation and the mean squared error (MSE) between the machine predicted scores and the human ratings.

4.3. Results

Experimental results are shown in Table 2. For all three speech representations, the comparison-based system obtains improvements relative to the template-based baseline in a range of 4.5% to 11.6% in correlation, and 3.8% to 15.9% in MSE. These results imply that the shape of the aligned path or the appearance of the distance matrix can provide more information about the quality of the pronunciation than alignment scores can do. These findings also agree with the findings we had in the task of mispronunciation detection. Using features extracted on the utterance level produces better results in both correlation and MSE than using features extracted on the phone level. A possible explanation is that aggregating the errors, i.e. the degree of mis-alignment, is better than averaging them. However, unlike our previous findings, there is no clear conclusion as to whether combining features from both levels can really achieve better performance.

Among the three speech representations, English phoneme state posteriorgrams gives the best result and also the largest improvement. This improvement most likely comes from the human supervision involved during English recognizer training for decoding posteriorgrams. The discriminative training process helps reduce mis-alignments from difference between speaker characteristics. Nevertheless, the high performance of the English phoneme state posteriorgrams suggests that high-resource language resources can be leveraged for training recognition on low resource languages in the context of a comparison-based approach. Because the alignment-based feature extraction process can be made independent from speech representation, a comparison-based approach can be feasibly integrated with the use of high-resource languages as training data.

4.4. Discussion

To further investigate how each type of alignment-based feature contributes to the task of pronunciation scoring, we focus on the English phoneme state posteriorgrams and repeat the 5-fold speaker-level cross validation by training on one single feature (extracted on both utterance-level and phone-level) at a time. Fig. 3 shows the correlation between the system output and human ratings for each feature.

Table 2: *Correlation and mean squared error between the machine predicted scores and the human ratings under different settings*

	MFCC	GP	English phoneme state posteriorgrams
Correlation			
Baseline	0.492	0.507	0.510
Utterance-level	0.526	0.536	0.559
Phone-level	0.511	0.526	0.534
Full system	0.523	0.535	**0.569**
Mean squared error			
Baseline	0.543	0.539	0.542
Utterance-level	0.513	0.509	0.491
Phone-level	0.519	0.516	0.508
Full system	0.522	0.507	**0.456**

First, note that the overall system performance is better than the results from using any single feature alone. This agrees with the results from several previous studies [6] which found that combining different scoring features can compensate for the weakness of each and produce a score that better correlates with human ratings.

Among the four different feature categories, the last one which compares the aligned path or the distance matrix with the self aligned path or the teacher's SSM obtains the best results on average. This could explain part of the reason why the comparison-based system can improve upon template-based approaches. Because the SSM from the teacher represents an optimal match, comparing it against the distance matrix can indicate proximity to a perfect match in a way that is different from template-based approaches relying only on alignment scores.

Moreover, a system based on *acc_path* or *acc_diag* performs better than a system based on *avg_path* or *avg_diag*. This again indicates that averaging or normalizing with respect to length may dampen the effect of high distortion regions. In line with previous work [7] indicating that utterance length is highly correlated with human ratings, the accumulated scores which have such information embedded also correlate better with human ratings. Although there is a chance that students may cheat the system by reading very quickly, there did not appear to be students circumventing the system in this way in our dataset.

5. Conclusion and Future Work

In this paper, we have explored the use of a comparison-based system in the task of pronunciation scoring. Experimental results have shown that, as in the task of mispronunciation detection, adopting alignment-based features that are extracted from the aligned path and the distance matrix can also improve system performance in predict-

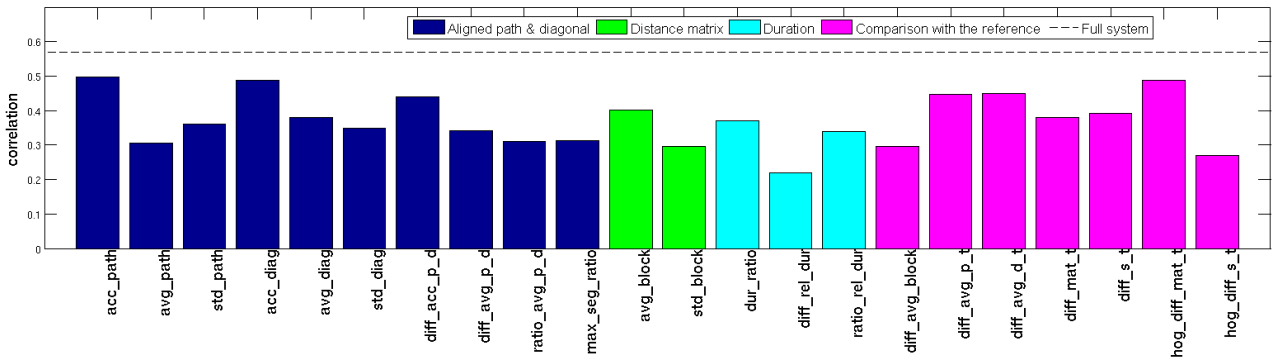


Figure 3: *Correlation between system output scores and human ratings based on a single feature*

ing pronunciation scores. The comparison-based system can be viewed as a combination of template-matching and classifier-based approaches. In fact, many of the alignment-based features are similar to ASR-based features that have been proved useful in pronunciation scoring. For example, comparing the structure of student and teacher SSMs is in some sense similar to comparing their phonetic structures [11, 12]. Features involving time comparisons might also reflect underlying durations of phoneme-like units.

Because the dataset we have collected is an initial attempt at gathering nonnative speech in a low-resource language, our current experiments are based on a relatively small dataset compared to that of previous work. As efforts continue to gather more data, we intend to examine system performance on larger-scale datasets, with the hope of enhanced performance due to greater amounts of training data. Running experiments on a dataset whose size is comparable with those in other studies can also allow us to have a fair comparison between absolute system performance. Future work should explore training the regressor from alignments in one language and testing on the other language to see whether misalignment patterns may be universal, and experimenting with speech representations that are more robust to different channel characteristics so that we can leverage more data from different sources.

6. Acknowledgements

The authors would like to thank Wade Shen for providing the dataset, and Ekapol Chuangsuwanich, Yu Zhang, Yaodong Zhang and Hung-An Chang for their help with the DBN-HMM recognizer.

7. References

[1] Eskenazi, M., "An overview of spoken language technology for education", in Speech Communication, 2009.

[2] Kewley-Port, D., Watson, C., Maki D. and Reed D., "Acoustic-articulatory inversion", in proc. ICASSP, 1987.

[3] Wohlert, H., "Voice input/output speech technologies for German language learning", in Die Unterrichtspraxis/Teaching German, 1984.

[4] Witt, S. M. and Young, S. J., "Phone-level pronunciation scoring and assessment for interactive language learning", in Speech Communication, 2000.

[5] Neumeyer, L., Franco, H., Digalakis, V. and Weintraub, M., "Automatic scoring of pronunciation quality", in Speech Communication, 2000.

[6] Franco, H., Neumeyer, L., Digalakis, V. and Romen, O., "Combination of machine scores for automatic grading of pronunciation quality", in Speech Communication, 2000.

[7] Cucchiarini, C., Strik, H. and Boves, L., "Automatic evaluation of Dutch pronunciation by using speech recognition technology", in proc. ASRU, 1997.

[8] Bernstein, J., De Jong, J., Pisoni, D. and Townshend, B., "Two experiments on automatic scoring of spoken language proficiency", in proc. Integrating Speech Technology in Learning, 2000.

[9] Cincarek, T., Gruhn, R., Hacker, C., Noth, E. and Nakamura, S., "Automatic pronunciation scoring of words and sentences independent from the non-native's first language", in Computer Speech and Language, 2009.

[10] Chen, J.-C., Jang, J.-S., Li, J.-Y. and Wu, M.-C., "Automatic pronunciation assessment for Mandarin Chinese", in proc. ICME, 2004.

[11] Minematsu, N., "Pronunciation assessment based upon the phonological distortions observed in language learners' utterances", in proc. ICSLP, 2004.

[12] Suzuki, M. Dean, L., Minematsu, N. and Hirose, K., "Improved structure-based automatic estimation of pronunciation proficiency", in proc. SLaTE, 2009.

[13] Lee, A. and Glass, J., "A comparison-based approach to mispronunciation detection", in proc. SLT, 2012.

[14] Lee, A., Zhang, Y. and Glass, J., "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams", in proc. ICASSP, 2013.

[15] Sakoe, H. and Chiba, S., "Dynamic programming algorithm optimization for spoken word recognition", in IEEE Trans. on Acoustics, Speech and Signal Processing, 1978.

[16] Hazen, T. J., Shen, W. and White, C., "Query-by-example spoken term detection using phonetic posteriorgram templates', in proc. ASRU, 2009.

[17] Zhang, Y. and Glass. J. R., "Unsupervised spoken keyword spotting via segmental DTW on Gaussian posteriorgrams", in proc. ASRU, 2009.

[18] Chang, C.-C. and LIN, C.-J., "LIBSVM: A library for support vector machines", in ACM Transactions on Intelligent Systems and Technology, 2011. Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.

[19] Dalal, N. and Triggs, B., "Histograms of oriented gradients for human detection", in CVPR, 2005.

[20] Muscariello, A., Gravier, G. and Bimbot, F., "Towards robust word discovery by self-similarity matrix comparison", in proc. ICASSP, 2011.

Predicting Gradation of L2 English Mispronunciations using Crowdsourced Ratings and Phonological Rules

Hao Wang, Xiaojun Qian and Helen Meng

Department of Systems Engineering and Engineering Management,
The Chinese University of Hong Kong, Hong Kong SAR of China
{hwang,xjqian,hmmeng}@se.cuhk.edu.hk

Abstract

Pedagogically, CAPT systems can be improved by giving effective feedback based on the severity of pronunciation errors. We obtained perceptual gradation of L2 English mispronunciations through crowdsourcing, and conducted quality control utilizing the WorkerRank algorithm to refine the collected results and reach a reliable consensus on the ratings of word mispronunciations. This paper presents our work on modeling the relationship between the phonetic mispronunciations and the actual word ratings. Based on phonological rules representing phonetic mispronunciation productions, we propose two approaches to predict the gradation of word mispronunciations. Reasonable correlation and agreement are found between the human-labeled and machine-predicted gradations for both approaches, which imply that the use of phonological rules in word-level mispronunciation gradation prediction is promising.

Index Terms: CAPT, crowdsourcing, mispronunciation gradation

1. Introduction

The success of computer-assisted pronunciation training (CAPT) technology is increasing due to the fact that CAPT systems can benefit learners by offering extra learning time and material, individualized feedback and the possibility of self-paced practice in a private and stress-free environment [1]. Furthermore, we observe increasing research interest in the pedagogical effectiveness of CAPT in recent years.

A key issue in CAPT concerns the generation of corrective feedback for L2 learning. Methodologists suggest teachers focus their attention on a few error types rather than try to address all the errors [2]. This can help learners discriminate errors by priority. Another reason is that if too many mispronunciations are presented at the same time, learners may get confused, be discouraged or even lose self-confidence, especially for beginner-level learners. One criterion for selecting errors is perceptual relevance – listeners may tolerate a few "*subtle*" mispronunciations because they do not affect intelligibility greatly; but perceptually "*serious*" errors which hamper communication must be indicated and corrected promptly. Hence, a CAPT system can be pedagogically improved by providing effective feedback through prioritizing detected errors in order of their severity. We believe that while variations exist across individual listeners, there is a general consensus in the perceptual gradation of pronunciation errors ranging from *subtle* to *serious*. Therefore, we are motivated to collect data on the severity of mispronunciations in L2 English speech and attempt to develop an automatic means of predicting the gradation of mispronunciations.

We used crowdsourcing to collect perceptual gradations of word-level mispronunciations and conducted quality control using the WorkerRank algorithm [3] to filter the crowdsourced data in terms of reliability. In this paper, we propose two approaches to predicting the gradation of word mispronunciations based on crowdsourced reliable data. The rest of this paper is organized as follows: Section 2 presents some related previous work. Section 3 reviews our previous effort on the collection of perceptual gradations of word-level mispronunciations and the procedure of quality control for selecting reliable data. Section 4 introduces our proposed approaches to predicting mispronunciation gradation. Experimental results are exhibited in Section 5, together with the discussion about the results. Section 6 presents the conclusions and future work.

2. Related Work

Our work collects human perceptual ratings of L2 English speech to develop some predictive model that can mimic human ratings according to the severity of word-level mispronunciations. Related previous work includes:

Kim at al. [4] made use of an acoustic model and generated probabilistic scores for specific phone segments based on a speech recognition system developed to help American adults learn the French language. A panel of five teachers of French were asked to rate the pronunciation of selected phone segments on a scale of 1 (unintelligible) to 5 (native-like). These collected ratings were mainly used for performance evaluation, but not for training for predictive scoring (as is done in our work).

In Neri et al. [5], a subset of speech material of low overall pronunciation quality was selected for annotators to label what they considered to be the most serious phonetic errors. The annotations were used for statistical analysis and to draw up a list of suggested priority of specified phonetic errors to be addressed. The work does not perform automatic predictions for prioritizing errors.

The measures of pronunciation quality in both the above studies were collected from human expert labelers. In recent years, crowdsourcing has become a popular technique widely used for data collection and labeling. Crowdsourcing is a process of obtaining needed services, ideas or content by soliciting contributions from an undefined large group of people. Amazon Mechanical Turk (AMT) [1] is one of the best known crowdsourcing platforms. It provides a convenient mechanism

[1] https://www.mturk.com/mturk/welcome

for distributing human intelligence tasks (HITs) via the web to an anonymous crowd of non-expert workers who complete them in exchange for micropayments [6]. Compared with traditional methods for data collection and labeling, crowdsourcing is considerably more efficient, cost-effective and diversified.

Kunath and Weinberger [7] collected English speech accent ratings from native English listeners on the AMT platform. AMT Workers were asked to rate accentedness of the given non-native speech on a five-point Likert scale (ranging from '1' for native accent to '5' for heavy, nonnative accent). This work mentioned about a research direction in using the collected data set to train an automatic speech accent evaluation system. However it only described the data collection procedure, and did not give information about how to train an automatic system.

Peabody [8] used AMT to collect word-level judgments of pronunciation quality for each utterance in the corpus. Each utterance was assigned to three Workers, who were asked to provide binary judgments for each word on whether it was mispronounced. The pronunciation quality of each word was classified based on the number of Workers who marked it as mispronounced (0 as good, 1-2 as ugly, 3 as mispronounced). These data were further used for mispronunciation detection.

Both efforts above used crowdsourcing techniques and considered that all the collected data were reliably labeled. Our current work proposes the WorkerRank algorithm [3] to assess the quality of crowdsourced data and we only preserve data of high quality in developing a model for predicting the severity of word-level mispronunciations.

3. Crowdsourced Mispronunciation Gradations

In our previous work [3], we used the AMT crowdsourcing platform to collect perceptual gradation of word-level mispronunciations in non-native English speech. This section presents a brief description of our crowdsourcing procedure, together with new corpus-specific data.

3.1. L2 corpus

The corpus we use is the Cantonese subset of the Chinese University Chinese Learners of English (CU-CHLOE) Corpus, which contains speech recordings by 100 Cantonese speakers (50 male and 50 female) reading several types of carefully designed material, as shown in Table 1.

Table 1. *Types of prompted speech in the CU-CHLOE English corpus.*

Group	# of prompts	Example
Confusable words	10	debt doubt dubious
Phonemic sentences	20	These ships take cars across the river.
The Aesop's Fable	6	The North Wind and the sun were…
Minimal pairs	50	look full pull foot book

The material is designed by experienced English teachers, aiming to cover representative examples of mispronunciations from Cantonese learners of English. Each of the 100 speakers reads 86 prompts that contain 436 unique words

3.2. Possible gradation of errors

We defined four grades of mispronunciations in terms of the severity, as follows:

1. No mispronunciation: As good as native pronunciation.
2. Minor/Subtle: Minor deviation in word pronunciation with the native pronunciation. Can accept the deviation even if it is not rectified in the learner's speech.
3. Medium: Noticeable deviation in word pronunciation with the native pronunciation. Would prefer that the deviation be rectified for better perceived proficiency of the learner's speech.
4. Major/Salient: Very noticeable deviation in word pronunciation with the native pronunciation, to the level that it is distracting and/or affecting communication with and understanding by the listener. Strongly advise that the deviation be rectified with high priority for improved proficiency of the learner's speech.

3.3. Overview of crowdsourcing procedure

We created 200 distinct HITs (see Figure 1), each of which contains a bunch of L2 English utterances for the AMT Workers to rate according to the gradation criteria described in Section 3.2. Each distinct HIT was assigned to 3 workers.

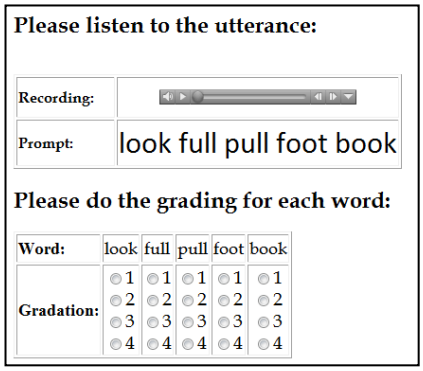


Figure 1: *An example of an utterance in an HIT.*

We ultimately obtained 600 sets of ratings (200 distinct HITs × 3 assignments) from 287 Workers.

3.4. Reliable ratings

We conducted quality control on the crowdsourced data by identifying and selecting *reliable Workers* and adopting their ratings based on an assumption that *reliable Workers* will always provide *reliable ratings*. The methodology we use for ranking the reliability of Workers is described as follows:

3.4.1. Graph-based representation for Workers

We represent the relations among Workers as an undirected weighted graph (see Figure 2) where a node is an individual Worker, a connection line between two Workers indicates that these two Workers completed common HITs and the weight for each connection is Cohen's weighted kappa [9] value, which

aims to measure the degree of agreement between two Workers on an ordinal scale.

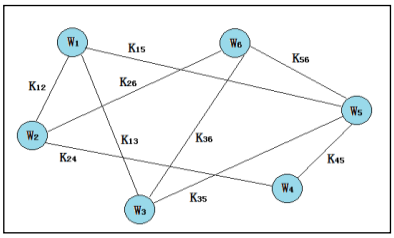


Figure 2: *A simple example of an undirected weighted graph representing AMT Workers and their relations.*

3.4.2. WorkerRank

We designed WorkerRank algorithm [3] to filter AMT Workers in terms of reliability. This algorithm is adapted from the well-known PageRank algorithm [10] that ranks web pages. We consider that a Worker is reliable if he/she gives ratings that are mostly consistent with other reliable Workers. The WorkerRank is defined in Equation 3:

$$\boldsymbol{W}(w_i) = \frac{1-d}{N} + d \cdot \left[\sum_{j:\{i,j\}\in E} \frac{k_{ij}}{\sum_{m:\{j,m\}\in E} k_{jm}} \boldsymbol{W}(w_j) \right], i = 1, \ldots, N, \quad (3)$$

where $\boldsymbol{W}$ is the resulting WorkerRank score vector, whose i-th component is the WorkerRank score associated to Worker w_i, N is the number of distinct Workers, d is the damping factor which controls the relative importance of the two involved terms, is the Cohen's weighted kappa value between Worker and Worker w_j.

Equation 3 is a recursive expression, thus we perform iterative calculation until convergence is reached, to obtain a list of individual AMT Workers sorted by their WorkerRank according to reliability. We wish to include ratings covering the entire vocabulary of the corpus based on the most reliable Workers. Therefore, we rank all Workers in descending order according to reliability. We start by including the ratings from the top-ranking Worker, and then proceed to include the ratings from the next best Worker, and continue this procedure until all the words in the corpus are covered. We ultimately include the top 190 Workers (with the damping factor $d = 0.99$) as reliable ones, which is the minimum set of Workers that provide ratings covering all the utterances.

According to the assumption presented at the beginning of Section 3, all the ratings from reliable Workers are regarded as reliable ratings. We derived reliable ratings for the whole Cantonese subset of the corpus which has 156,709 reliable ratings. The distribution across 4 possible grades (see Section 3.2) is shown in Table 4.

Table 4. *Distribution of reliable ratings for each grade, based on Cantonese subset of CU-CHLOE corpus.*

Grade	Count	Percentage
1	109,226	69.70%
2	26,762	17.08%
3	12,109	7.73%
4	8,612	5.49%
TOTAL	156,709	100.00%

4. Predicting Mispronunciation Gradations

4.1. Baseline prediction based on table lookup for mispronunciation transcriptions

The entire corpus is phonetically transcribed by trained linguists. Based on the transcriptions, we conduct our first trial of predicting the gradations of word-level mispronunciations using the following approach: For each transcription of a word, we aggregate all the reliable ratings of the articulated words carrying the same transcription. Then, we take the average of the aggregated values and treat it as the gradation score of the corresponding transcription of that word. An example is shown in Table 5.

Table 5. *An example of how a transcription of a word mapped to crowdsourced ratings.*

Word	Transcription	Ratings from reliable Workers	Average of aggregated scores
rate	r ey t	1,1,1	1
rate	r iy t	3,3,4	3.4
rate	r iy t	4,3	
rater	r iy t	4,4	4
rater	r iy t ax	4,3,4	3.67

To predict the gradation of a given word mispronunciation, we adopt the mapped average score as its predicted rating, e.g., for an articulated word "rate" with the transcription "r iy t", we look it up in the obtained rated transcription list (See Table 5), and map it to the value 3.4 which is assigned as its predicted gradation score.

The above approach is straightforward but has an obvious limitation that the prediction can only take effect for those pronunciations (transcriptions) of words that have been observed. To solve the limitation, we attempt to use phonological rules to make the prediction have a more general coverage of mispronunciations.

4.2. Approaches based on phonological rules

The processes of phonetic mispronunciation productions are usually modeled by phonological rules. We assume that the phonological rules present in a word mispronunciation have a strong impact on the gradation of the word, so that associating each phonological rule with a certain score can help derive the gradation of word-level mispronunciations. In this section, we propose two prediction approaches by modeling the relationship between phonological rules and the crowdsourced reliable word ratings (see Section 3.4): one heuristic is to equate the word gradation with the score of the most salient phonological rule (one with the maximum score) in a word; the other one is based on linear regression – the word gradation yields from a linear combination of all the scores of rules found in a word mispronunciation

As described in [11], phonetic mispronunciation productions can be represented as context-dependent phonological rules of the form:

$$\alpha \rightarrow \beta / \sigma \_ \lambda,$$

which denotes that phone α is substituted by the phone β, when it is preceded by the phone σ and followed by the phone λ. The insertion rule can be represented by replacing α with null symbol

0 while the deletion rule is to replace β with null symbol 0. For σ and λ, they can be replaced with symbol # as a word boundary.

As mentioned previously, all speech data of the corpus are phonetically labeled by trained linguists; and the canonical pronunciations of all words can be readily obtained from electronic dictionaries (e.g., TIMIT, CMUDict, etc.). By aligning the canonical pronunciations with manual transcriptions of the corpus using phonetically-sensitive alignment [12], context-dependent phonological rules can be generated for all phonetic mispronunciations in the corpus. These derived rules are used to predict word-level mispronunciation gradation by the following two approaches.

4.2.1. Maximum gradation score

For each phonological rule, we aggregate the derived reliable ratings (see Section 3.4.2) of word mispronunciations that include this phonological rule; then we simply take the average of the aggregated values and treat it as the gradation score of the rule. An example is illustrated in Tables 6a and 6b.

Table 6. *An example of how a phonological rule is mapped to the crowdsourced ratings.*

(a). *word-to-rules mapping.*

Word	Phonological rules	Rating
rate	ey → iy / r _ t	3,4,4
rater	ey → iy / r _ t er → ax / t _ #	4,4

(b). *rule-to-ratings mapping.*

Phonological rule	Rating set	Average
ey → iy / r _ t	3,4,4,4,4	3.8
er → ax / t _ #	4,4	4

Using the rated phonological rules derived above, we can predict the gradation of a word mispronunciation by following the principle that the most serious error (phonological rule with the highest gradation score) dominates the gradation of the word mispronunciation. Therefore, we predict mispronunciation gradation according to the steps below:

1. get the transcription of a word mispronunciation;
2. derive a set of phonological rules of this word mispronunciation;
3. map each of the derived rules to a gradation score by referring to the rated rule list obtained previously;
4. assign the gradation score of the most serious error in a word to this word as its predicted mispronunciation gradation.

An example of the above steps is given as follows:

1. we get a mispronunciation "ae ch ih ng" of word "aching";
2. the derived phonological rules are "ey → ae / # _ k" and "k → ch / ey _ ih";
3. rule "ey → ae / # _ k" is associated with a score of 3.26, rule "k → ch / ey _ ih" is associated with a score 3.57 by referring to the rated rule list;
4. the gradation of this word mispronunciation is assigned as 3.57 which is the higher gradation score of the two rules derived previously.

4.2.2. Linear regression

Another approach is to model the gradation of a word mispronunciation as a linear combination of the gradation scores of the corresponding phonological rules that the word mispronunciation includes. This relationship can be expressed as:

$$G_w = \sum_r \big(G_r \cdot \delta(r)\big) + b, \tag{4}$$

where G_w is the gradation of an uttered word mispronunciation w; G is the gradation score of the rule r; $\delta(r)$ is an indicator function, i.e. $\delta(r) = 1$ if r occurs in w, and $\delta(r) = 0$, otherwise; b is the offset term. The summation is taking over all in the system.

Multiple word mispronunciation gradations can be expressed in a matrix form as follows:

$$\boldsymbol{w} = \boldsymbol{A}\boldsymbol{r} + b\boldsymbol{e}, \tag{5}$$

where $\boldsymbol{w}$ is a vector containing the gradation score of each uttered word, which is calculated by averaging the crowdsourced "reliable" ratings of that word; $\boldsymbol{A}$ is a matrix with binary elements A_{ij} indicating whether the phonological rule j occurs in the uttered word i; $\boldsymbol{r}$ is a vector that contains the gradation scores of each rule; $\boldsymbol{e}$ is the all-one vector.

We run least-square linear regression analysis. A rule score vector $\boldsymbol{r}$ and an offset term b are obtained, and are used for predicting word mispronunciation gradation by Equation 4, e.g. for the mispronunciation "s ae l ax n t" of word "salient", we derive two phonological rules: "ey → ae / s _ l" and "iy → 0 / l _ ax"; the corresponding gradation scores of these two rules obtained from the previous regression analysis are 0.74 and 1.27; thus, with the trained offset term $b = 1.64$, the gradation of this mispronunciation is the summation of the above three scores, which is 3.65.

5. Experiments

5.1. Procedure

The experiments are carried out using the Cantonese subset of CU-CHLOE corpus. We split the corpus by speakers into disjoint training (25 male and 25 female) and test (25 male and 25 female) sets. 2,347 distinct context-dependent phonological rules are generated, which fully cover all phonetic mispronunciations in the training set. The gradation scores of all generated rules are trained on rated word mispronunciations using each of the two approaches as described in Section 4.2. The rules that are generated from the training set of the corpus may not cover all the mispronunciations in the test set. Thus, during prediction, we simply skip those mispronunciations that include untrained rules.

We calculate correlation and Cohen's weighted kappa between human-labeled gradations (i.e. the average of crowdsourced "reliable" ratings for each uttered word) and machine-predicted gradations by each prediction approach for the test set. To calculate kappa values, we first quantify all the word gradation scores (by rounding) to 4 integer values {1,2,3,4} which represent 4 possible grades of mispronunciations (see Section 3.2); some (less than 2% of total number of word mispronunciations) of the gradation scores obtained from linear regression approach exceed the range from 1 to 4; we quantify those gradation scores to their nearest grade values (1 or 4). For

the purpose of comparison, we include the baseline approach (See Section 4.1) in the following Table.

Table 7. *Evaluation results for different prediction approaches.*

2,347 rules	**Baseline**	**Maximum score**	**Linear regression**
# of tested words	15934	15934	15934
# of predicted words	13766	14736	14736
% of predictions	86.39%	92.48%	92.48%
Correlation (*r*-value)* **95% CI**	0.627 (0.617, 0.637)	0.644 (0.635, 0.653)	0.644 (0.635, 0.653)
Kappa	0.561	0.550	0.588

5.2. Discussion

From Table 7 we see that all correlation values are above 0.6 and all kappa values exceed 0.5, which reflects a reasonable consistency between human-labeled and machine-predicted gradations. If we compare all the prediction approaches, the two approaches based on phonological rules outweighs the baseline approach in almost all evaluation measures in Table 7, and the linear regression approach has the best performance. Table 7 also illustrates that prediction based on phonological rules have a better coverage of mispronunciations than the baseline approach based on table lookup for transcriptions.

The presented approaches to predicting the gradation of word-level mispronunciations are based on the detailed phonetic transcriptions of L2 speech labeled by trained linguists. This guarantees that the phonetic mispronunciations of L2 speech are identified accurately. However, in a practical system, it is not easy to obtain accurate (manual) phonetic transcriptions of L2 utterances immediately. In that case, an acoustic model can help obtain possible transcriptions (with acoustic scores) automatically, though usually with the trade-off of lower accuracy.

Conclusions and Future Work

Giving effective feedback based on the severity of mispronunciations is of core pedagogical importance in a CAPT system. We used crowdsourcing to collect the perceptual gradation of word-level L2 English mispronunciation, and conducted quality control with the WorkerRank algorithm on the crowdsourced data to derive reliable gradations. In this paper, we propose two approaches to predict gradation of word mispronunciations using the derived reliable gradations and phonological rules. Reasonable correlation and agreement found in the experimental results shows that our proposed approaches using phonological rules for predicting word mispronunciation gradation is promising.

In future work, we will try other regression analysis to seek a better model for prediction. Directly optimizing the correlation or kappa on the training set is also an interesting direction to pursue. Besides, we also plan to use acoustic model to assist scoring the mispronunciations which cannot be covered by the phonological rules derived from the training set or whose transcriptions are not immediately available.

* p-value associated with each of the three correlation values is less than 0.0001.

Acknowledgements

The work is partially supported by the grant from the Hong Kong SAR Government's Research Grants Council General Research Fund (Project No. 415511).

References

[1] Neri, A., Cucchiarini, C. and Strik, H., "ASR corrective feedback on pronunciation: Does it really work?", in Proc. of Interspeech, 1982-1985, 2006.

[2] Ellis, R, "Corrective Feedback and Teacher Development", L2 Journal, 1: 3-18, 2009.

[3] Wang, H. and Meng, H., "Deriving Perceptual Gradation of L2 English Mispronunciations using Crowdsourcing and the WorkerRank Algorithm", in Proc. of the 15th Oriental COCOSDA, Macau, China, 9-12 Decembeer 2012.

[4] Kim, Y., Franco, H. and Neumeyer, L., "Automatic pronunciation scoring of specific phone segments for language instruction", In Fifth European Conference on Speech Communication and Technology, 1997.

[5] Neri, A, Cucchiarini, C, and Strik, H, "Segmental errors in Dutch as a second language: how to establish priorities for CAPT", in Proc. of InSTIL/ICALL Symposium, 2004.

[6] McGraw, I., Glass, J., Seneff, S., "Growing a Spoken Language Interface on Amazon Mechanical Turk", in Proc. of Interspeech2011, Florence, 2011.

[7] Kunath, S. A. and Weinberger, S. H., "The wisdom of the crowd's ear: speech accent rating and annotation with Amazon Mechanical Turk", in Proc. of CSLDAMT '10, Association for Computational Linguistics, 2010.

[8] Peabody, M. A., "Methods for pronunciation assessment in computer aided language learning", [dissertation], US -- MA: Massachusetts Institute of Technology, 2011.

[9] Shoukri, M. M., "Measures of interobserver agreement", 2004.

[10] Page, L., Brin, S., Motwani, R., and Winograd, T., "The pagerank citation ranking: Bringing order to the web", Technical report, Stanford Digital Library Technologies Project, 1998.

[11] Lo, W., Zhang, S. and Meng, H., "Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System," in Proc. of Interspeech, Makuhari, Japan, 26-30 September 2010.

[12] Harrison, A. M., Lo, W. K., Qian, X. J. and Meng, H., "Implementation of an Extended Recognition Network for Mispronunciation Detection and Diagnosis in Computer-Assisted Pronunciation Training," in Proc. of the 2nd ISCA Workshop on Speech and Language Technology in Education, Warrickshire, 2009.

Determining Sentence Pronunciation Difficulty for Non-native Speakers

Jeesoo Bang and Gary Geunbae Lee

Department of Computer Science and Engineering
Pohang University of Science and Technology, South Korea
{jisus19, gblee}@postech.ac.kr

Abstract

This paper investigates the features that determine the sentence pronunciation difficulty for Korean speakers of English. We selected three types of features: length, word frequency, and phonemes that Korean speakers generally replace with other phonemes. We used support vector machines and a multiple linear regression model to determine the pronunciation difficulty of given sentences, and measured the results with a five-fold cross validation. We demonstrated that these features could determine sentence pronunciation difficulty with an accuracy and a correlation coefficient sufficient for computer-assisted pronunciation training (CAPT) systems. The combination of all three feature types had the highest accuracy and correlation coefficient in determining sentence pronunciation difficulty. For single features, the length-based feature type was the most accurate in determining sentence pronunciation difficulty. The phoneme-specific feature type also had high accuracy. Length, phoneme, and word features can be used to guide the automatic choice of sentences for CAPT systems that depend on users' proficiency levels.

Index Terms: sentence level decision, pronunciation level, pronunciation difficulty feature, CAPT sentence level

1. Introduction

Computer-assisted pronunciation training (CAPT) is beneficial for learning a foreign language because it provides a private, stress-free environment. Most CAPT systems include an automatic speech recognition component, and can therefore provide individual feedback to users. Pronunciation is one of the most difficult skills to acquire for adults learning foreign languages [1], [2]. For that reason, numerous studies have considered pronunciation training systems, and assessment of students' pronunciation is an important problem in a CAPT system.

Many studies have focused on automatic evaluation of students' proficiency improvement [3], [4]. The CAPT users' proficiency level must be considered when presenting sentences to them. If the presented sentence is too easy, the CAPT user would feel bored, and if the sentence is too difficult, the user may become discouraged. Therefore, the sentences should be classified into pronunciation difficulty levels. The CAPT system cannot use sound information when classifying sentences' pronunciation levels, because the sentences cannot be uttered beforehand. Thus, we explore the automatic determination of sentence pronunciation difficulty not using sound information.

The goal of this project is to provide appropriate sentences according to an individual's proficiency level. To that end, sentences should be classified into several pronunciation difficulty levels without sound information. In this paper, the pronunciation difficulty level of a sentence is classified using several features extracted from the sentence.

This paper is organized as follows. Section 2 introduces the sentence corpus used in our study. Section 3 describes the features and methods used in this study. Section 4 describes the evaluation of the features. Section 5 discusses the results and Section 6 provides a conclusion.

2. Data

We collected data for the experiment because we could not find an appropriate corpus with pronunciation difficulty scores for each sentence. For this reason, we built a new corpus, named the pronunciation difficulty test (PDT) corpus. We selected 240 sentences, and 20 experiment participants recorded and rated each sentence.

2.1. Data collection

The 240 sentences used in the experiments were obtained from various English text sources: English textbooks from Korean elementary, middle, high schools; the reading part of certified English tests, such as TOEIC and TOEFL; and English newspapers. We removed sentences containing digits, family names, or special characters because these sentences could confuse the experiment participants, whether or not they are pronouncing the sentences correct.

We recruited 20 native Korean university students to rate the sentences. The students' English proficiency levels were intermediate to advanced; most students had scored approximately 800 on the TOEIC. The students were asked to

Table 1: *Example sentences from the PDT corpus with difficulty levels and average difficulty score*

Difficulty Level	Average Difficulty	Sentence
Easy	1	Can you help me, please?
	1.4	The most important thing to me now is my friends.
Medium	2.05	The difficulty is that all orders must be delivered within a week.
	2.3	Won't that make for some interesting marketing campaigns?
Difficult	3.05	But political analysts said the strategy remains dicey.
	3.9	My life has been synched to a chemotherapy calendar ever since my leukemia diagnosis last year.

record the 240 sentences and rate each sentence from 1 (easy pronunciation) to 5 (difficult pronunciation) (Table 1). The students could look through the sentences before reading them aloud and recording them. Also, the students re-recorded a sentence if they stammered.

2.2. Data analysis

The PDT corpus has 240 sentences and 2,791 words. The mean and standard deviation of word counts in the sentences are 11.6 and 6.4, respectively. The students' ratings of pronunciation difficulty varied from 1 to 3.9 with a mean of 1.98 and a standard deviation of 0.72.

We rated the 4,800 utterances with a pronunciation error detection and feedback system [5] to see how much the students' ratings and their speaking ability agree. The system is an English pronunciation simulation and phoneme error detection system for non-native speakers. The scoring accuracy [6] of the system is 82.4% for mispronunciation detection. The system generates a score for a given utterance, based on the mispronounced phoneme counts over all the phonemes in the utterance.

We calculated the correlation of the average rating score for each sentence and the average system score for each sentence. The correlation coefficient was 0.90, which is quite high. This correlation coefficient indicates that the subjective rating score of pronunciation difficulty is relevant to the objective pronunciation score, which means that if a student felt that he or she pronounced a certain sentence poorly, the probability is high that the student actually mispronounced it.

We also calculated the Pearson's correlation coefficients among the students' ratings for inter-annotator agreement. We calculated the correlation coefficient for each rating pair, resulting 190 pairs for 20 students, and then took average for the 190 correlation coefficients. The correlation coefficient among the students' rating scores was 0.53.

2.3. Data preparation

We partitioned the PDT corpus into three classes to solve the determining sentence difficulty problem by classification. The partitioning was performed using the average student rating. Because the ratings had a continuous distribution, we partitioned the PDT corpus almost equally. We labeled the three classes Easy, Medium, and Difficult according to the ratings. The Easy, Medium, and Difficult classes were assigned 74, 87, and 79 sentences, respectively.

3. Experimental design

Three types of features were extracted to determine sentence pronunciation difficulty. There were length-based features, phoneme-specific features, and word identity features. Additionally, two machine learning methods, support vector machines (SVMs) and multiple linear regression (MR), were used to determine the pronunciation difficulty.

3.1. Features

Only text-based features were extracted from the PDT corpus because we wanted to automatically provide sentences that are appropriate for CAPT users, and the sentences could not have acoustic features beforehand. For feature extraction, we utilized the sentence text and its corresponding phoneme sequence, which was generated using the CMU pronouncing dictionary[1].

3.1.1. Length-based features

Sentence length is a common feature in readability measurement tasks [7]-[9]. Although readability and pronunciation difficulty are different concepts, reading and reading aloud, that is, pronunciation, can be considered similar tasks. We used seven length-based features: the number of words, characters, and phonemes, the maximum and average number of characters per word, and the maximum and average number of phonemes per word.

3.1.2. Phoneme-specific features

To effectively capture pronunciation difficulty, we extracted several phoneme-specific features from the phoneme sequences of the PDT corpus. Because Korean was the mother tongue of experiment participants, we extracted Korean-specific phoneme features.

Table 2: *List of possible phoneme substitutions, which Korean speakers commonly mispronounce*

Consonant			Vowel		
/tʃ/	→	/t/	/ɪ/	→	/i/
/ð/	→	/d/	/ɔɪ/	→	/i/
/θ/	→	/t/	/ɝ/	→	/ɹ/
/θ/	→	/s/	/ʊ/	→	/oʊ/
/ʒ/	→	/dʒ/	/ɛ/	→	/æ/
/f/	→	/p/	/ɑ/	→	/ɔ/
/ɹ/	→	/l/	/ɔ/	→	/oʊ/
/v/	→	/b/	/ʌ/	→	/ɑ/

Korean speakers commonly mispronounce particular phonemes. The most common are eight consonant replacements and eight vowel replacements (Table 2) [10]. The eight consonant substitutions have seven unique phonemes because /θ/ can be mispronounced as /t/ or /s/. These 15 unique phonemes were counted for each sentence.

We also counted complex word-syllable coda counts. The coda comprises the consonant sounds of a syllable that follows the nucleus, which is usually a vowel. The Korean language has four simple syllable types: V, VC, CV, and CVC; consonant clusters are not allowed in Korean [11]. Because Koreans are not familiar with complex codas, we counted the complex codas. We counted codas for each sentence if the number of codas in a word exceeded two because words with two codas, such as *land* and *lunch*, are common and easy to pronounce but words with three or four codas, such as *fifths* and *lengths*, are not.

The third phoneme-specific feature was weighted difficulty

[1] The CMU pronouncing dictionary is currently available at http://www.speech.cs.cmu.edu/cgi-bin/cmudict, and version 0.7a is used in this work

Table 3: *Results of classification and regression on the PDT corpus with all possible combinations of the feature types*

Features	SVMs				MR
	Accuracy (%)	Precision (%)	Recall (%)	F1-measure (%)	Correlation coefficient
All	85.4	85.3	85.4	85.3	0.90
Phoneme +word	82.5	82.3	82.5	82.2	0.90
Length + word	82.1	81.9	82.1	82.0	0.90
Length + phoneme	80.0	79.8	80.0	79.9	0.88
Length	79.6	79.3	79.6	79.4	0.89
Phoneme	73.3	73.0	73.3	73.1	0.86
Word	56.3	55.4	56.3	55.5	0.63

[12]. We assigned a weight to each feature and summed the weights for each sentence. The weights are as follows: commonly mispronounced phonemes – 1, two codas – 1, three codas – 3, and four codas – 5.

We calculated the fraction of consonant counts over vowel counts (cov counts) in a word. The cov counts are similar to coda counts, but the cov counts can capture abstract syllable types. Because the Korean language does not have consonant clusters, Korean speakers of English could have trouble pronouncing words with a consonant count much greater than the vowel count, indicating that they need to practice this type of English. We calculated the maximum and average number of cov counts per word in each sentence. Also, we extracted the maximum and average number of syllables per word in each sentence.

We used seven phoneme-specific features in total: the number of erroneous phonemes, the number of complex coda counts that exceeds two, the weighted sum of commonly mispronounced phonemes and complex codas, the maximum and average number of cov counts, and the maximum and average number of syllables per word.

3.1.3. Word identity features

Unigrams of words have been used as a feature to calculate reading difficulty [13]. We considered that infrequent words may indicate the pronunciation difficulty of English sentences.

The British national corpus (BNC) was used to calculate the word frequencies for each word of the PDT corpus. The BNC is considered one of the most reliable corpus resources available, with more than 100 million words. The BNC reflects present day English usage for speech and publications in the UK [14].

A leading corpus lexicographer, organized lemmatized and unlemmatized frequency word lists of the BNC. We used the unlemmatized frequency word list "all.num.o5", which can be downloaded from Kilgarriff's website[1]. The list contains items that occurred more than five times in the BNC corpus (including both spoken and written material). The file contains 208,656 different words and each entry is ranked by frequency.

We refined the PDT corpus to use the BNC frequency list. We changed the sentences in the PDT corpus from American spellings to British spellings to maximize the effectiveness of the frequency list. We extracted the word frequency ranking for each word in the sentences in the PDT corpus. We calculated the maximum word frequency ranking for each sentence because the ranking decreases as the frequency of word decreases in the BNC. Also, we calculated the average word frequency ranking for each sentence to determine the number infrequent words in the sentence.

We set the frequency rankings to vary from 1 to 5 to have a reasonable feature range because both the maximum and the average frequency rankings varied significantly. The maximum frequency rankings had a mean of 28,997 and a standard deviation of 51,445, and 75% of the PDT corpus sentences had frequencies less than the mean. The average frequency rankings also showed the same phenomenon. We divided the sentences with frequencies less than the mean into three levels, 1, 2, and 3, and divided the remaining 25% of the sentences into two levels, 4 and 5. We used two word identity features: the maximum and average frequency ranking for each sentence.

3.2. Models

The purpose of determining the pronunciation difficulty is to automatically provide CAPT users with appropriate sentences according to their pronunciation levels, and a CAPT system can use levels or scores to evaluate the readers. If a CAPT system has a user with a certain proficiency level, the system provides sentences according to the user's level. If another CAPT system evaluates the user with a scoring method during training, the system can provide sentences according to the user's scoring. To ensure that the CAPT system made these adjustments in sentence difficulty, we used both classification and regression methods to determine the pronunciation difficulty of the sentences.

We used SVMs [15] to classify the sentences into pronunciation difficulty levels. SVMs were chosen as classifiers due to their superior classification performance in many machine learning tasks. For the regression method, we used MR.

4. Experiment and Result

We used SVMs and MR to determine the pronunciation difficulties of the PDT corpus. All possible combinations of the three feature types were examined. We performed a five-fold cross-validation using Weka 3.7 [16], which provides a collection of machine learning algorithms.

The final results are presented as accuracy, precision, recall, and F1-score for the SVMs, and a Pearson's product-moment correlation coefficient for MR (Table 3). In Table 3, the feature type names are abbreviated as *length* (length-based feature), *phoneme* (phoneme-specific feature), and *word* (word identity

[1] http://www.kilgarriff.co.uk/bnc-readme.html

feature). We used total 16 features for the experiment: seven length-based features, seven phoneme-specific features, and two word identity features (Section 3.1). The correlation of MR was higher than the accuracy of SVMs because we divided the PDT corpus into three classes that did not have explicit gaps among the average ratings. The best performing set of features was the combination of all three feature types for both of SVMs and MR.

The length-based features and phoneme-specific features classified better than the word identity features. The students rated sentences as difficult to pronounce if the sentences had many words, long words, or phonemes that are difficult for Korean speakers. In contrast, the word difficulty alone did not closely correspond to the pronunciation difficulty, which is different than reading difficulty [13].

For the classification problem, F1-measures of the three classes can be calculated because the classification was done using SVMs, which used multiple binary SVMs to classify multiple classes. The PDT corpus was divided into three classes (Section 2.3). The F1-measures for the Easy, Medium, and Difficult classes were 0.92, 0.79, and 0.85, respectively. The F1-measures of the Easy and Difficult classes were higher than that of the Medium class because the Easy and Difficult classes have only one boundary with the Medium class, whereas the Medium class has two boundaries. The confusion matrix, which Weka provides, shows that the Easy and Difficult classes were each classified into two classes: the Easy and Medium or the Difficult and Medium; but the Medium class was classified into all three classes: the Easy, Medium and Difficult (Table 4).

Table 4: *Confusion matrix of the SVM classifier using all features*

Original	Classified as		
	Easy	Medium	Difficult
Easy	71	3	0
Medium	9	67	11
Difficult	0	12	67

5. Discussion

The accuracy of the classification result was 85.42%, and the correlation coefficient of the regression result was 0.90. The classification accuracy and the regression correlation had similar results for various feature sets.

We determined the pronunciation difficulty of the PDT corpus with three types of features. The classification accuracy and the regression correlation increased with increase in the number of features combined. The combination of all three feature types had the highest accuracy and correlation, the combinations of two feature types had accuracy and correlation that were higher than the single-feature type sets.

For single-feature type sets, the length-based feature type had higher accuracy than the other two feature types. With only the length feature, SVMs acquired approximately 80% accuracy, and MR had a correlation coefficient of 0.89, which is almost the same as the value when all features were used. Sentence length had a significant effect on determining the pronunciation difficulty. The phoneme feature alone had a 73.3% classification accuracy, which is below the length-based feature. This result indicates that the phoneme-specific feature alone is not sufficient to determine the pronunciation difficulty of sentences. Although the word identity feature type alone was not sufficient to determine the pronunciation difficulty (56.3% and 0.63), the word feature had an effect on the overall accuracy. Without the word feature, the accuracy and correlation coefficient declined.

A classification accuracy of 85.4% and a regression correlation coefficient of 0.90 are sufficient to determine the pronunciation difficulty of sentences to be provided to CAPT users. In the classification, the classifier did not confuse any of the sentences in the Easy class with the Difficult class or vice versa. CAPT users in the Easy level would only be provided with sentences from the Easy level and a small number from the Medium level, but none from the Difficult level. Medium level CAPT users can be provided ten sentences containing eight Medium level sentences, one Easy level sentence, and one Difficult level sentence. However, the inclusion of a few sentences from unsuitable levels is not harmful to CAPT users because CAPT, as a pronunciation training system, is intended for learning, not for evaluation.

6. Conclusions

The purpose of this paper was to use text information to determine the sentence pronunciation difficulty for non-native speakers of English. The problem was solved using classification and regression based on length-based, phoneme-specific, and word identity features. The highest performance was achieved with the combination of all three feature types. For single feature types, the length-based feature had the highest accuracy and correlation coefficient, followed by the phoneme-specific feature. We demonstrated that determining sentence pronunciation difficulty with these features could be used in CAPT systems with satisfactory accuracy (85.4%) and correlation coefficient (0.90).

7. Acknowledgements

This work was supported by the Industrial Strategic technology development program 10035252, Development of dialog-based spontaneous speech interface technology on mobile platform funded By the Ministry of Trade, industry & Energy(MI, Korea). This work was supported by the MKE (The Ministry of Knowledge Economy), Korea and Microsoft Research, under IT/SW Creative research program supervised by the NIPA (National IT Industry Promotion Agency) (NIPA-2012-H0503-1201-1002).

8. References

[1] T. Scovel, *A time to speak: A psycholinguistic inquiry into the critical period for human speech*: Newbury House Cambridge^ eMA MA, 1988.

[2] J. Setter and J. Jenkins, "Teaching pronunciation: A state of the art review'," *Language Teaching,* vol. 17, pp. 1-17, 2005.

[3] A. Neri, C. Cucchiarini, and H. Strik, "The effectiveness of computer-based speech corrective feedback for improving segmental quality in L2 Dutch," *ReCALL,* vol. 20, pp. 225-243, 2008.

[4] R. Hincks, "Speech technologies for pronunciation feedback and evaluation," *ReCALL,* vol. 15, pp. 3-20, 2003.

[5] J. Lee, J. Bang, M. Chung, S. Kim, and G. G. Lee, "A Pronunciation Variants Prediction Method for ASR-based Mispronunciation Detection," *Computer Speech and Language,* submitted.

[6] S. Kanters, C. Cucchiarini, and H. Strik, "The Goodness of Pronunciation algorithm: a detailed performance study," *Proceedings of SLATE,* 2009.
[7] X. Liu, W. B. Croft, P. Oh, and D. Hart, "Automatic recognition of reading levels from user queries," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 2004, pp. 548-549.
[8] E. Dale and J. S. Chall, "A formula for predicting readability," *Educational research bulletin,* pp. 11-28, 1948.
[9] G. R. Klare, *The measurement of readability*: Iowa State University Press, 1964.
[10] D.-H. Ahn and M. Chung, "One-pass semi-dynamic network decoding using a subnetwork caching model for large vocabulary continuous speech recongnition," *IEICE TRANSACTIONS on Information and Systems,* vol. 87, pp. 1164-1174, 2004.
[11] J.H. Kim, *Unowa moonwha [Language and Culture]*: Seoul, Korea: Youkrack Publishing, 2001.
[12] K. Zechner, D. Higgins, R. Lawless, Y. Futagi, S. Ohls, and G. Ivanov, "Adapting the Acoustic Model of a Speech Recognizer for Varied Proficiency Non-Native Spontaneous Speech Using Read Speech with Language-Specific Pronunciation Difficulty," *Proceedings of INTERSPEECH 2009,* pp. 604-607, 2009.
[13] K. Collins-Thompson and J. Callan, "A language modeling approach to predicting reading difficulty," in *Proceedings of HLT/NAACL*, 2004.
[14] G. N. Leech, P. Rayson, and A. Wilson, *Word frequencies in written and spoken English*: Longman Harlow,, UK, 2001.
[15] V. Vapnik, *The nature of statistical learning theory*: springer, 1999.
[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *ACM SIGKDD Explorations Newsletter,* vol. 11, pp. 10-18, 2009.

Automated Content Scoring of Spoken Responses Containing Multiple Parts with Factual Information

Wenting Xiong[1], *Keelan Evanini*[2], *Klaus Zechner*[2], *Lei Chen*[2]

[1]Department of Computer Science, University of Pittsburgh, Pittsburgh, PA, USA

[2]Educational Testing Service, Princeton, NJ, USA

wex12@cs.pitt.edu,{kevanini,kzechner,lchen}@ets.org

Abstract

This paper presents approaches to automated content scoring of spoken language test responses from non-native speakers of English which contain multiple parts addressing factual information that the test taker has previously heard via auditory stimulus materials. While previous work relating to content scoring of spontaneous, unpredictable speech has focused only on entire responses and on general topic matching approaches, such as content vector analysis, the specific nature of spoken responses in our data requires response segmentation and extraction of features that indicate the relevance and correctness of the facts contained in the different parts of the response. Our best content features, based on similarity with key facts and concepts, achieve correlations of $r = 0.615$ (for speech recognition output) and $r = 0.637$ (using human transcriptions) with expert human rater scores. Furthermore, we show that these content features outperform traditional vector space based features. Finally, we demonstrate that the performance of a scoring model based on a combination of features developed previously and some of the newly designed content features improves significantly from $r = 0.624$ to $r = 0.664$ on an unseen evaluation set when using speech recognition output.

Index Terms: spoken language assessment, automated scoring, content appropriateness

1. Introduction

The research reported in this paper falls into the domain of automated scoring of non-native speech, and focuses on the scoring of the content of spoken responses in a test of English for non-native speakers. Determining the content accuracy of highly predictable speech (e.g. reading text aloud) is straightforward, in that one can use the output hypothesis of an automatic speech recognition (ASR) system as a proxy for what the speaker said, and then compute the edit distance to the stimulus passage or sentence to obtain an estimate for content correctness [1, 2]. Given that ASR systems for such highly predictable speech perform very well even for non-native speech (word error rates are generally below 10% [3]), these estimates of content accuracy are fairly reliable.

The situation is quite different, however, when processing spontaneous, highly unpredictable speech from non-native speakers. Typically, word error rates can range from 20-40% for this type of input. Furthermore, no particular text reference is available that can be used to compare the speaker's response to. For these reasons, most approaches to evaluating content correctness for spontaneous responses in spoken language tests have relied on methods originally proposed in the field of Information Retrieval, such as Content Vector Analysis (CVA) or Latent Semantic Analysis (LSA) [4]. In these approaches, the words in a spoken response are treated as an unordered list ("bag-of-words"), and comparisons between previously obtained training vectors and a spoken test response are made, e.g., by using the cosine similarity between weighted word vectors.

The test items used for this study are spontaneous in nature; however, the test taker is required to mention a set of specific facts in the response. Accordingly, generic approaches which regard one complete response as a single unit, such as CVA, may not be suitable, since responses for the test items in this study typically consist of multiple units (or text spans), each of which addresses a specific concept (content element) requested by the test item.

We therefore implement and evaluate several approaches that take into account the specific nature of these test items and their responses, in particular their localized content elements for specified concepts, to be able to assess the content accuracy more reliably and more specifically. Our aproach broadly consists of two main steps: (a) automated segmentation of the spoken test response into units of coherent content; and (b) computation of a set of content features based on scoring each text unit first separately and then comprehensively according to a set of content facts provided by test developers and further annotated by experts.

The remainder of this paper is organized as follows: Section 2 presents related work in the area of automated content scoring; Section 3 describes the data we use for this study; Section 4 provides details about our approach and the methods used in this research; Section 5 describes the experiments and evaluations we conducted; Section 6 discusses our findings; finally, Section 7 concludes the paper and provides an outlook into future work.

2. Related Work

There have been many previous studies about measuring the relevance and accuracy of content in the domains of automated assessment of essays and short textual responses. These efforts can be grouped into the following two sets. The first group relies on extracting patterns associated with the correct answers from the responses and matching them with pre-defined scoring rules [5, 6]. For example, the c-rater system described in [5] parses test-takers' responses and uses a pattern-matching algorithm to match the parsed constituents with manually written rules. The degree of match is used as a measure of content accuracy. To cope with the demand of manually generating these pre-defined patterns, [6] developed a bootstrapping method to generate patterns from a set of keywords and synonyms. Later, [7] compared several machine learning methods, e.g., decision trees and Bayesian learning, to the pattern matching method and

reported that the machine learning methods provide encouraging results.

The second type of method used for content scoring relies on a variety of text similarity measurements to compare a response with model responses [8]. Compared to the first group, such methods can bypass the labor intensive pattern-building step. A widely used approach to measuring text similarity between two text strings is to convert each text string into a vector of word counts and then use the angle between these two vectors as a similarity metric. For example, CVA has been successfully utilized to detect off-topic essays [9] and to provide content-related features for essay scoring [10]. For this group of methods, how best to measure the semantic similarity between two terms is a key question. A number of metrics have been proposed, including metrics derived from WordNet [11], a semantic ontology [12], and metrics related to the co-occurrence of terms in corpora or on the Web [13].

Recently, other novel NLP methods have been applied to the task of content scoring. For example, methods from the related NLP task of textual entailment [14], which attempts to find directional inference relations between two strings of text, have been applied to content scoring. [15] combined several graph alignment features with lexical semantic similarity measures using machine learning techniques and showed that answers can be more accurately scored in this way than by using semantic measures alone.

Compared to the research on content scoring for written text, there is only a small amount of research on scoring tests of spoken language based on content accuracy. In one example, [4] investigated using CVA, Pairwise Mutual Information (PMI), and Latent Semantic Analysis (LSA) to score spontaneous speech responses. In addition, [16] investigated the use of different semantic similarity measures, including PMI_IR from web queries, to score short spoken responses. The current study expands on these previous studies by introducing a new approach to segmenting the spoken response into discrete sections prior to assessing the content in each section and by introducing novel features to evaluate the accuracy of the content in the response.

3. Data

The data for this study was drawn from a pilot administration of the TOEFL® Junior™ Comprehensive test, an international assessment of English proficiency targeted at middle school students (aged from 11 to 15) which contains sections addressing the four components of English proficiency: Reading, Writing, Speaking, and Listening. The content-based Speaking test questions, or *items*, investigated in this study involve a task in which the test-taker first listened to an audio stimulus of a lecture or conversation containing several facts about a particular topic. The topics were drawn either from activities that are frequently done in middle school classes (e.g. writing a book report) or academic subjects appropriate for the targeted age range. While the test-taker is listening to the audio stimulus for each item, keywords are displayed on the computer screen to highlight the concepts that the test-taker will be asked to explain. After listening to the audio stimulus, the test-taker is given a fixed amount of time to prepare and then 60 seconds to provide a spoken response. The prompts emphasize that the response should contain information about each of the concepts highlighted during the presentation of the audio stimulus, and these keywords remain on the screen while the test-taker provides a spoken response.

During the course of developing the test material, additional resources were manually created for each item, including transcripts of the stimuli and "key points" (sample responses that contain the information that should be present in a high-scoring response). Each response was provided with a holistic score on a scale of 0 - 4 by expert human raters. The scoring rubrics addressed the following three main aspects of speaking proficiency: delivery (pronunciation, fluency, prosody), language use (grammar and lexical choice), and content.

The participants in this study were mostly students of middle school age residing in non-English speaking countries (average age = 13.1 years; s.d. = 2.2 years), and 15 different native language backgrounds were represented in the group. A total of 1700 participants are included in the study: 967 were used for the training corpus and 733 were used for the evaluation corpus. In the Speaking section, each participant responded to 3 content-based items, and a total of 6 different test forms were used. In this study, we focus only on a subset of 11 items that contain exactly four concepts (one for background information, which we refer to as *General*, plus three concrete fact-based concepts), and we exclude all responses that were flagged by raters as anomalous (e.g., because they contained a language other than English, the response was inaudible due to a technical difficulty, etc.). In total, 2048 spoken responses were available for training and 1568 were available for evaluation.

4. Methods

In our study we observed that the test-takers tended to organize the concepts in their responses in discrete segments of the response corresponding to the order in which they were discussed in the audio stimulus. Given this typical structure, we take an analytic approach to assessing the quality of the content contained in a given response: we propose to score a response's content with respect to each concept separately. This is accomplished by comparing the content of the response to pre-defined components that are expected from a proficient, on-topic response to each item (as described in Section 4.1) and an automatic segmentation of the transcriptions of the responses (as described in Section 4.2).

4.1. Item Content Analysis

To create a gold standard of the factual information that each item contains, we manually annotate the concepts for each item based on the relevant stimulus and response points provided by test developers. For each concept, we code its related factual information from four components: the *Name* of the concept (a keyword/phrase used in the prompt), the descriptive *Facts* (phrases that are likely to be included in a high quality response), the *Key Points* (a sample model response), and the *Context* (portions of the relevant stimulus that address this specific concept).

Take the *Frogs* item as an example.[1] After hearing a teacher present a lecture about the life cycle of frogs, the test-taker is presented with the following prompt: *"Talk about the physical changes a frog goes through. What happens at each stage? Be sure to include as many details as you can about each stage: tadpole, tadpole with legs, froglet, adult frog."* So, the student is expected to provide factual information about each of the four stages in a frog's life cycle that were discussed in the lecture, and these are the four concepts that constitute the item content

[1]The full content of this item can be viewed at *http://toefljr.caltesting.org/sampletest/s-frogs.html*.

Table 1: Item content analysis of the "Frogs" item.

Concept	Name	Facts
1	*General*	frog, physical changes, life, water, land, born, grows, moves
2	tadpole	first stage, water, little fish, tail, swim, gills, breath
3	tadpole with legs	second stage, little legs, like a frog, back, front
4	froglet	third stage, small frog, fully developed, tail, shorter, lungs, out of water, land
5	adult frog	last stage, adult, no tails, become, live on land, breathe air though lung

for this item. For a specific concept, e.g. "tadpole", the Name is *"tadpole"*; the Facts are *"first stage, water, little fish, tail, swim, gills, breath"*; and the Key Points regarding this concept are *"In the first stage, a frog is called a tadpole. A tadpole lives in water; it has a tail and breathes with gills."*; and the corresponding Context is *"In the first stage, when the frog is born in the water, it's called a tadpole. A tadpole looks a lot like a tiny little fish. Like a fish, it has a tail, and the tail is important because it helps it swim. It also has gills. That's another thing that a tadpole has that's like a fish. The gills are to help it breathe in the water. But this is just the first stage... The tadpole will go through more physical changes over the next few weeks."*.

The result of a complete concept analysis for this item is illustrated in Table1 for the *Name* and *Facts* content components. Due to space limitations, we do not present the *Key Points* and *Context* components.

4.2. Response Segmentation

We consider the segmentation problem within the context of our automated content scoring tasks as a pre-processing step on the transcripts. Before content feature extraction, we split responses into multiple self-contained text spans, each of which addresses a specific item concept (e.g. the concepts in Table 1). Specifically, we train a 1-order Hidden Markov Model[2] on manually transcribed and segmented responses to label each token of a response in terms of four labels (using indices from 1 to 4), and then split the response at the places where the labels of two successive tokens differ. Finally, each segment takes the same concept label as its tokens.

The automated segmentation model is trained on the features listed in Table 2;[3] we further tune the model for the best parameter settings on a held out sample set. In this study, three annotators segmented 625 human transcribed responses, which are used as the segmentation gold standard. For agreement analysis, all three annotators annotated a subset of 80 responses, in which $\kappa = .92$ on concept labels at the word level.

4.3. Content Feature Extraction

The proposed content scoring approach contains two steps for each response: first, we compute different features to measure the content information at the segment level using each of the four components of the item content analysis (as described in Section 4.1); second, we use various scoring functions to aggregate these segment-level content features to determine a score of content appropriateness for the entire response.

[2]We use the SVM representation of HMMs provided by SVM^{hmm} [17].

[3]To match a token with a n-gram fact, we compare all n-grams that contain that token against the fact to see if any of them match.

Table 2: Segmentation features

#	Description
1	Token index
4	Indicator of the presence of each concept name
4	Indicator of the presence of any fact from each concept
4	Signed distance to the closest occurrence of each concept name
4	Signed distance to the closest occurrence of any fact of each concept
32	Part-Of-Speech tags

4.3.1. Segment content features

We develop four segment content features (one for each component in the item content analysis) to measure the content information about concepts *c* contained in a segment, *s*, of a response, listed in order of increasing amount of information:

- $f_{name}(s, c)$: number of occurrences of the *Name* of concept *c* in the segment *s*.
- $f_{fact}(s, c) = f^{abs}_{fact}(s, c)/f^{abs}_{fact}(c_{points}, c)$: the scoring of the segment *s* with respect to the set of facts of concept *c* and then normalized by the score computed from the *Key Points* of *c*, where $f^{abs}_{fact}(s, c)$ computes the number of the *Facts* of *c* occurred in *s* weighted by the fact token length, plus the sum of the average unigram frequency of each *Fact*.
- $f_{points}(s, c)$: text to text similarity between *s* and the *Key Points* of *c* using WordNet.
- $f_{context}(s, c)$: text to text similarity between *s* and the *Context* of *c* using WordNet.

$f_{name}(s, c)$ captures whether a particular concept is mentioned in a response, while $f_{fact}(s, c)$ gives credit to relevant descriptive details including phrases that are matched exactly as well as their variations, thus adding more robustness to ASR errors in comparison to features based on exact string matching or syntactic features. In addition, we also consider traditional content scoring methods based on text to text similarity metrics derived from WordNet [18], where we use $f_{points}(s, c)$ and $f_{context}(s, c)$ to compare a segment against our "model response" and the concept's context, respectively.

4.3.2. Response content features

For aggregation, we first group the segments of a response by concept labels (denoted as $r(c_i)$), and compute the maximum feature value within each $r(c_i)$ for each concept c_j from each

component a:

$$h_a(i,j) = \max_{s \in r(c_i)} f_a(s, c_j) \quad (a \in \{name, concept, points, context\}) \tag{1}$$

For a given component a, we then summarize all $h_a(i,j)$ into a n by n matrix H_a, where n is the number of related item concepts.[4] Intuitively, H models how much information of concept c_j is covered by the part of the response $r(c_i)$ that is supposed to address concept c_i, as indicated by the value at position (i,j):

$$H_{n,n} = \begin{pmatrix} h(1,1) & h(1,2) & \cdots & h(1,n) \\ h(2,1) & h(2,2) & \cdots & h(2,n) \\ \vdots & \vdots & \ddots & \vdots \\ h(n,1) & h(n,2) & \cdots & h(n,n) \end{pmatrix}$$

We propose two scoring functions over H to aggregate the segment content features into response content features for any component. The first scoring function S_{mean} computes the mean of the sum of every column, which measures on average how each concept is addressed in the whole response.

$$A_{n,n} = \begin{pmatrix} 1 & 1 & 0 & \cdots & 0 \\ 1 & 1 & 1 & \cdots & 0 \\ 0 & 1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{pmatrix}$$

In contrast, the second scoring function S_{matrix} (Formula 2) takes the segment difference into consideration: it constrains the scoring of the response's content for c_j to be based on the corresponding segments which address c_j, which are the numbers on the diagonal ($H(j,j)$). Considering possible segmentation errors and the transition from one concept to another in responses, we relax the constraint to also enable scoring on the segments of the preceding concept ($H(j-1,j)$) and the following one ($H(j+1,j)$), when they are available, in the process of scoring the response's content on c_j. This scoring matrix is denoted as A. In addition, we introduce a penalty component, which penalizes when a response regarding a specific concept carries more information about another concept (indicating that the response's content is inaccurate).[5] Finally, we denote the penalty matrix as $B-I$, where B is an indicator matrix of which concept is developed most in the segments of each concept, and I is an identity matrix. Ideally, $B = I$ and thus $B - I = 0$.

$$S_{matrix}(H) = H \cdot A - 0.5H \cdot (B - I) \tag{2}$$

We compute $S_{mean}(H)$ and $S_{matrix}(H)$ for each component to measure the response content quality for automated scoring. Note that our construction of the content features as described above can be easily adapted to items of any number of concepts, as long as responses can be segmented correspondingly.

5. Experiments

To evaluate our analytic-based content features, we conduct two intrinsic evaluations based on a sample set of three items (200 responses), and one extrinsic evaluation by means of a scoring model on all 11 items selected for analysis (see Section 3 for details). For the intrinsic evaluations, we used the responses from only the training set of the scoring model: we first investigate the utility of our content features, based on their correlation with scores provided by human experts across four content components between two scoring functions; we then compare our best content feature based on automated segmentation with the best CVA-based feature. For the extrinsic evaluation, we evaluate our best content features in the context of a scoring model containing both the proposed content features and features related to other components of a non-native speaker's proficiency.

5.1. Analysis of content features on human transcriptions

First we compare different content features on 200 manually transcribed responses for 3 items. We randomly select 92 responses to train the HMM segmentation model, and compute the content features $S_{mean}(H)$ and $S_{matrix}(H)$ on 108 testing responses based on various segmentation output: 1) human segmented results (S-manual), 2) HMM model predictions (S-auto) and 3) no segmentation, i.e., considering the whole response as one single segment (S-no). Note that we evaluate our segmentation model (S-auto) within our analysis of the utility of the content scoring features, as it is designed to be the pre-processing step performed prior to content feature extraction. For comparison, we visualize the performance of the content features computed from three segmentation models in a group, across 4 components for the two features $S_{mean}(H)$ (Figure 1)[6] and $S_{matrix}(H)$ (Figure 2).

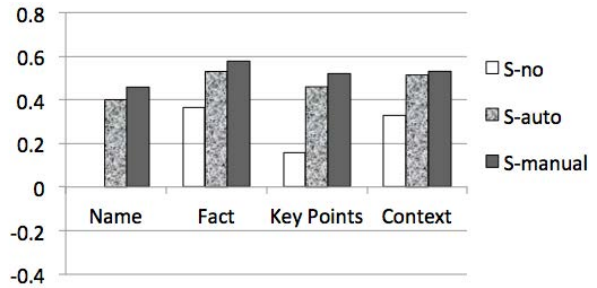


Figure 1: *Features' correlation with response scores using scoring method $S_{mean}(H)$. The segmentation models used for each component are S-no, S-auto and S-manual, from left to right.*

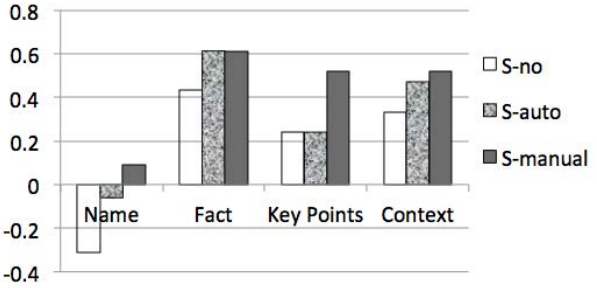


Figure 2: *Features' correlation with response scores using scoring method $S_{matrix}(H)$.*

[4] n=4 in this study.

[5] As a preliminary study, we set the coefficient of the penalty item (0.5) based on our intuition.

[6] The value of the $S_{matrix}(H_{name})$ feature is undefined for the S-no segmentation condition and is thus not included in Figure 1.

As Figure 1 and Figure 2 show, the *Fact* component yields the best performance regardless of the segmentation model and the scoring function. Although $S_{matrix}(H)$ results show a less consistent pattern across components (Figure 2), it yields better *Fact*-based content features compared to $S_{mean}(H_{fact})$ for all segmentation models. With respect to the impact of segmentation, in general, manual segmentation always works better than machine-based segmentation models, and segmentation improves the features' performance for all components. More importantly, for the best feature component, the performance of $S_{matrix}(H_{fact})$ generated by manual segmentation and machine segmentation are almost the same. This suggests that even though automated segmentation may introduce errors, we can still assess response content based on factual information quite reliably, by applying S_{matrix}. (In other words, $S_{matrix}(H_{fact})$ can work equally well when moderate segmentation errors exist.)

5.2. Human transcriptions vs. ASR output

To test whether our approach works equally well on ASR output, we compute the content features based on the automated segmentation of the ASR output of the sample testing set, which contains 3 items.[7] Here we retrain S-auto on all available gold-standards ($N = 625$).[8] Due to space limitations, we only present the results of $S_{mean}(H_{fact})$ and $S_{matrix}(H_{fact})$ – the two best-performing features identified in previous sections.

Table 3: Comparison of the *fact*-based content features between human transcriptions and ASR output.

	Trans_human		Trans_ASR	
	S_{mean}	S_{matrix}	S_{mean}	S_{matrix}
S-no	.367	.437	.403	.459
S-auto	.443	.637	.545	.615

5.3. Comparison with CVA

To validate the approach involving manual item content analysis, we compare our best content features computed using automated segmentation with the features generated by Content Vector Analysis (CVA). CVA was chosen as a baseline since it is a widely used alternative for text scoring and it only requires the response texts and the associated gold-standard scores for training.

In this experiment, we compute the two proposed content features and the CVA features from using human transcriptions from the sample data set (200 responses), in order to be able to make a fair comparison based on the upper bound of the features' performance. Furthermore, the CVA features are computed as leave-one-out evaluation, which includes the similarity between a given response and the model vector at each score level (e.g. $SimToScore\_1$), as well as the score level of the model vector that it is closest to. For a given response, the model vectors are computed on all the other responses of the specific item.[9]

[7]Similar patterns were observed when we extended the comparison to the other 8 items.

[8]The word-level accuracy of the updated segmentation model is 77% when trained and tested on all available gold-standards.

[9]Because the response is item dependent, we only consider the responses to the same item when computing the model vectors.

Table 4 compares the features' performance on the testing set. It shows that our best content feature ($S_{matrix}(H_{fact})$, $r = 0.612$) outperforms the best CVA feature ($SimToScore\_3$, $r = 0.440$) by 40% relative.

Table 4: Comparison between the best proposed content feature and the best CVA baseline feature.

	Proposed features	CVA features
Best r	.612	.440

5.4. Automated Scoring

Finally, we evaluate the proposed content features by examining their influence on a scoring model that is designed to predict holistic English speaking proficiency scores provided by expert raters. The raters provided a single discrete score for each response on a scale of 1 - 4, and were instructed to take into account detailed rubrics for the following components of speaking proficiency while providing each score: fluency, pronunciation, stress, intonation, grammar, word choice, and content. Based on a set of 2048 responses with scores, a linear regression model was trained with 8 features extracted using SpeechRater, an existing automated scoring system designed for spontaneous speech [19]. These features included measurements of a speaker's rate of speech (fluency), words per breath group (fluency), rate of long silences (fluency), acoustic model score (pronunciation), phone duration score (pronunciation), rate of stressed syllables (stress), rate of lexical types (fluency and word choice), and language model score (grammar and word choice). The model's correlation on the unseen set of 1568 responses (no speaker overlap) is $r = 0.624$.

To evaluate the contribution of the proposed content scoring approach to this model, the two top-performing content features, $S_{mean}(H_{fact})$ and $S_{matrix}(H_{fact})$, were added to the model. After the addition of these two features, the updated linear regression model obtained a correlation of $r = 0.664$ on the same evaluation set of 1568 responses, representing a statistically significant improvement of 0.04 over the baseline model with no content features ($t = 6.3$, $p < 0.001$).[10]

6. Discussion

In this paper, we propose a method for assessing the quality of the content of spoken responses based on segmenting the response and evaluating how the content contained in each of the segments relates to the test question at various levels of detail. Our experimental results show that the content features based on factual information relating to each concept perform best with both automated and manual concept-based segmentation, although the performance of the other features is typically within around 10% of the fact-based features. As described in Section 4.1, we define Facts to be short keywords or phrases that provide supporting details for the content related to each concept in a response. One drawback of this approach is that these facts must be manually extracted from the test questions prior to automated scoring; however, one can imagine that such information could be automatically constructed given the relevant resources using NLP techniques. In this paper we mainly focus on content feature engineering, and thus use these manually provided resources in our experiments.

[10]We use William's test of significance for dependent correlations.

Another important finding from this study is the fact that the proposed content features performed better after the segmentation procedure, and the automated segmentation model works comparable to a human gold standard in terms of their impact on $S_{matrix}(H_{fact})$ (our best content feature). This finding suggests a structured nature of the spoken responses which is induced by the way in which the test question is asked. Thus, the design of this speaking task may enable the use of a more targeted approach to content evaluation than has typically been employed in the past.

7. Conclusion

We have shown in this paper that it is feasible to evaluate the content of spontaneous spoken responses in a language test automatically, using features related to factual information in the responses and automated speech recognition to obtain transcriptions of test takers' spoken responses. While human experts are needed to generate the phrases containing factual information for each test item, this effort only needs to be undertaken once for each new test form, is not very time consuming, and could conceivably be done in a semi-automated manner in the future.

We also demonstrated that these content features are beneficial for developing automated scoring systems for spontaneous speech since they improve the agreement with expert human ratings and expand the aspects of speaking proficiency that are covered by the scoring system.

In future research, we plan to investigate methods of automatically extracting the Facts contained in the stimulus materials and extending the approach to additional types of structured, content-based spoken test responses, such as narrative retellings.

8. References

[1] P. Price, J. Tepperman, M. Iseli, T. Duong, M. Black, S. Wang, C. Boscardin, M. Heritage, P. David Pearson, S. Narayanan *et al.*, "Assessment of emerging reading skills in young native speakers and language learners," *Speech Communication*, vol. 51, no. 10, pp. 968–984, 2009.

[2] M. Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

[3] J. Balogh, J. Bernstein, J. Cheng, and B. Townshend, "Automatic evaluation of reading accuracy: assessing machine scores," in *Proceedings of the ISCA Tutorial and Research Workshop on Speech and Language Technology in Education (SLaTE)*, 2007, pp. 1–3.

[4] S. Xie, K. Evanini, and K. Zechner, "Exploring content features for automated speech scoring," in *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Montréal, Canada: Association for Computational Linguistics, 2012, pp. 103–111.

[5] C. Leacock and M. Chodorow, "C-rater: Automated scoring of short-answer questions," *Computers and the Humanities*, vol. 37, no. 4, pp. 385–405, 2003.

[6] J. Z. Sukkarieh, S. Pulman, and N. Raikes, "Auto-marking 2: An update on the UCLES-Oxford university research into using computational linguistics to score short, free text responses." in *International Association of Educational Assessment*, 2004.

[7] S. Pulman and J. Z. Sukkarieh, "Automatic short answer marking," in *Proceedings of the 3rd NAACL-HLT Workshop on Innovative Use of NLP for Building Educational Applications*, 2005, pp. 9–16.

[8] M. Mohler and R. Mihalcea, "Text-to-text semantic similarity for automatic short answer grading," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, 2009, pp. 567–575.

[9] D. Higgins, J. Burstein, and Y. Attali, "Identifying off-topic student essays without topic-specific training data," *Natural Language Engineering*, vol. 12, 2006.

[10] Y. Attali and J. Burstein, "Automated essay scoring with e-rater v.2.0." in *Presented at the Annual Meeting of the International Association for Educational Assessment*, 2004.

[11] C. Corley and R. Mihalcea, "Measuring the semantic similarity of texts," in *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, 2005, pp. 13–18.

[12] C. Fellbaum, *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press, 1998.

[13] P. D. Turney, "Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL," in *Proceedings of the Twelfth European Conference on Machine Learning (ECML)*, Freiburg, Germany, 2001, pp. 491–502.

[14] I. Dagan, O. Glickman, and B. Magnini, "The PASCAL recognising textual entailment challenge," in *Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment*, J. Quiñonero-Candela, I. Dagan, B. Magnini, and F. d'Alché Buc, Eds. Springer, 2006, pp. 177–190.

[15] M. Mohler, R. Bunescu, and R. Mihalcea, "Learning to grade short answer questions using semantic similarity measures and dependency graph alignments," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011, pp. 752–762.

[16] F. Huang and L. Chen, "Scoring spoken responses based on content accuracy," in *Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications, NAACL-HLT*. Montréal, Canada: Association for Computational Linguistics, 2012.

[17] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *Journal of Machine Learning Research*, vol. 6, no. 2, p. 1453, 2006.

[18] C. Leacock and M. Chodorow, "Combining local context and wordnet similarity for word sense identification," in *WordNet: An Electronic Lexical Database*, C. Fellbaum, Ed. MIT Press, 1998, pp. 305–332.

[19] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

Naturalness on Japanese Pronunciation before and after Shadowing Training and Prosody Modified Stimuli

Rongna A [1], *Ryoko Hayashi* [2], *Tatsuya Kitamura* [3]

[1]Department of Rehabilitation for Sensory Functions, Research Institute of National Rehabilitation Center for Persons with Disabilities, Tokorozawa, Japan
[2]Graduate School of Intercultural Studies, Kobe University, Kobe, Japan
[3]Faculty of Intelligence and Informatics, Konan University, Kobe, Japan

hohchahar-aruna@rehab.go.jp, rhayashi@kobe-u.ac.jp, t-kitamu@konan-u.ac.jp

Abstract

This study attempts to investigate the change of naturalness impression for the Japanese utterance by Japanese as foreign language learners (JFL) before and after pronunciation training (shadowing / repeating), and to discuss the crucial prosodic cue for the naturalness judgment.

The speech of 8 JFL learners before and after pronunciation training was used, and their durational pattern, pitch pattern, or both of them were replaced with those of the model speech. 52 Japanese native speakers (JNS) assessed the naturalness of these stimuli. The results showed JNS judge the duration AND F0 modified stimuli most natural. In addition, the shadowing trained group tended to have been valued highly than the repeating trained group after the training. Furthermore, the acoustical analysis of speech material showed a difference of moraic structure and pitch accent between the shadowing and repeating group.

Index Terms: Japanese as a foreign language learners, naturalness, shadowing, repeating, prosodic modified stimuli

1. Introduction

Japanese is known as a mora-timed language and has lexical pitch-accent. These characteristics cause the learners of Japanese (JFL) difficulty to control speech timing and realize correct word accents [1, 2]. Wrong timing control and pitch accent, however, causes misunderstanding of word meaning and unnaturalness of speech. The acquisition of mora and pitch accent is considered to be very important for JFL learners despite the existence of regional varieties of pitch accent [3, 4]. In recent years, the importance of pronunciation training has received remarkable attention in teaching Japanese. In our previous studies, we used shadowing training to practice JFL's pronunciation [5, 6, 7]. Shadowing requires learners to listen to the model sentences while repeating almost simultaneously: The learners should repeat the sentences as exactly as possible while listening to the in-coming information [8]. Shadowing is said to improve prosodic features of learners' pronunciation [9]. Our previous studies showed that the speech rate and accuracy of pitch-accent in the learners' pronunciation rose radically during the *shadowing* training and this effect lasted also after the training [5]. However, it remained unknown how the naturalness of the learners' pronunciation was changed by the shadowing training.

Furthermore, the crucial prosodic cue for Japanese native speakers' (JNS) naturalness judgment is still in discussion. Sato [10] conducted a naturalness judgment test of Japanese utterances of a Korean native speaker and a Chinese native speaker and modified stimuli, i.e., one or all of the prosodic features were replaced by that or those of the native speaker's utterance: pitch pattern, durational pattern or intensity of the sentence. As a result, he concluded that pitch pattern is the most important prosodic feature for JNS. On the other hand, Tsurutani [11] showed that Japanese native speakers put more weight on accuracy in timing (durational pattern) than in pitch when judging the naturalness of JFL learners' speech. However, in [11] a Japanese-English bilingual speaker, with near native degree of fluency in both languages, was asked to utter the speech materials with perfect model pronunciation, and also with absolute beginner's pronunciation containing all the classical errors. The two previous studies used different types of stimuli, and they have different results.

The purpose of this paper is, thus, to compare the naturalness judgments by JNS before and after shadowing training. As mentioned above, after the shadowing training, the pitch accent pattern and the speech rate (durational structure) are easily improved [5]. Our question is to what extent the corrected pitch accent and durational structure are important for JNS to perceive the speech as natural. To resolve this question, the naturalness judgment of the learners' utterance was compared with the synthesized speech and the crucial prosodic cue for the naturalness judgment was also explored at the same time.

2. Methods

2.1. Speech materials

The natural stimuli were taken from our previous studies [6, 7], in which there were thirty three Chinese and Mongolian JFL learners who participated in two kinds of short term pronunciation training. All participants majored in Japanese language department at Inner Mongolian University. They learned Japanese for three years, and their level of Japanese was intermediate then. 19 of them trained to pronounce sentences with shadowing (shadowing group) and the other 14 with repeating (repeating group). In shadowing training, the JFL were instructed to imitate the model speech almost simultaneously as soon as possible. In the repeating training, the learners are required to repeat the model speech, presented as a short phrase, after listening to the end. In both training methods, the speech produced by a native male speaker of Japanese was used as the model speech. The text the participants read was always the same at pre-, post- oral reading and during training, and was chosen from [12], consisting of 656 morae, with the total

duration of the model speech for 117.6 seconds. The data were recorded both before and after the training at a sampling rate of 48,000 Hz with 16-bit resolution using an IC recorder (Roland EDIROL, R-09) individually.

At our preliminary study [13], the naturalness of their utterances before and after training was evaluated with a Likert scale (ranging from 1: extremely unnatural to 5: extremely natural = native-like) by ten JNS. The sentences taken from the utterances by JFL were two sentences and consisted of 47 moras. *Minasan wa okashi o yoku tabemasu ka? Moo otona dakara okashi wa amari tabenai toiu hito mo ooi deshoo* (Do you all eat snacks often? Perhaps someone says that he doesn't eat snacks so much because he is already grown-up).

Ten JNS listened to the speech of JFL, and judged the naturalness from three viewpoints: *overall impression*, *the correctness of speech* and, *the correctness of rhythm*. The present study used the evaluation of *overall impression*, and chose four JFL learners' utterance (two Chinese and two Mongolian) from each group of which naturalness was evaluated as almost the same (2.8 and 2.9 points) at pre-reading.

2.2. Stimuli

The following five types of stimuli were used:

pre: speech data recorded before the training,
post: speech data recorded after the training,
dur: speech data with the duration converted into that of the model speech,
F0: speech data with the F0 converted into that of the model speech
durF0: speech data with the duration and F0 converted into those of the model speech.

Stimuli dur, F0, and durF0 were synthesized using Praat [14]. The source signal of dur, F0, and durF0 was pre. First, the speech data were labeled at the phoneme level. Stimulus dur was then synthesized so that the duration of each phoneme of the speech data was corresponded to that of the model speech. Time-Domain Pitch-Synchronous Overlap-and-Add (TD-PSOLA) method [15] was employed to the phoneme-level time warping. Stimulus F0 was synthesized by replacing the F0 of the speech data with that of the target speaker. The F0 of the target speaker was time-warped at the phoneme level to fit the time structure of the learner's speech. Stimulus durF0 was synthesized by replacing the F0 of Stimulus dur with that of the model speech. The last stimulus thus has the duration and the F0 of the model speech and the spectral envelopes of the original speaker.

2.3. Naturalness judgment

Two types of natural stimuli (pre and post) and three types of modified stimuli (dur, F0 and durF0) were used in the present experiment. Each of the 5 type of stimuli includes 8 speakers' stimuli: 4 from shadowing group and 4 from repeating group. Together with the model speech and five dummies, 46 stimuli in total were randomized and evaluated by 52 native Japanese speakers. Among them, 26 were Kansai dialect speakers (12 females, 14 males, average age 18.6) and 26 were Tokyo dialect speakers (15 females, 11 males, average age 21.3).

The JNS were required to assess the naturalness of stimuli using a Likert scale with potential responses ranging from 1 (extremely unnatural) to 7 (extremely natural = native-like). Before the real assessment experiment, we presented 5 sample stimuli for practice.

3. Results

There are no significant differences in pair-wise comparisons between Kansai dialect speakers and Tokyo dialect speakers (shadowing group: $F(1, 6) = 0.06$, n.s.; repeating group: $F(1, 6) = 0.01$ n.s.) at 5 type of stimuli. Therefore, we use the scores of naturalness from all of the 52 native Japanese speakers.

Figure 1 shows the mean evaluation score for each stimuli type. The naturalness judgment for the model speech was evaluated as 7 (SD=0). The shaded bars indicate the shadowing groups, and the white bars indicate the repeating groups. Two-way analysis of variance with group (shadowing and repeating group) and type of stimuli (pre, post, dur, F0, durF0) showed a significant main effect in type of stimuli ($F(4, 24) = 29.51$, $p < .001$) and marginal significance in interaction ($F(4, 24) = 2.62$, $p = .06$). The results of multiple comparisons of Bonferroni-test show the scores of durF0 were significantly higher than others in both training groups ($p < .05$). For stimuli post, the score of the shadowing group tends to be higher than that of the repeating group ($F(1, 6) = 5.73$, $p = .05$). The mean score of dur was higher than that of post in the repeating group ($p < .05$).

In other words, the JNS judged the stimuli durF0 most natural. The score became significantly higher only if both prosodic cues - duration AND pitch pattern - of the utterance were corrected. The results showed also that in the stimuli post, that is, the utterance after training, the shadowing group tended to get higher one point score than did the repeating training group.

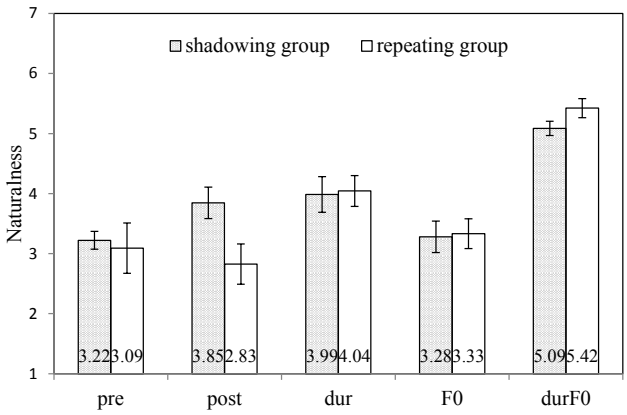


Figure 1: *Result of Naturalness judgment. Error bars indicates standard errors.*

4. Discussion

In the present study, it is revealed that in the natural speech after training (stimuli post) the shadowing group tended to get higher scores than the repeating group. In order to clarify the reason, the durational pattern and accent pattern of the natural stimuli pre and post was analyzed.

4.1. The duration of moras

As is well known, Japanese is a mora-timed rhythm language it has the characteristics of isochronal mora timing [16, 17]. At this present study, we measured the duration of each mora, but the

moras prior to the pause (the end of *bunsetsu*) were excluded because of the prepausal lengthening [18]. The results show the mean duration of a mora in model speech was 98.2ms (SD=28.6, naturalness judgment in the former section was 7), while that in JFL speech before training was 133.7ms (SD=50.0, naturalness 3.2) in the shadowing group and 134.4ms (SD=45.5, naturalness 3.1) in the repeating group. After the training, the difference between the model speech and JFL' speech became smaller in the shadowing group: the mean duration of a mora in the shadowing group was 117.1ms (SD=38.8, naturalness 3.8) and in the repeating group was 132.5ms (SD=45.2, naturalness 2.8). The mean duration and standard deviation of mora in the JFL's utterance were longer than those in the model speech. However, after the shadowing training, the mean duration and standard deviation of mora had been got closer to model speech.

To see the deviation of durational patterns in the utterance by JFL from model speech, the duration of each mora was measured and the deviation of the moraic structure (DM) was calculated as follows:

$$\frac{}{}$$

m: number of moras in utterance
d: duration of mora in model speech
d': duration of mora in JFL speech

Figure 2 shows the mean value and standard deviation of DM. Two-way repeated measures ANOVA was performed between groups (shadowing, repeating) and stimuli (pre, post). The result showed a significant main effect ($F(1, 6) = 33.41$, $p < .01$) and interaction ($F(1, 6) = 10.38$, $p < .05$). Post hoc analysis showed that the deviation of the shadowing group is smaller than the repeating group in the stimuli post ($p < .05$). That is to say, after the shadowing training the durational deviation from model speech was reduced at the mora level.

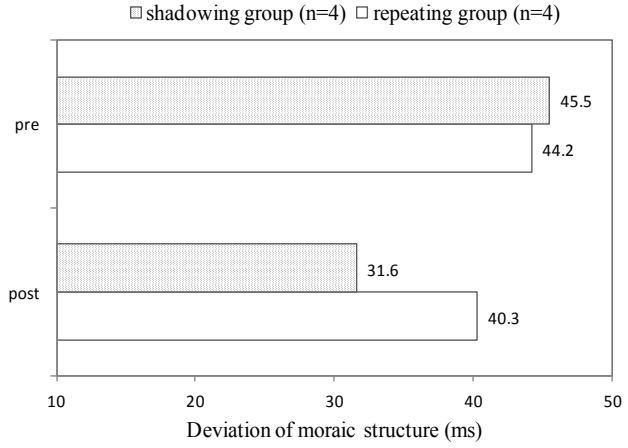


Figure 2: *The deviation of moraic structure*

4.2. Accuracy of word accent

To investigate the importance of correct pitch accent, a native Japanese speaker who had received phonetic training (the second author) listened and judged if the pitch pattern of JFL was same with the model speech. The words to be judged were 13, e.g., all nouns, verbs, adjectives and adverbs underlined: *Minasan wa okashi o yoku tabemasu ka? Moo otona dakara okashi wa amari tabenai toiu hito mo ooi deshoo* .

Mean accuracy of pre and post in both groups are shown in Figure 3. Two-way repeated measures ANOVA was performed between groups (shadowing, repeating) and stimuli (pre, post). The accuracy of word accent among the tasks showed no significant difference between shadowing group and repeating group ($F(1, 6) = 1.05$, n.s.), and between pre and post ($F(1, 6) = 0.05$, n.s.). However, the shadowing group tended to show higher accuracy than the repeating group after training (post) about 11.6 points.

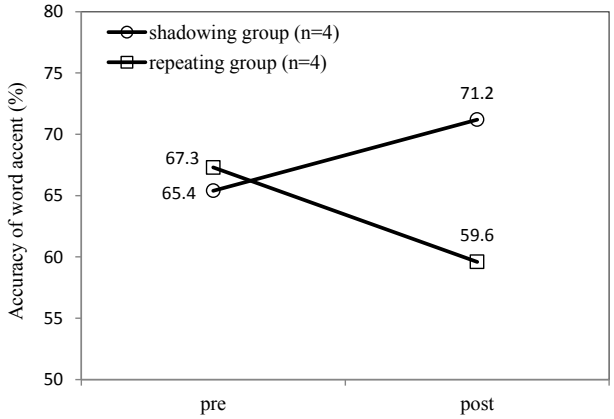


Figure 3: *The accuracy of pitch accent.*

Our previous study observed the effectiveness of shadowing training for JFL, they showed that speech rate and accuracy of pitch accent were substantially changed after training [5]. The present study attempted to see the improvement of JFL speech by shadowing training from the view point of the native speakers' impression and to find the crucial prosodic cue for the naturalness judgment. The shadowing group got higher score of naturalness judgment after training than the repeating group. This result could refer to the difference of moraic duration and accuracy of pitch accent.

The present study showed a small tendency that dur stimuli were judged more natural than F0 stimuli, but there was no significant difference like [11]. The reason why Sato [10] concluded that pitch pattern is more important than duration pattern for JNS could probably be because of the difference of the learners level, which was intermediate. Therefore, the timing control in their utterances could have been better than that in the stimuli we used in the present study. In fact, in [10], the judgment for natural stimuli was 5.21 point (7 point scale) and higher than that of pre in the present study.

The other reason for the slight increase of the naturalness in dur could be thought to be the effect of pause. Analyzing the original speech of JFL learners, it was often found that the duration of pause was shorter than the model speech. This fact is in agreement with [19]. In addition, the number of pauses was quite different in each utterance. In the present study, however, we used two sentences and the pause between the sentences in dur stimuli was modified the same as that of the model speech.

The shadowing training requires the learners to imitate as soon as possible after the model speech was heard. This task can affect easily the durational structure (speech rate) including pauses.

Moreover, it is revealed that neither durational pattern nor F0 pattern alone caused significant increase of naturalness judgment scores by native speakers. The crucial prosodic cue is neither the durational and nor the F0 pattern, but both. For the JFL education it is important to pay attention to both prosodic cues and they cannot be treated separately.

5. Conclusions

The present paper reports about native speakers' impressions of JFL utterances before and after shadowing/repeating training. Evaluation was also done for the synthesized stimuli in order to examine what the crucial prosodic cue might be. The results suggest that both pitch patterns and durational patterns are important in order for the utterances to be heard as natural Japanese and it is difficult to discuss the cues separately.

This results of the present study could provide useful suggestion for developing pronunciation training for JFL learners.

6. Acknowledgements

The authors would like to thank for staffs and students at Inner Mongolian University, Kokugakuin University and Kobe University. This research was supported by Grants-in-Aid for Scientific Research, project No. 21242013, No. 24652101 and No. 23242023, from the Japan Society for the Promotion of Science.

7. References

[1] Ôtsubo, K., "Onsei kyôiku-no mondaiten [The problems of speech education]", Kôza Nihongo-to nihongo kyôiku 3 Nihongo-no onsei on'in (ge) [The course of Japanese and Japanese Education 3: Japanese Phonetics and Phonology], Meijisho'in, 23-46, 1990.

[2] Ayusawa, T., "Gaikokujin gakushusha-no nihongo akusento / intonêshon shûtoku [Acquisition of Japanese Accent and Intonation by Foreign Learners]", Onsei kenkyû [Journal of the Phonetic Society of Japan], 7(2), 47-58, 2003.

[3] Tanaka, S., Kubozono, H., "Nihongo-no hatsuon kyôshitsu: Riron-to renshû [Introdution to Japanese Pronunciation Theory and Practice]", Kurosio, 1999.

[4] Isomura, K., Onsei-wo oshieru [Teaching Japanese Phonetics], Hitsujishobo, 2009.

[5] A, R., Hayashi, R., "Accuracy of Japanese pitch accent rises during and after shadowing training", Proceedings of the 6th International Conference on Speech Prosody 2012, 214-217, 2012.

[6] Arona, Hayashi, R., "Shadôingu renshû-niyoru nihongo hatuon-no henka: mongorugo/chûgokugonogowasha-wo taishô-ni [The effect of shadowing training for Mongolian and Chinese learners of Japanese]", IEICE Technical Report 109, 451, 19-24, 2010.

[7] A, R., Hayashi, R., "Shadôingu kunren-niyoru nihongo gakushûsha-niokeru go-akusento-no henka [The change of word accent in shadowing training for Japanese as foreign language learners]", Kotoba-no Kagaku kenkyû [Journal of the Japan Society for Speech Sciences], 12, 57-71, 2011.

[8] Tamai, K., "Risuningu shidôhô toshiteno shadowing kôka nikansuru kenkyû [A study on the effects of shadowing as an instructional method of listening]". Tokyo: Kazamashobo, 2005.

[9] Mori, Y., "Shadowing with Oral Reading: Effects of Combined training on the Improvement of Japanese EFL Learners' Prosody", Language Education & Technology, No.48, pp. 1-22, 2011.

[10] Sato, T., "Tan-on to inritdu ga nihongo onsei no hyooka ni Ataeru eikyoo [Comparison of the influence of segmental information and prosody on the assessment of Japanese pronunciation]", Sekai no Nihongo kyoiku [The global Japanese-language education], 5, 139-154, 1995.

[11] Tsurutani, C, Ishihara, S., "Naturalness Judgement of Prosodic Variation of Japanese Utterances with prosody Modified Stimuli", Proceedings of Interspeech 2012, 2012.

[12] Miyagi, S. *et al.*, "Mainichi no kikitori purasu 40 [Everyday listening plus 40]", Bonjinsha, 2003.

[13] Yasuda, R., A, R., Hayashi, R., "Training efficacy of shadowing vs. reading/repetition for Chinese and Mongolian learners of Japanese: A perceptual study", Proceedings of the auditory research meeting, 41(3), 223-227, 2011.

[14] Praat, http://www.fon.hum.uva.nl/praat/

[15] Moulines, E., Charpentier, F., "Pitch-Synchronous Waveform Processing Techniques for Text-to-Speech Synthesis Using Diphones", Speech Communication, 9(5), 453-467, 1990.

[16] Kubozono, H., Oota, S., On-in koozoo to akusento [Phonological structure and Accent], Kenkyusha Syuppan, 1998.

[17] Kubozono, H.: Speech segmentation and phonological structure. In T. Otake and A. Cutler (eds.) Phonological Structure and Language Processing: Cross-linguistic Studies. Berlin : Mouton de Gruyter, 77-94, 1996.

[18] Takeda, K., Sagisaka, Y., Kuwahara, H., "On sentence-level factors governing segmental duration in Japanese", J.Acoust. Soc. Am., 86(6), 2081-2087, 1989.

[19] Ishizaki, A., "Nihongo no ondoku nitsuite gakushusha wa donoyôni pôzu wo okuka - eigo/huransugo/chugokugo/kankokugo wo bogo to suru gakushusha to nihongobogowasha no hikaku-[How the JFL learners are put the pause in Japanese oral reading: a comparison in Japanese learners with English, France, Chinese, Korean native speakers]", Sekai no nihongo kyoiku [The global Japanese-language education], 15, 75-89, 2005.

Quantifying and Evaluating the Impact of Prosodic Differences of Foreign-Accented English

Hansjörg Mixdorff [1] *and Murray J. Munro*[2]

[1] Department of Computer Science and Media, Beuth University Berlin, Germany
[2] Department of Linguistics, Simon Fraser University, Vancouver, Canada
mixdorff@beuth-hochschule.de, mjmunro@sfu.ca

Abstract

The identification and correction of prosodic deviations in second-language speech still poses a significant challenge for computer-aided language learning. With this ultimate goal in mind, the current study compares utterances by Cantonese speakers of Canadian English with those of native English subjects through both acoustic analysis and perceptual evaluation. We aim to find measurable prosodic differences accounting for the perceptual results. Our outcomes indicate, inter alia, that unstressed syllables are relatively longer compared to stressed ones in the Cantonese corpus than in the Canadian English corpus. Furthermore, the correlations of syllabic durations in utterances of one and the same sentence are much higher for Canadian English subjects than for Cantonese speakers. The latter use a similar range of *F0*, but produce more and longer pitch-accents than Canadian English speakers. In a perception study we found that applying native durations together with F0 contours to the foreign-accented speech led to significantly improved listener judgments of prosodic goodness. Adjustments to duration alone also tended to yield better ratings, though the effect was not statistically significant. When durations of native English utterances were adjusted to those of Cantonese speakers, significant decrements in ratings were observed.

Index Terms: foreign accent, prosodic analysis, perception tests

1. Introduction

Although foreign accent is commonly attributed to segmental deviations from the native norm, prosodic differences certainly account for many difficulties in understanding accented speech (see, for instance, [1][2]). In the current paper we examine speech from Cantonese users of Canadian English collected for two earlier studies [3][4]. We perform an acoustic prosodic analysis of the material and compare their speech with corresponding utterances by native Canadian English subjects in order to establish objective parameters that best reflect foreign accent, and are correlated with listeners' judgments of prosodic accuracy. Whereas English is often classified as a stress-timed language, Cantonese is a syllable-timed tone language, a contrast which poses a number of prosodic problems for learners of the other language.

Although pedagogical specialists often identify accurate prosody as a critical aspect of pronunciation teaching [5], so far only a handful of empirical studies have evaluated its role in non-native speech. Hahn [6] reported that accurate stress placement enhanced listeners' recall of main ideas from L2 speakers' utterances; Derwing, Munro & Wiebe [7] found that instruction on prosodic features improved comprehensibility of L2 narrative productions, while segmental instruction did not; and Tajima, Port, and Dalby [8] observed that temporal adjustments to L2 speech increased intelligibility.

In the first part of the current paper we present a comparative acoustic analysis of prosodic features of Cantonese and Canadian English speakers. The second part explores the perceptual impact of native prosody in terms of duration and F0 transplanted to the utterances of Cantonese speakers, as well Cantonese duration characteristics applied to Canadian English speakers.

2. Speech Material and Method of Analysis

The original corpus consists of readings of a short English passage adapted from [9] by 77 native adult speakers of Cantonese with English pronunciation skills ranging from poor to good. All were born in Hong Kong and had moved to Canada after the age of 16, where they had been residing for 1 to 4 years. We selected this pre-existing corpus for the large number of available subjects as well as the fact that the reading-style material facilitated sentence-wise comparisons. Recordings were made in a sound-treated booth and sampled at 22.05 kHz/16 bit. To identify a suitable subset of items for the current study, the second author rated the prosodic performance (rhythm and intonation) of individual participants on a scale from 1–9 (1 = excellent; 9 = very poor). This preliminary assessment was designed only to ensure that a final stimulus set with considerable variability in prosodic goodness would be obtained. The same story was read by 32 native speakers of Canadian English. For the purposes of this study we selected 41 speakers of Cantonese representing the full range of performance ratings, 20 male and 21 female, as well as 30 Canadian English speakers, 15 male and 15 female. Results presented in this paper are based on the first five sentences from the short story with a total duration of between 11 and 22 seconds. The part of the data uttered by Cantonese speakers shall henceforth be referred to as *CANT*, that of the Canadian English speakers as *CNDE*.

In the first step, all recordings were force-aligned at the syllable-level using an *HTK* [10] based system provided by Yuanfu Liao, NTUT, Taipei, trained on the *TIMIT* corpus [11]. The target of alignment was a canonical *SAMPA* transcription of the underlying text, produced with the grapheme phoneme converter of the first author's multilingual TTS system [12]. Strictly speaking, due to reductions, especially in the native speakers' utterances, not all of the identified segments corresponded to phonetic syllables. However, for the sake of the following acoustic comparison it was required that the label sequence be identical for all utterances of the same sentence. The automatic segmentations were converted to *PRAAT* TextGrid format [13] and syllabic boundaries hand-corrected. We were not interested in the identity and exact boundaries of phones actually realized, but rather in the rhythmic structure of the utterances.

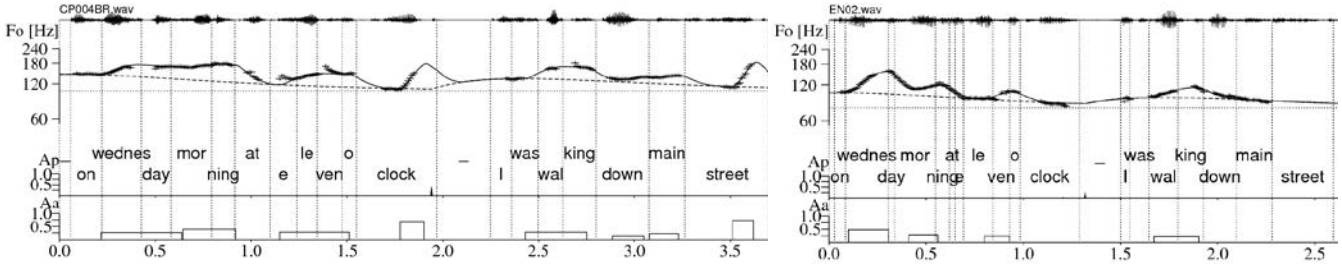


Figure 1: *Example of analysis of sentence 1, uttered by a Cantonese (left) and Canadian English speaker (right), both male: "On Wednesday morning at 11 o' clock I was walking down Main Street." From the top to the bottom: speech waveform, F0 contour (+extracted, - modeled), underlying phrase and accent commands. Vertical lines indicate boundaries of text syllables. Fb is denoted by the horizontal line underneath the F0 pattern.*

In order to analyse the intonational properties of the two corpora, *F0* values were extracted at a step of 10 ms using the *PRAAT* default pitch extraction settings and subjected to manual inspection and correction. Syllables exhibiting creaky voice were marked on the syllabic tier.

All utterances were subjected to Fujisaki model [14] parameter extraction [15] as shown in Figure 1 (sentence 1, produced by Cantonese speaker CP004 and Canadian English speaker EN02, both male). The figure displays the following, from the top to the bottom: the speech waveform, the *F0* contour (+signs: extracted, solid line: model-based), the text, the underlying phrase and accent commands. This methodology has the great advantage of providing an accurate fit of the original *F0* contour, while breaking it down into a limited number of parameters which can then be related to phonologically relevant prosodic landmarks such as accents and boundary tones. It is also very useful when resynthesizing stimuli with modified prosody, as it provides *F0* values for each point in time. The amplitudes *Aa* of the box-shaped accent commands (see Figure 1) are correlated with the amount of emphasis given to an accented syllable, wheras the magnitudes *Ap* of the impulse-wise phrase commands reflect the amount of *F0* reset at the onset of new prosodic phrase.

3. Prosodic Parameters and Results of Analysis

The objective of the analysis was to identify systematic differences between the *CANT* and *CNDE* data at the rhythmic as well as the intonational levels and relate them to the proficiency ratings.

Timing: When we examined how the number of speech pauses in an utterance influenced the expert rating, we observed a Pearson $r = .54$ ($p < .01$). The correlation is slightly less with the total duration of speech pauses in the passage ($r = .42$)

For further quantitative analysis, the syllabic labels from the *CANT* and *CNDE* corpora were compared with respect to means and standard deviations of durations, as well as rhythmic properties of the utterances. Analysis showed a considerably higher rate of 5.7 syllables/second for the *CNDE* speakers against 4.4 syllables/second for the *CANT* group. If we correlate the syllable rate of the learners with the expert rating on their prosodic performance, we obtain a Pearson *r* of -.67, suggesting fluency as a key factor in the judgment.

We next investigated whether the syllable-timed nature of Cantonese as opposed to the stress-timing of English also affected the realizations of the Cantonese speakers. Based on the text underlying the utterances we categorized all syllables as belonging to one of three classes: *Unstressed*, *stressed (verbs)* and *stressed (other content words)*. Due to the fact that verbs often become deaccented in a sentence context this classification was mainly based on the superordinate part-of-speech. Although individual realizations varied, results showed that stressed syllables in the *CNDE* corpus were relatively longer than unstressed syllables than in the *CANT* corpus. The following table shows the results of comparison.

Table 1: *Means and standard deviations of syllabic durations for unstressed and stressed syllables.*

Group	Syllable type	Mean [ms]	N
CANT	unstressed	181	1230
	stressed (verbs)	280	451
	stressed (others)	310	451
CNDE	unstressed	121	840
	stressed (verbs)	219	300
	stressed (others)	271	315

The ratio of mean durations unstressed/stressed is 0.45 for the Canadian English speakers, whereas it is 0.58 for the Cantonese speakers. The values are 0.55 versus 0.65 for the stressed syllables in verbs. This suggests that Cantonese speakers are aware of the English stress system, yet they tend to produce syllables of more uniform lengths than the Canadian English speakers.

Looking more closely at the rhythmic patterns of individual sentences we correlated the syllabic durations in one realization of a sentence with the syllabic durations in all the other realizations of the same sentence. The advantage of this approach is that the effect of the speech rate on this measure is rather small. In fact, this measure was previously used for evaluating the quality of a duration model for text-to-speech synthesis [16] and later applied to the analysis of foreign-accented English [17].

By averaging the inter-utterance correlations over all utterances of a given sentence for each group (*CANT, CNDE*) we obtain a measure of similarity within that group, as well as the mean inter-group correlation. Results indicate that the *CNDE* realizations (mean $r = 0.91$) are much more similar in their rhythmic structure (more highly correlated) than the *CANT* ones (mean $r = 0.75$). Also the cross-correlation between the two groups is only moderate (mean $r = 0.77$).

In order to test whether the observed sentence-based correlations were valid indicators of prosodic goodness, we calculated the centroid of all *CNDE* utterances for each sentence. That is, for each syllable in a given sentence we averaged over all observed instances in the *CNDE* data set, yielding prototypical syllabic durations for each sentence

("duration norm"). Subsequently we calculated the correlations between each of the Cantonese utterances and their corresponding Canadian English duration norm. Statistical analysis showed that this rhythmic correlation was significantly (r = -.479, p < .01) correlated with the original prosodic goodness rating on the 9-point scale. We have to take into account that the Cantonese data spans a range of proficiency levels so that some speakers might already have attained very high rhythmic proficiency, with others exhibiting almost "Cantonese" rhythm.

Intonation: As mentioned earlier, the extracted *F0* contours were parameterized using the Fujisaki model in order to establish the differences between the *CNDE* and *CANT* data sets. To this effect, automatic parameter extraction was performed [15]. Then the analysis results were inspected and, if necessary, corrected using the interactive *FujiParaEditor* [18]. As can be seen from the examples in Figure 1, the resulting model *F0* contours are very close copies of the original natural ones, the *RMSE* being less than 2 semi tones.

If we look at mean *F0* we find that it is generally higher for Cantonese than for Canadian English speakers, for both male (127 vs 115 Hz) and female subjects (220 vs 197 Hz). 3.5% of syllables in *CNDE* exhibit creaky voice, and 3.1% in *CANT*. This difference, however, is mostly due to the Canadian male speakers.

Some numerical results of analysis are displayed in Table 2, which shows means and standard deviations of accent command amplitude *Aa* and duration for the *CANT* and *CNDE* data. As can be seen - though mean amplitudes of *Aa* are quite similar - the *CANT* group produce longer accent commands than the Canadian English speakers. These appear as plateau-like gestures spanning several syllables at high pitch (compare Figure 1, left)

Table 2: *Mean and standard deviation of accent command amplitude Aa and accent command duration.*

group		*Aa*	duration [ms]
CNDE	mean	.28	227
	s.d.	.15	102
	N	443	
CANT	mean	.26	286
	s.d.	.15	165
	N	785	

Table 3: *Means and standard deviations of accent command amplitude Aa associated with stressed syllables in verbs and other content words.*

stress	group	N	Mean	S.D.
stressed (verbs)	*CANT*	485	.21	.15
	CNDE	307	.20	.15
stressed (others)	*CANT*	424	.23	.16
	CNDE	291	.25	.18

If we look at the frequency of accent commands there are 1.75 commands per second in the *CNDE* group, but 1.60 for the *CANT* group. The syllable-based frequency is one command every 2.7 syllables in the *CANT* group, but one command every 3.3 syllables in the *CNDE* data.

If we calculate mean *Aa* for stressed syllables in verbs as opposed to other content words we find that these values are quite similar within the Cantonese group, as the means differ by only 8%, but for the Canadian English speakers, the difference is almost 29%, that is, verbs generally receive lower prominence than other content words (see Table 3).

Table 4 shows means and standard deviations for the phrase command magnitude *Ap* indicating the amount of *F0* reset at the onset of a new phrase. Apparently, the two groups do not differ with respect to the strength of rephrasing. This result, together with that from the accent command amplitudes *Aa,* suggests that the Cantonese speakers use a quite similar *F0* range as the Canadian English speakers.

Table 4: *Mean and standard deviation of phrase command magnitude Ap.*

group	means	S. D.	N
CNDE	.29	.14	206
CANT	.29	.13	363

However, the Cantonese speakers rephrase more frequently, on the average once every 5.9 syllables compared to 7.1 syllables for the Canadian English speakers. This result is partly due to the higher speech rate of the *CNDE* group (compare Figure 1), as well as the lower fluency of the *CANT* group, who insert additional phrase boundaries.

4. Stimuli and Design of the Perception Study

Stimuli for perceptual evaluation were created by applying a number of processing steps to the original speech recordings. We selected seven *CANT* speakers, as well as two *CNDE* subjects, EN11 and EN14, to serve as "donors" of the prosodic features to be transplanted to the Cantonese-accented recordings, while utterances from four other *CNDE* subjects were added to serve as anchoring points for the listeners. Table 5 shows a list of the speakers whose utterances were chosen for prosodic manipulations, as well as some of their prosodic characteristics. It also lists the expert prosodic goodness ratings for the Cantonese speakers.

Table 5: *Utterances manipulated for the perception test. The list shows ratings of prosodic goodness, as well as means of some prosodic parameters.*

Subject	Sex	Rating	Mean syll.dur. (ms)	Corr.w. dur. norm	Mean Ap	Mean Aa	Fb [Hz]
CP015	M	7	284	.57	.17	.20	85
CP023	F	2	207	.89	.20	.42	140
CP025	F	6	212	.70	.14	.14	195
CP027	F	5	205	.91	.21	.37	150
CP028	M	1	183	.95	.16	.16	110
CP034	M	8	250	.78	.35	.36	100
CP053	M	8	266	.92	.40	.28	90
EN11	F	native	186	.96	.36	.31	140
EN14	M	native	178	.95	.31	.33	85

The following types of stimuli (see Table 6) were created:

(1) **Plain resynthesis**: In the *PR* condition, all utterances were resynthesized using the Fujisaki model-based *F0* contours, employing the *FujiParaEditor* and the *PRAAT* PSOLA resynthesis capability.
(2) **Duration modification A**: Durations of Cantonese stimuli were adjusted to match those of the Canadian English target speakers EN11 and EN14 (conditions *DA11* and *DA14*, respectively), and the Canadian English duration norm *DNE*. In order to compensate for the change in *F0* slope due to the higher target speech rate, the amplitudes of the Fujisaki model parameters for the *CANT* utterances were rescaled accordingly.
(3) **Duration and *F0* modification**: Utterances from (2) whose duration characteristics now matched those of EN11 and EN14, were further manipulated with respect to *F0*, by applying the original Fujisaki model parameters of EN11 and EN14, respectively, giving conditions *DF011* and *DF014*. The base frequency *Fb* was kept at the original value of the Cantonese speakers. Hence the resulting *F0* contour had the shape of the donor speaker's production, while the *F0* values remained within the range of the Cantonese speaker. This also facilitated the "cross-gender" modifications.
(4) **Duration modification B**: The durations of EN11 and EN14 were adjusted to those of Cantonese speakers CP015, CP034 and CP053 as in (2), giving the *DB* condition. They were also adjusted to half-way between source and target; that is EN11 and EN14 were slowed down and syllable duration ratios and pause structures of the Cantonese targets applied, but the resulting utterance duration was the average between the original Canadian English and the targeted Cantonese utterance. This approach, the *DBH* condition, was taken because further deceleration tended to render the *CNDE* utterances extremely slurred and unnatural.

Finally, in order to equalize the quality of the stimuli and tone down some artifacts of time stretching and pitch shifting, all stimuli were filtered to telephone bandwidth (340-4000 Hz) and resampled at 8k Hz.

The recordings were prepared for perceptual evaluation by separating the five sentences in all stimuli into individual tokens. These were informally prescreened by two research associates, who deemed that any distortions and artifacts in the *CANT* tokens were minimal. Several of the *CNDE* tokens in the DB condition, however, were judged to sound too unnatural for use. On the basis of the prescreening, only the fourth sentence from speakers CP015 and CP053 was selected for inclusion in both the *DB* and *DBH* conditions. The final stimulus set thus consisted of the 248 stimuli summarized in Table 6.

Table 6: *Summary of stimuli in the perception task*

Name	Description	*CANT* Tokens	*CNDE* Tokens
PR	Plain Resynthesis	35	30
DA11	Duration of EN11	35	-
DA14	Duration of EN14	35	-
DNE	Mean Duration of CNSE	35	-
DF011	Duration+F0 of EN11	35	-
DF014	Duration+F0 of EN14	35	-
DB	Duration of CP015, CP053		4
DBH	Half-way adjustment of DB		4

Twelve native speakers of Canadian English, all with training in phonetics, were recruited as judges. During prescreening it was found that the 9-point evaluation scale used by the expert rater for the original stimulus set was too large to allow satisfactory ratings of the items in the subset.We therefore refined our approach such that the judges rated each production for prosodic goodness on a 5-point scale (1 = poor, 5 = excellent) by focusing on intonation and temporal characteristics. Stimuli were blocked on sentence and presented in self-paced sessions via headphones under quiet conditions. On hearing each item, the listeners responded via a keyboard press. After a 12-item warm-up session, each token was judged twice in a randomized presentation; the order of blocks was counterbalanced across listeners. Prior to assigning each rating, listeners were allowed to replay the stimulus up to three times. Total time for the task was about 35 minutes, though rest breaks between blocks were permitted.

5. Results of the Perception Study

Mean ratings for the two presentations of each item were computed for each listener. The intraclass correlation for the resulting scores was .924 ($p < .001$), indicating a high level of inter-judge reliability. For the purposes of the following analyses, ratings were pooled over the five sentences to yield a mean score by speaker for each condition. A paired samples *t*-test revealed that the plain resynthesized (*PR*) *CNDE* tokens were rated significantly higher (M = 4.7) than the *CANT PR* (M = 2.8) productions, $t(11) = 13.49$, $p < .001$.

Figure 2 shows the mean goodness ratings on the *CANT* productions for each stimulus condition. In all manipulated conditions, the values were higher than in the *PR* condition. For the simultaneous manipulations of duration and *F0*, *DF014* and *DF011*, increases of .78 and .71, respectively, were observed, while for the duration-only manipulations, *DNE*, *DA14*, and *DA11*, the increases were smaller, with values of .17, .2, and .17, respectively. A one-way repeated measures ANOVA revealed a significant effect of condition, $F(5, 55) = 15.43$, $p < .001$. However, according to post hoc Bonferroni-adjusted *t*-tests ($p < .05$), only the *DF014* and *DF011* conditions yielded significantly higher scores. Also, scores in *DF014* and *DF011* were both significantly higher than those for the three other manipulated conditions. No other pairwise differences proved significant.

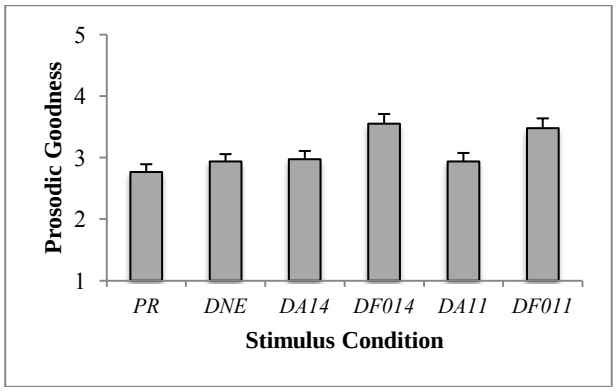


Figure 2: *Mean goodness ratings (with standard error) in the 6 conditions for the CANT productions*

A visual inspection of the ratings of individual *CANT* speakers suggested that those who were rated most poorly in the *PR* condition were the ones who benefited most from the

manipulations. Figure 3 presents mean ratings across all conditions for the speakers with the highest and lowest scores in the *PR* condition. Data for speaker CP028, who was rated almost as high as some of the native English speakers in the *PR* condition, indicate only small differences in ratings, while those for the lowest-rated speaker in *PR*, CP015, show greater variation.

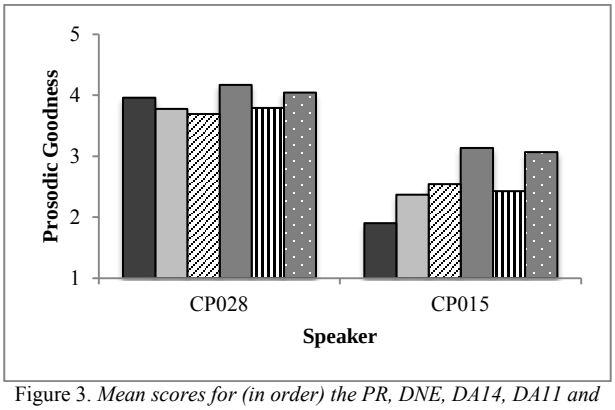


Figure 3. *Mean scores for (in order) the PR, DNE, DA14, DA11 and DF011 conditions for the speakers with the highest (CP028) and lowest (CP015) ratings in PR.*

Mean goodness ratings for the *CNDE* productions are given in Figure 4. In all cases, ratings of the manipulated stimuli were lower, with differences from the *PR* condition ranging from .93 to 1.64. A repeated measures ANOVA once again yielded a significant effect of stimulus condition, $F(4,44) = 16.86$, $p < .001$, with post hoc tests indicating that all manipulated conditions had significantly lower ratings than *PR*. However, none of the other pairwise differences were significant.

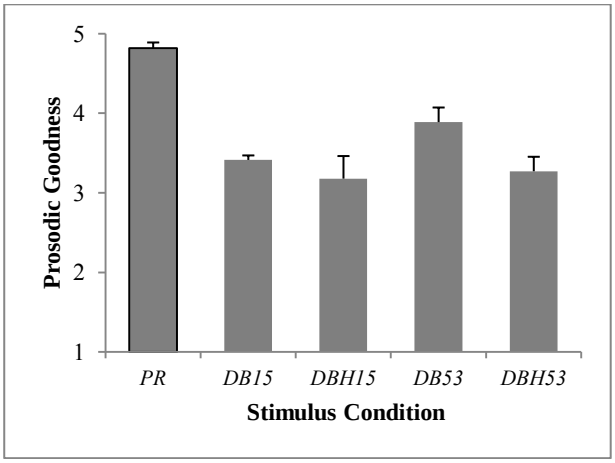


Figure 4. *Mean goodness ratings (with standard error) in the 5 conditions for the CNDE productions*

6. Discussion and Conclusions

The current study concerned the prosodic analysis of accented English speech data produced by Cantonese users of English. We found that the number of pauses in an utterance, a typical indicator of disfluency, as well as the speech rate, were correlated with expert perceptual judgments of prosodic goodness. At the rhythmic level, Cantonese learners of English produced relatively longer unaccented syllables than Canadian English speakers, which suggests that their rhythm was influenced by the syllable-timed structure of Cantonese. The syllabic durations in the Canadian English group were more uniform than within the Cantonese group expressed by the durational correlations between individual productions of the same sentence. At the intonational level, Cantonese speakers produced comparable excursions of *F0* and used a similar range of *F0* as the Canadian English controls. They placed pitch-accents more frequently and exhibited less vocal fry than their Canadian English counterparts. We found that the degree of prominence in terms of accent command amplitude, *Aa*, assigned to stressed syllables was more uniform in the Cantonese subjects, whereas it was considerably lower for the Canadian English speakers in the case of verbs. This might suggest that though the Cantonese learners were aware of the stress system, they did not replicate the information structure the same way as the native speakers.

In general, the results of the perceptual study indicated that transplanting native-like prosodic characteristics on Cantonese-accented English speech led to better judgments from phonetically sophisticated listeners. For the sample of speakers under study here, statistically significant improvement was observed for simultaneous manipulation of duration and F0, but not for duration alone, although there was a tendency toward better ratings in the duration-only manipulations. It is important to note that the degree of improvement appears to have depended on the nature of the speech prior to manipulation, with speech that was already prosodically good showing less improvement than speech that was especially poor. Expansion of the stimulus set to include more Cantonese speakers with poor prosody may uncover more subtle, but statistically significant, effects of duration manipulations alone. Also worthy of note is the fact that the effects of the manipulations did not vary according to donor speakers. Again, expansion of the study will be useful in establishing whether this is generally the case. The magnitude of improvement yielded by transposing durations and *F0* is comparable to that found in an earlier study [19] (.8 on a 5 point scale).

In the case of the duration manipulations of the native English productions, a significant decrement was seen in the ratings in all conditions. Despite a tendency for the items that were completely slowed to Cantonese-accented rates to be judged somewhat worse than those slowed to a rate intermediate between native English and Cantonese-accented, the differences between manipulated conditions were not significant. However, only a limited number of stimuli were used because of the difficulty in creating natural-sounding stimuli in the completely slowed condition. Once again, there was no indication of an effect of donor speaker.

Future work will entail perceptual experiments with segmentally and prosodically manipulated stimuli in order to identify the factors that contribute most to the percepts of strong foreign accent and reduced intelligibility. In addition, the effects of the manipulations on phonetically untrained listeners should be evaluated to determine whether comparable patterns of perception are observed. In future work, we plan to evaluate the applicability of our findings for computer-aided pronunciation training. Work by Pfitzinger et al. has shown that re-synthesizing speech in the learner's voice, but with corrected prosody can assist in training compound accents in German, for instance [20]. A variety of common language laboratory exercises could potentially make use of resynthesized utterances, including discrimination and

identification tasks, and shadowing. The use of the learners' own speech in such tasks may prove especially beneficial in addressing individual speakers' pronunciation needs.

7. Acknowledgements

This work was supported by DFG international collaboration grant no. Mi 625-17/1 funding Mixdorff's stay at Simon Fraser University. We thank Gloria Mellesmoen for assistance with stimulus preparation and perceptual data collection. Thanks also go to Yuanfu Liao for providing the force-alignment segmentations.

8. References

[1] Anderson-Hsieh, J., Johnson, R. and Koehler, K. "The relationship between native speakers judgements of nonnative pronunciation and deviance in segmentals, prosody and syllable structure", Language Learning, 42: 529-555, 1992.

[2] Magen, H.S., "The perception of foreign-accented speech", Journal of Phonetics, 26: 381-400, 1998.

[3] Munro, M.J, & Derwing, T.M. "The functional load principle in ESL pronunciation instruction: An exploratory study", System, 34: 520-531, 2006.

[4] Munro, M.J., Derwing, T.M., & Burgess, C.S. "Detection of nonnative speaker status from content-masked speech", Speech Communication, 52: 626-637, 2010.

[5] Celce-Murcia, M., Brinton, D., Goodwin, J. & Griner, B. Teaching Pronunciation: A Coursebook and Reference Guide. New York: Cambridge University Press, 2010.

[6] Hahn, L. "Primary stress and intelligibility: Research to motivate the teaching of suprasegmentals", TESOL Quarterly, 38: 201-223, 2004.

[7] Derwing, T. M., Munro, M. J., & Wiebe, G. E. "Evidence in favor of a broad framework for pronunciation instruction", Language Learning, 48: 393-410, 1998.

[8] Tajima, K., Port, R. & Dalby, J. "Effects of temporal correction on intelligibility of foreign accented English", Journal of Phonetics, 25: 1-24, 1997.

[9] Mellgren, L., & Walker, M. New Horizons 4. Reading, Mass.: Addison-Wesley Publishing Company. (p.106), 1973.

[10] Pye, D.,Woodland, P. and Young S. "Large Vocabulary Multilingual Speech Recognition using HTK." Proc. Eurospeech, Madrid, 1995.

[11] Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallett, D., and Dahlgren, N., "DARPA TIMIT acoustic-phonetic continuous speech corpus CDROM," NIST Order No. PB91-505065, National Institute of Standards and Technology, Gaithersburg, MD, 1990.

[12] Hilbert, A., and Mixdorff, H.,"Weiterentwicklung eines Sprachsynthesesystems", in G. Görlitz [Ed.], Nachhaltige Forschung in Wachstumsbereichen Band I, Logos Verlag, Berlin, 35-42, 2011.

[13] Boersma, P. "Praat, a system for doing phonetics by computer", Glot International 5(9/10): 341-345, 2001.

[14] Fujisaki, H. and Hirose, K., "Analysis of voice fundamental frequency contours for declarative sentences of Japanese", Journal of the Acoustical Society of Japan (E) 5(4): 233-241, 1984.

[15] Mixdorff H., "A novel approach to the fully automatic extraction of Fujisaki model parameters", Proceedings of ICASSP 2000, vol. 3, 1281-1284, Istanbul Turkey, 2000.

[16] Mixdorff H. and Jokisch, O., "Evaluating the quality of an integrated model of German prosody", International Journal of Speech Technology 6(1): 45-55, 2003.

[17] Mixdorff, H. and Ingram, J, "Prosodic Analysis of Foreign-Accented English", In Proceedings of Interspeech 2009, Brighton, England, 2009.

[18] http://public.beuth-hochschule.de/~mixdorff/thesis/fujisaki.html

[19] Ingram, J., Mixdorff, H. and Kwon, N., "Voice morphing and the manipulation of intra-speaker and cross-speaker phonetic variation to create foreign accent continua: A perceptual study." In Proceedings of SLATE 2009, Wroxall Abbey, England, 2009.

[20] Bissiri, M. P.; Pfitzinger, H. R. (2009). Italian speakers learn lexical stress of German morphologically complex words. *Speech Communication*, 51(10): S. 933–947.

Prosodic Chunking of German as a Foreign Language

Hansjörg Mixdorff, Hamurabi Gamboa Rosales

[1] Department of Computer Science, Beuth University Berlin, Germany
[2] Unidad Académica de Ingeniería Eléctrica, Universidad Autónoma de Zacatecas, Mexico
mixdorff@beuth-hochschule.de; hamurabigr@hotmail.com

Abstract

This study concerns the perception of boundaries and accented syllables by native German subjects as compared to foreign non-speakers and learners of the language at different proficiency levels. To this effect six-syllable sequences excised from a context of three poly-syllabic words of German were presented to participants who had to select the syllables they perceived as accented, as well as the locations of word boundaries. Results show that German native subjects perform well at the word boundary task, but mark correctly less than two thirds of accented syllables. Chinese and Mexican non-learners still detect a considerable number of word boundaries and accented syllables. Learners of German show improvement at the task with growing experience though they often pick legal subword units that do not necessarily form a plausible sequence. Correlation analysis of factors for syllable and boundary selection performed for non-learners and German subjects – as expected – shows considerably different behaviours. Whereas the boundary location does not influence the Germans' decision on the accent location, Chinese and Mexican non-learners show a preference to mark an accent when the syllable is followed by a word boundary. We also found that the acoustic properties of the syllables had a larger impact on the non-learners' decisions since they could not operate on linguistic knowledge of German.

Index Terms: Prominence, accent and boundary perception, L2 learning

1. Introduction

It is a well-known fact that non-speakers and native speakers of a language perceive and process stimuli from that language quite differently. When a person studies a foreign language (L2) (s)he makes a transition between the two states as (s)he acquires a growing competence as to L2 linguistic structures and vocabulary. An important competence for communication is the ability to process and chunk the speech stream into meaningful units. The segmentation of an utterance into words depends heavily on prosodic cues (see, for instance, [1]) such as F0 and duration – features which might be employed quite differently in the L2 – as well as the growing L2 lexicon on the part of the learner.

Lexical stress is a property of each poly-syllabic word and in German – in contrast to certain other languages – its location in the word is rather flexible, i.e. not predefined by default [2]. Hence three-syllable words, for instance, can exhibit lexical stress either on the first, second or third syllable. There exists a small group of words which are segmentally identical, but differ as to the lexical stress location (set in bold face): compare, for instance, 'um-**fah**-ren' (*to go around*) vs. '**um**-fah-ren' (*to run over*). In the context of an utterance the lexically stressed syllables become potential loci of accentuation, usually associated with prosodic cues such as F0 transitions and lengthening. Therefore the learner is required to memorize this feature for each word and decode it from the speech stream.

This study elaborates on an initial experiment reported in [3], which concerned the perception of boundaries and accented syllables by native German subjects as compared to Chinese non-speakers and learners of the language at different proficiency levels. We aimed to investigate how subjects performed on a task for which they either had to rely – to varying degrees – on their linguistic knowledge as well as the acoustic properties of the stimuli.

Furthermore, we explored the perceptual interrelationship between boundaries and accented syllables. This work was originally inspired by Gilbert et al. [5] who showed in a learning experiment that speech chunking in French is performed in rhythm groups. However, whereas French words exhibits a default prominence on the ultimate syllable of a word, next to the word boundary, the situation in German as explained above is quite different.

The current study expands the work in [3] three-fold: (1) We perform an acoustic prosodic analysis of the stimuli and examine the dependency of listener judgements on several acoustic parameters, such as *F0*, syllable duration and intensity, (2) incorporate results from a different language group, namely native speakers of Mexican Spanish with or without knowledge of the German language, (3) reevaluate our results with respect to subword units which can be legal words of German. As [3] to this date is still in press we will first present the experiment in greater detail.

2. Experiment Stimuli and Design

The stimuli employed in this study are all six-syllable sequences excised from the context of a sentence with the structure "Menschen können A, B, C sein", *("People* can *be A, B, C.")* where A, B and C are possible characteristics of people, either adjectives or past participles. These real words possess either two or three syllables, with the lexical accent on the first, second or third syllable, respectively. For each of the conditions we selected 10 unique words none of which contained syllables of the others. From this set of words we constructed groups of three words A, B and C, where A and C were always three syllables long, and B either contained two or three syllables. Here is an example: "Menschen können be-**le**-sen, **schlag**-fer-tig, **lang**-wei-lig sein." (*"People can be erudite, quick-witted, boring".*) Figure 1 displays a stimulus example demarked by grey vertical lines inside the surrounding carrier phrase, syllable boundaries are indicated by dotted vertical lines. By combining three words we yielded groups with all possible positions of the lexical accents, including conditions off accent clash.

We intentionally omitted the conjunction 'and' before the third item in order not to supply a morphemic marker of the boundary. The stimuli were then constructed in such a way, that only the B word was completely preserved in the stimulus whereas only parts of word A and C were present.

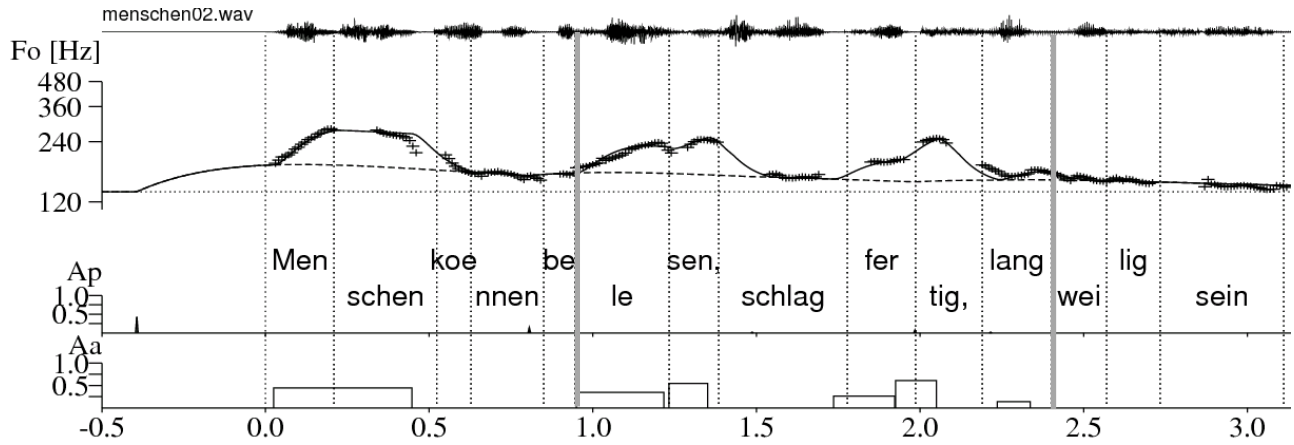


Figure 1*: Example of carrier sentence, from which the stimuli were excised. Stimulus boundaries indicated by thick grey lines.*

TEIL 1	PLAY BUTTON	TIMES PLAYED	SYLL. 1	BND	SYLL. 2	BND	SYLL. 3	BND	SYLL. 4	BND	SYLL. 5	BND	SYLL. 6	WORD IDENTIFIED	WORD IDENTIFIED
	PLAY		LISH ☐	☒	PO ☐	☐	PU ☐	☐	LER ☒	☒	BE ☐	☐	LE ☒	POPULÄR	
	PLAY		ER ☐	☐	KANT ☐	☒	SHLAK ☒	☐	FER ☐	☐	TICH ☐	☒	RE ☐	SCHLAGFERTIG	
1.	PLAY		TER ☐	☐	KYLT ☐	☐	Y ☐	☐	BER ☐	☐	LEKT ☐	☐	UN ☐		
2.	PLAY		BE ☐	☐	FOL ☐	☐	ENT ☐	☐	SHI ☐	☐	DEN ☐	☐	GLAUP ☐		
3.	PLAY		BER ☐	☐	LEKT ☐	☐	FO ☐	☐	TO ☐	☐	GEN ☐	☐	KOM ☐		
4.	PLAY		DEN ☐	☐	UN ☐	☐	TER ☐	☐	KYLT ☐	☐	FO ☐	☐	TO ☐		
5.	PLAY		LE ☐	☐	GEN ☐	☐	GE ☐	☐	RI ☐	☐	SEN ☐	☐	GROS ☐		
6.	PLAY		ZEN ☐	☐	SHLAK ☐	☐	FER ☐	☐	TICH ☐	☐	LANG ☐	☐	VAI ☐		
7.	PLAY		FA ☐	☐	REN ☐	☐	FER ☐	☐	LE ☐	☐	GEN ☐	☐	PRAG ☐		

Figure 2: *Format of the experiment section of the questionnaire.*

A stimulus created from the sample sentence could hence be "…**le**sen, **sch**lagfertig, **lang…**", with the lexical accent syllables set in bold face.The stimuli were recorded by one male and one female native speaker of German and the target sequences excised using *PRAAT* [1] after performing peak scaling to 98%.

In order for the syllables to be accessible for people without knowledge of German, we used a pseudo-transcription based on German SAMPA in which all non-letter symbols where replaced by other letters or deleted (in the case of the lengthening symbol ":"): LE-ZEN-SHLAK-FER-TICH-LANG. The stimuli, a total of 106 tokens, were randomised and an MS-Word-based questionnaire was developed which contained a list of the stimuli with PLAY links to the audio files.

We checked the frequency of the target words and found only 13 of the 30 trisyllabic words in a list of the 10000 most frequent German words [6], and seven of the 20 disyllabic ones. However, 15 of these words are ranked in the last third of the list. Hence, we expected most of them to be unknown to students of German at the beginner's and even at the intermediate level.

In the header of the questionnaire we inquired about some personal details such as age and gender, learning history of German, time spent in Germany as well as languages spoken at home. Then the experiment design and its aims were explained, namely, to determine the boundaries between words of German as well as the accented ("strong") syllables of these words which had been cut out of context. Participants were advised to play stimuli a maximum number of four times and note down any words they might have perceived. They were also informed that each stimulus contained at least one accented syllable and at least two boundaries.

The pseudo-transcriptions of the syllables were listed with checkboxes that enabled the selection of each of the syllables/boundaries as shown in the excerpt of the questionnaire displayed in Figure 2 with two illustrative examples in grey and the first seven trials.

Subjects were then asked to work through the examples one by one and make their choices based on their perception. They were also asked to note down how often they played the stimuli and whether they had identified any real words in the sequence. The completed questionnaires were saved in Word format by the participants and later on converted to RTF format for further evaluation of results.

3. Measurements of Acoustic Parameters

All sentences from which the stimuli were excised were segmented manually on the syllable level inside the *PRAAT* TextGrid, yielding syllabic durations for all stimulus utterances. Means and standard deviations of syllable durations were 245ms/99ms for the male subject and 230ms/93ms for the female.

F0 contours were extracted at a step of 10ms using the *PRAAT* default pitch extraction settings and subjected to manual inspection and correction.

All utterances were subjected to Fujisaki model [7] parameter extraction [8], see example in Figure 1. The figure displays from the top to the bottom: The speech wave form,

the *F0* contour (+signs: extracted, solid line: model-based), the text, the underlying phrase and accent commands. We employed the accent command amplitudes for quantifying the interval of *F0* transitions occurring in each of the stimulus syllables.

Intensity contours were extracted in *PRAAT* with default settings, and mean intensities in dB, as well as maxima employing parabolic interpolation were determined for each syllable.

4. Results of Analysis

From [3] we had results from eight native German listeners, eight Chinese (CN) learners after their first year, 15 Chinese learners after their second year, three Chinese students after year 5 and eleven Chinese non-speakers of the language.

For the current paper we expanded the German group by two more subjects, as well as added 21 Mexican native speakers of Spanish (MX), 12 without knowledge of German, five after year 1 and 4 after year 3 of German classes.

We determined how many of the intended items (boundaries and accented syllables) had been selected by the participants. This measure, however, does not reflect additional selections that were erroneous, that is, insertions.

We therefore calculated an error score, in analogy to ASR performance evaluations by defining the error as follows:

error = 100 x (insertions + deletions)/total number of tokens

If we assume that two prominences and two boundaries are randomly selected on each trial, chance level for the current data set would be at 176.7 for the prominences and 175 for the boundaries.

Table 1: *Percentage correct and insertion and deletion errors.*

group	percentage correct bound. mean/s.d.	percentage correct prom. mean/s.d.	error bound. mean/s.d.	error prom. mean/s.d.	N Subj. (m/f)
Germans	96.0/5.6	63.5/15.0	6.4/8.4	56.7/17.3	4/6
CN non-learners	69.1/10.2	46.0/13.5	64.3/19.4	96.2/17.3	3/8
CN after year 1	80.1/13.1	38.6/22.2	38.1/28.4	81.0/9.9	4/4
CN after year 2	88.8/7.1	35.1/12.8	20.8/8.8	84.0/11.6	9/6
CN after year 5	96.2/2.4	78.9/21.3	4.4/2.7	27.0/24.1	0/3
MX non-learners	70.3/20.3	37.9/14.6	62.3/38.8	91.4/11.8	7/5
MX after year 1	75.2/15.5	30.9/10.0	39.4/17.3	101.5/13.2	3/2
MX after year 3	88.6/9.2	45.6/13.4	25.8/10.8	84.8/21.7	2/2

As can be expected (see Table 1), native speakers fare well at identifying word boundaries (96.0% correct) whereas the rate is considerably lower for accented syllables (63.5%). Examination of individual trials shows that there are many cases when even the identification of the word boundaries does not facilitate selection of the accented syllable. The Chinese 5th year students achieve even better results than most of the German natives (96.2%/78.9% correct for boundaries and accents) whereas the ratings are still 69.1% and 46.0% for the Chinese non-speakers of German and quite similar for the Mexican non-speakers. The reason why the Chinese 5th year surpassed the German native subjects might be that they memorized the accent location for each lexical entry whereas German speakers acquire it with their native language and might not always be aware of the concept of word accent.

CN students on lower levels, that is, after year 1 and 2 improve with respect to word boundaries to 80.1% and 88.8%, respectively. This seems to suggest that their vocabulary and knowledge of word structures expands. However, their performance regarding the accented syllables (38.6% and 35.1%) seems to even deteriorate when compared to the non-learners. This appears like a rather unexpected outcome which can be partly explained by differences in the strategy employed by the subjects. When we divide the mean numbers of insertions by the total number of expected accented syllables, the result is .38 for CN non-learners, and only .19 for CN students after year 2, for instance. In contrast, the number of omissions divided by the total number of accented syllables is .51 for non-learners and .65 after year 2. This means that in general non-learners marked more syllables as accented than the students of German who apparently operated more conservatively and therefore missed a considerable number of accented syllables. We can only speculate about the reasons for this behaviour. It might be the case, that the hint in the instructions mentioning *at least* one accented syllable in each stimulus influenced the judgments of the learners.

The results of correctness stated so far are based on the rigorous assumption that all original two- and three syllable words from which the stimuli were taken were identified correctly. However, since the left and right words were most often incomplete, these truncations sometimes produced subword units that corresponded to legal German words. In addition also the central word often lends itself to further segmentation, compare, for instance, "**lie**bevoll" (*affectionate*) and "**Lie**be **voll**" (*love full*), though these sequences are not necessarily plausible and require a different assignment of stress. However, for learners at the initial stage these shorter words are probably much easier to recognize than the rather infrequent larger ones. In some cases, subword chunking, however, did not produce meaningful results. The word "ver**le**gen" (*embarrassed*) was sometimes subdivided into "ver" and "**le**gen" where only the right part bears a meaning of it its own (*to lay*). We also observed cases where the compound accent syllable was missing in the stimulus, compare "**groß**zügig" – *generous,* and "groß|**zü**gig" (*speedy*)", and subsequently the secondary stress of the original word became the primary one of the resulting new item.

In order to assess the effect of subword chunking we related the number of insertions on the boundary and accent level to the number of additional boundaries and accents due to the potential subword units. The results for the different groups are listed in Table 2. Here we see a clear difference between non-learners and learners of German in that many more insertions of the learners create legal subwords. Another interesting result is that the vast majority of all boundary

insertions by the Mexican learners can be attributed to these subword units. With respect to non-words created by additional word boundaries we see that the proportion of insertions explained is quite similar for all groups, except for the advanced Chinese students who are already almost perfect. However, the ratio of these insertions compared to the total number of additional boundaries due to subwords is only 24.0% for Chinese and 20.9% for Mexican learners after year 1, for instance, but 35.8% for the Chinese non-learners. This reflects the greater phonological competence of the learners.

Table 2: *Percentage of insertions explained by subword units.*

group	% corr. bound. subword	% bound. creating non-words	% corr. prom. subword
Germans	70.3	16.7	58.7
CN non-learners	48.3	18.0	30.0
CN after year 1	72.4	17.3	46.2
CN after year 2	73.2	17.0	49.8
CN after year 5	66.7	33.3	93.0
MX non-learners	50.6	17.4	30.9
MX after year 1	80.8	16.9	25.2
MX after year 3	81.4	12.2	44.6

We examined which factors triggered the selection of an item and compared Germans with Chinese and Mexican non-learners because these two groups are maximally different with respect to the kind of information they can draw on to access the stimuli. To this effect we calculated the ratio for each item – either prominent syllable or boundary - to be selected by the three groups.

In the following analysis we do not yet consider any acoustic measurements, only the structure of the stimuli. Each syllable was classified regarding the following features: (1) lexical stress, (2) vowel length, (3) lexical stress on left/right neighbour, (4) boundary type left/right, (5) number of phones in onset/coda. Boundaries were classified regarding the following features: (1) word boundary, (2) lexical stress on syllable to the left/right. Table 5 shows the results of correlation analysis between the above-mentioned features and the ratio at which a syllable or boundary was selected. Comparison of figures shows that both groups show similar tendencies. The fact that vowel length is correlated with the ratio at which a syllable was marked as accented is probably as much due to the structural properties of German as much as due to vowel length being a prominence-lending acoustic feature. In the set of words we selected for this study 68% exhibit lexically stressed syllables with long vowels. An important difference between Germans and non-learners is the effect of adjacent word boundaries on the tendency for a syllable to be perceived as accented. We find negative correlation with the left syllable boundary and positive correlation with the right syllable boundary which is considerably stronger for the non-learners. This means a stronger preference for word-final syllables to be perceived as accented than for word-initial ones. In contrast, the effect of the neighbouring syllables' stress status appears stronger in the German subjects, indicating a preference for avoiding stress clash. We also calculated a regression model based on the factors chosen and find that it explains 87% of the variance for the German listeners, but only 68% for the non-learners.

With respect to the boundaries, the status of the boundary as either being an inter- or intra-word boundary clearly guides the decision whereas the stress status of the adjacent syllables is irrelevant to the judgments of the Germans. In contrast, a stressed syllable to the left of the boundary still has a significant effect on the judgments of the non-learners. This interrelationship matches the result for the accented syllables.

In terms of the learning effect during the experiment we compared the correctness of the results on the first and the second half of stimuli, but did not find any significant differences. We also examined the relationship between the judgments of the subjects and the acoustic properties of the stimuli. First of all we looked at how the acoustic correlates varied with the status of a syllable regarding stress and its position with respect to a word boundary. Table 3 lists some of the results for syllabic duration, accent command amplitude *Aa*, as well as mean and maximum intensity in *dB*. As expected, stressed and pre-boundary syllables are longer and exhibit larger *F0* excursions as reflected by *Aa.* The values calculated from the *PRAAT* intensity contours reflect only slight differences.

Table 3: *Means and standard deviations of some prosodic parameters with respect to the status of the underlying syllable.*

Syllable Status	Syllable Duration [ms]	Accent command amplitude *Aa*	Mean intensity [dB]	Max Intensity [dB]
Stressed	313/82	.30/.20	70.8/4.6	79.6/3.8
Unstressed	193/74	.28/.20	70.2/4.7	78.1/3.9
Pre-boundary	301/92	.40/.14	70.4/4.4	78.5/4.1
Post-boundary	210/86	.21/.22	69.6/4.7	78.4/3.9

Table 4: *Correlations (Pearson's r) between the ratio of a syllable being selectedas accented and its underlying prosodic features.*

group	syllable duration [ms]	accent command amplitude *Aa*	mean intensity [dB]	max intensity [dB]
Germans	.61**	.10*	.06 (n.s.)	.17**
CN non-learners	.54**	.29**	.08*	.18**
MX non-learners	.60**	.30**	.13**	.28**

Table 5: *Factors facilitating the selection of accented syllables and boundaries for Germans as well as Chinese and Mexican non-learners..*

Accented Syllables				Boundaries			
feature	corr. r with ratio selected (Germans)	corr. r with ratio selected (CN non-learners)	corr. r with ratio selected (MX non-learners)	feature	corr. r with ratio selected (Germans)	corr. r with ratio selected (CN non-learners)	corr. r with ratio selected (MX non-learners)
lexical stress	.76**	.46**	.56**	inter-word boundary	.97**	.72**	.82**
long vowel	.53**	.38**	.40**	stress left	.03(n.s)	.12**	.12**.
stress left	-. 34**	-.25**	-.29**	stress right	.01(n.s.)	.00 (n.s.)	-.04 (n.s.)
stress right	-.31**	-.17**	-.18**				
boundary left	-.13**	-.30**	-.23**				
boundary right	.11*	.36**	.31**				
n onset	.15**	.13**	.10**				
n coda	.21**	.24**	.30**				
variance.explained	.87	.68		variance explained	.97	.74	.83

Subsequently we correlated the ratio selected for prominent syllables with their associated prosodic features. The result is displayed in Table 4. As can be seen, syllabic duration is the strongest cue for non-learners, followed by *F0* and maximum intensity of the syllable. It is interesting that the Mexicans seem to respond more strongly to intensity than the Chinese. A regression model based on the three factors duration, *Aa* and max intensity explains 58.2% of the variance for the Chinese learners and 65.1% for the Mexicans.

5. Discussion and Conclusions

This study examined the performance of Chinese and Mexican learners and non-learners of German, as well as German native subjects on a boundary and accent syllable identification task. Our results are still preliminary due to the relatively small number of subjects.

Boundary identification was much more reliable than accent identification for all groups with even the non-learners reaching around 69% correct. Chinese learners on an advanced level even outperformed naïve German listeners. We analysed the difference between the judgments of the German subjects and Mexican as well as Chinese non-learners and found – among other results – in the latter a preference to mark syllables adjacent to a word boundary as accented. We also observed that many choices were valid with respect to legal subwords of German and therefore reduced the error rates compared to our earlier study. Finally, we examined the acoustic properties of prominent syllables selected by the non-learners and determined the factors which guided theirs judgment. Results seem to indicate that syllabic duration was the strongest cue, followed by *F0* and intensity.

Future work will concern the testing of larger and more homogeneous student groups on various levels and with other L1. We will also perform a more detailed analysis of the words that were recognized and see whether recognition at some point during the experiment changes the behaviour of the participants. Finally we will examine the acoustic properties of the word boundaries to see their impact on the learners' judgements. In conclusion we think that the chunking paradigm is a useful tool to investigate the perceptual competences of L2 learners at various stages.

5. Acknowledgements

Many thanks go to Hongwei Ding, Tongji University, Shanghai, China, for collecting the data of Chinese subjects. Thanks also to Angelika Hönemann for help with questionnaires and some of the data analysis.

6. References

[1] Christophe, A., Gout, A., Peperkamp, S., & Morgan, J. (2003). Discovering words in the continuous speech stream: the role of prosody. *Journal of Phonetics, 31*(3-4), 585–598.

[2] Kohler, K. (1977) *Einführung in die Phonetik des Deutschen*. Berlin: Erich Schmidt.

[3] Mixdorff, H., Hönemann, A. and Ding, H. (forthcoming). Perception of phrase boundaries and prominent syllables in German. In Eva Liina Asu & Pärtel Lippus (Eds.), *Nordic Prosody. Proceedings of the XIth conference, Tartu 2012* (pp. 245-254). Frankfurt am Main: Peter Lang.

[4] Boersma, P., & Weenink, D. (2012). *Praat: doing phonetics by computer* (Version 5.1.26) [Computer program]. Retrieved April 4, 2012, from <http://www.praat.org/>.

[5] Gilbert, A. C., & Boucher, V. J. & Jemel, B. (2011). The role of rhythmic chunking in speech: Synthesis of findings and evidence from statistical learning. In W.-S. Lee & E. Zee (Eds.), *Proceedings of the 17th International Congress of Phonetic Sciences* (pp. 747–750). Hong Kong: Department of Chinese, Translation and Linguistics, City University of Hong Kong.

[6] http://wortschatz.uni-leipzig.de/Papers/top10000de.txt, retrieved on 14 April 2013.

[7] Fujisaki, H. and Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. Journal of the Acoustical Society of Japan (E) 5(4): 233-241.
[8] Mixdorff H. (2000). A novel approach to the fully automatic extraction of Fujisaki model parameters. Proceedings of ICASSP 2000, vol. 3, 1281-1284, Istanbul Turkey.

Applying Rhythm Metrics to Non-native Spontaneous Speech

Catherine Lai[1], *Keelan Evanini*[2], *Klaus Zechner*[2]

[1]School of Informatics, University of Edinburgh, Scotland, UK
[2]Educational Testing Service, Princeton, NJ, USA
clai@inf.ed.ac.uk, kevanini@ets.org, kzechner@ets.org

Abstract

This study investigates a variety of rhythm metrics on two corpora of non-native spontaneous speech and compares the non-native distributions to values from a corpus of native speech. Several of the metrics are shown to differentiate well between native and non-native speakers and to also have moderate correlations with English proficiency scores that were assigned to the non-native speech. The metric that had the highest correlation with English proficiency scores (apart from speaking rate) was rPVIsyl (the raw Pairwise Variability Index for syllables), with $r = -0.43$.

Index Terms: Rhythm metrics, non-native speech, fluency

1. Introduction

Various studies have investigated metrics for quantifying rhythmic differences between languages based on properties of segmental durations [1, 2, 3]. These metrics present summary statistics about the variation of consonantal and vocalic durations throughout utterances. For example, [1] consider the percentage of the speech that was vocalic (%V) versus the standard deviation of consonantal intervals (ΔC, similarly ΔV for vocalic segments) in a sentence, while [2] employ a Pairwise Variability Index to quantify differences in adjacent intervals. [3] propose normalization of ΔC and ΔV by dividing this measure by the mean durations of consonantal and vocalic intervals respectively (VarcoC, VarcoV).

The main motivation behind the development of these measures has been to investigate the idea that languages fall into discrete rhythm classes (i.e. syllable- vs. stress-timed). However, experimental evidence has cast doubt on the idea that cross-linguistic differences in these measures represent categorical 'rhythm class' differences [4, 5]. In general, it seems that languages may vary on several rhythmic dimensions. [4] examine how well these metrics discriminate five different languages using a large corpus of read data. They find that, while pairs of languages can be discriminated, no pair of metrics separates all pairs of languages. Similarly, [5] find speech rate to be a dominant factor in the perception of language differences, but still find that listeners can discriminate the original language of utterances with identical segmental and intonational content even when speech rate is controlled for.

In addition to categorizing speech from different languages with respect to their rhytmic properties, these types of rhythm metrics have also been used to measure the rhythmic closeness of non-native speech to a given target language. Recent studies have found some of these rhythm metrics, especially ones measuring vocalic durations, to be useful in characterizing differences between native and non-native speech. For example, [6] found VarcoV and %V to be the most discriminating with respect to L1 and L2 (English-Spanish, English-Dutch). Moreover, they found that the VarcoV of English from native Spanish speakers had an intermediate value between the values of native English and Spanish speech (and similarly for English speakers of L2 Spanish). This suggests that these metrics can be used to quantify the effect of L1 on non-native productions. [7] investigated several different rhythm metrics on Cantonese-accented English and Mandarin-accented English and found that some of the metrics grouped the non-native English speakers with English speakers whereas other metrics grouped them with Mandarin and Cantonese speakers. [8] attempted to discriminate L1 British English speakers and two classes of L2 English (L1 French) speakers using scores from all metrics. They also suggest that %V and VarcoC give the best discrimination between the L1 classes. Their best reported SVM-based classification score used %V, ΔV, VarcoV and nPVI-V and achieved a 67% accuracy rate. Finally, [9] defined a new rhythm measure, the Pairwise Variability Error, and used it, along with several standard rhythm measures, to classify native English speech and Japanese-accented English. They found that the Pairwise Variability Error was the single best-performing rhythm measure, with a classification accuracy of 69.4%.

In addition to studies that have discriminated between native and non-native speech using rhythm measures, some previous studies have also used rhythm measures to score the proficiency of non-native speech. [10] used several standard rhythm measures (in combination with additional features) to predict English proficiency scores for a set of Korean English learners. [11] studied the differences between Spanish-accented English and native English with regard to phrasal prominence. They defined a rhythm measure based on the differences in mean vowel durations for syllables with primary stress and secondary stress and found that this measure correlated with phrasal prominence scores at a rate of 0.683. Finally, [12] studied read aloud English produced by native Mandarin speakers and found that rhythm metrics improved the prediction accuracy of holistic English proficiency scores when they were added to regression models built on features assessing fluency, pronunciation, and reading accuracy. The vocalic measures had the best correlations with human scores; furthermore, the correlations were negative, which indicates that a higher relative percentage of vocalic intervals in the non-native speech led to lower scores.

The methodology of using rhythm features to evaluate a non-native speaker's speaking proficiency in this study is similar to the approaches taken in these previous studies. However, most previous studies were based on restricted speech (read aloud or repeat aloud tasks) and only included non-native speakers from a single language background. In the current work we examine what these measures can tell us about spontaneous speech from speakers representing a large range of L1 backgrounds. This study will thus provide more direct evidence of

the usefulness of rhythm measures for assessing a non-native speaker's communicative competence in English, since a more naturalistic speaking task is examined. Furthermore, the inclusion of speakers from many different L1 backgrounds in this study means that conclusions about more general test taker populations can be drawn.

This paper is organized as follows: first, Section 2 describes the data sets and the methodology used for extracting the rhythm metrics; Section 3 presents the results of the experiments as follows: Section 3.1 shows how the individual rhythm metrics are correlated with proficiency ratings in non-native speech, Section 3.2 compares the rhythm metrics for non-native speakers to native speakers, Section 3.3 investigates the robustness of the syllable-level rhythm metrics by comparing two different non-native corpora to native speech, and Section 3.4 examines how the metrics vary based on the L1 of non-native speakers; finally, Section 4 summarizes the main contributions of this paper and discusses directions for future research.

2. Data and Methodology

In this study, we examined spoken responses from three data sets related to the TOEFL iBT assessment, an international assessment of English proficiency. The non-native speech sets were drawn from two sources: 1) responses to the TOEFL Practice Online assessment (henceforth TPO) and 2) responses to the TOEFL Academic Speaking Test (henceforth TAST). The native speech was drawn from a study in which native English speakers responded to TOEFL test questions in a laboratory setting (henceforth TOEFL-NS). In all of these data sets, the speakers responded to open-ended prompts that elicited spontaneous speech on a variety of topics. All of the responses are either 45 or 60 seconds in duration. Table 1 summarizes the sizes of these three data sets by listing the number of responses and speakers contained in each, as well as the number of different L1s represented (for the two non-native data sets).

Data set	# Responses	# Speakers	# L1s
TPO	1019	239	50
TAST	87	60	23
TOEFL-NS	182	34	N/A

Table 1: *Summary of the three data sets used in the study*

The TPO responses were each subsequently provided with holistic scores of English proficiency by two independent, expert raters using scoring rubrics that reflected several aspects of a speaker's speaking proficiency including delivery (fluency, pronunciation, intonation, etc.), language use (vocabulary, grammatical accuracy, etc.), and content appropriateness. The raters gave each response a proficiency score on a scale of 1 - 4, with 4 indicating the highest proficiency level.

Table 2 summarizes the rhythm metrics that are investigated in this section. The metrics were calculated over consonantal (C), vocalic (V) and syllabic (Syl) intervals except for speech rate, which is defined only in terms of syllables. In addition, we use the proportion of utterance medial silence as a feature (%sp); this metric is equivalent to the inverse of the proportion of syllabic intervals in a response (i.e., %sp = 1 - %syl).

Phone boundaries were derived automatically using the Penn Phonetics Lab Forced Aligner [13] on manual transcriptions of the spoken responses. The phones were grouped into syllables using a rule-based, onset maximization approach.[1] Disfluencies (such as filled pauses) were not removed from the data sets before the calculation of rhythm measures.

Metric	Description
ΔX	Standard deviation of X intervals.
%X	Percentage of X speech.
VarcoX	ΔX $\times$ 100/ mean(X)
nPVI-X	Normalized Pairwise Variability Index: $100 \times \sum_{k=1}^{n-1} \lvert x_{k+1} - x_k/(x_{k+1} + x_k/2)\rvert/n-1$
rPVI-X	raw PVI. $\sum_{k=1}^{n-1} \lvert x_{k+1} - x_k\rvert/n-1$
srate	Syllables per second.

Table 2: *Summary of rhythm metrics. We calculate these measures over V=Vocalic and C=consonantal intervals, as well as Syl=syllables.*

3. Results

3.1. Rhythm Metrics and Proficiency Scores

Metric	r1	r2
ΔV	-0.22	-0.20
ΔC	-0.18	-0.17
ΔSyl	-0.27	-0.24
%V	-0.20	-0.26
%C	0.20	0.26
VarcoV	*n.s.*	*n.s.*
VarcoC	-0.15	-0.16
VarcoSyl	-0.16	-0.11
nPVIV	*n.s.*	*n.s.*
rPVIC	-0.30	-0.25
nPVISyl	-0.22	-0.16
rPVISyl	-0.44	-0.36
%sp	-0.38	-0.27
srate	0.41	0.40

Table 3: *Correlation with scores (r1, r2) on the TPO data set (N=1019)*

Table 3 shows the Pearson correlations between the various rhythm metrics and the holistic English proficiency scores from two different raters (r1 and r2) for the TPO data set. The C- and V-based metrics that correlated best with the human scores were rPVIC, ΔV and %V. This indicates that a greater amount of variability in segment lengths or a greater proportion of vocalic speech was associated with lower scores. However the correlations associated with these metrics are relatively low compared to those associated with the syllable-based metrics, in particular speaking rate (r=(0.41, 0.40)) and rPVIsyl (r=(-0.44, -0.37)). This level of correlation is not too far off the inter-rater correlation for this data set (r=0.50). This suggests that lower proficiency scores correlate with slower speech and with greater duration changes from syllable-to-syllable. Note, these two metrics are also correlated (r=-0.49). The proportion of the utterance medial silence, %sp, was also relatively highly correlated with the pronunciation scores. We can take this to be a more traditional measure of fluency.

[1] `https://p2tk.svn.sourceforge.net/svnroot/p2tk/python/syllabify/syllabifier.py`

All of the metrics were significantly correlated with the human scores ($p < 0.001$) except the normalized vowel measures: VarcoV and nPVIV. This is somewhat unexpected given that [6] found VarcoV to be the most useful metric for discriminating L2 speech (Spanish, English). This may be due to the larger range of L1's associated with the L2 speech in this data set. However, the fact that neither of the two rate-normalized vowel measures correlated with the scores suggests that the rhythmic information provided by these metrics is dominated by influence of speaking rate on the pronunciation scores. So, while these metrics may highlight L1 features of L2 speech, this may not actually be very important for how fluent the speech is perceived to be by human raters.

%sp appears to be more or less independent of the other rhythm metrics: it was only significantly correlated with ΔV, VarcoV, and ΔSyl and all correlations were reasonably low. Interestingly, %sp did not have as high correlations with the scores as speaking rate. All metrics were significantly correlated with speaking rate except VarcoSyl. Given that speaking rate had a much higher correlation with human scores, it appears that segment level rhythm measures are not so useful for automated scoring. However, they may still be useful for understanding the underlying rhythm differences between different L2 speakers.

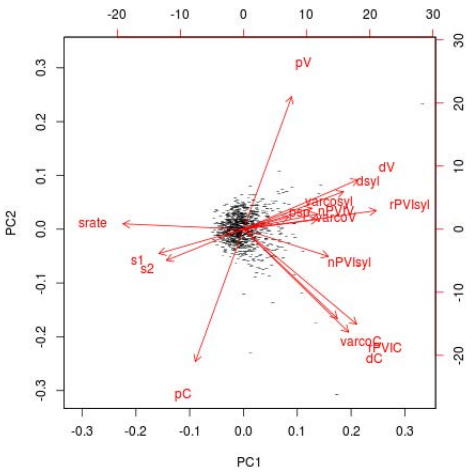
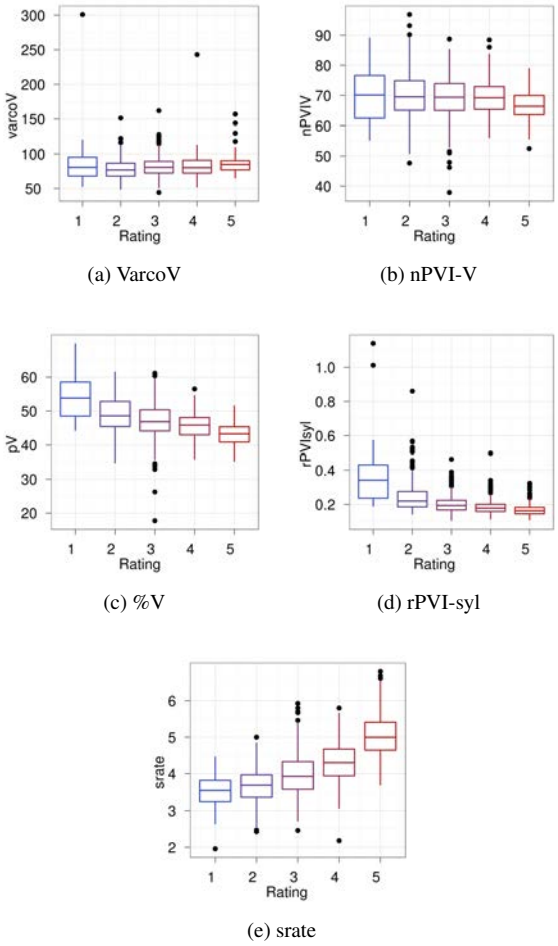


Figure 1: *Biplot of TPO rhythm metrics: rhythm data projected onto the first two components from a PCA.*

Principal components analysis on this data give us more information about the relationship between these measures. Figure 1 shows rhythm data projected onto the first two components from a Principal Components Analysis. The red arrows represent the original metrics with respect to these components. We see that syllable and vowel based measures cluster together and point in the opposite direction to the scores. Consonantal measures, on the other hand, appear orthogonal to the score vectors. %V appears independent of the other vowel measures.

3.2. Comparison to native speech

In order to investigate how the correlations between rhythm metrics and pronunciation scores relate to differences between native and non-native speech, we also applied the rhythm measures to the TOEFL native speaker forced alignment data. Figure 2 shows how the rhythm metrics differ for different score groups. In these figures, human scores from the first set of raters (r1 in Table 3) are used for the TPO responses, and the TOEFL-NS responses were each given an arbitrary rating of 5 to highlight the differences between the two corpora. We previously saw that neither VarcoV nor nPVI-V were significantly correlated with TPO scores. In Figure 2a we see that VarcoV (standard deviation/mean vowel duration) doesn't appear to really differ across scores or corpora. The similarity in the means for VarcoV seems to reflect the fact that what mattered for the scoring was not the variance in vocalic intervals, but rather the fact that the less fluent speakers spoke slower. In Figure 2b, we can see that nPVI-V distributions are not significantly different across the TPO data. However, the nPVI-V values for TOEFL-NS are significantly lower. This suggests that adjacent vocalic durations are more similar in L1 English, but this was a rhythmic factor not captured by the non-native speakers in this task. Figures 2c-2e confirm that %V, rPVIsyl, and speaking rate are good indicators of closeness to L1 English: L1 English speakers have a lower percentage of vocalic segments, lower syllable-to-syllable duration differences and a faster speaking rate.



(a) VarcoV (b) nPVI-V

(c) %V (d) rPVI-syl

(e) srate

Figure 2: *TPO and TOEFL-NS data: TPO data is rated between 1-4. TOEFL-NS data is rated 5 to highlight the differences between corpora.*

In order to compare the performance of all of the features across the two data sets, Figure 3 shows the correlations be-

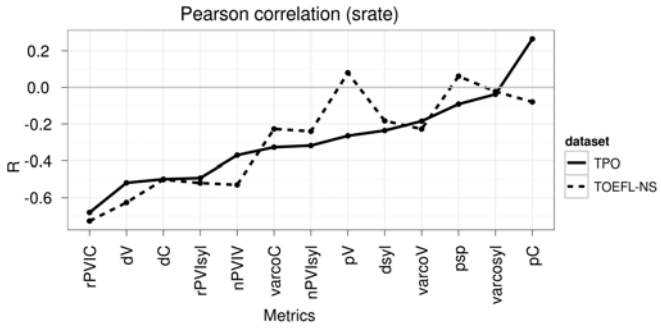
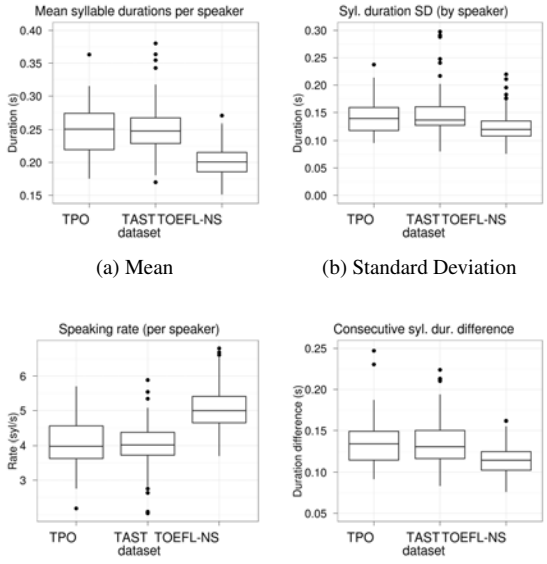


Figure 3: *Correlation with srate: TPO (non-native) and TOEFL-NS (native speakers) data sets*

tween the various metrics and speaking rate.[2] Looking at the correlations, we note that, unlike for the TPO data, %V was not significantly correlated with speaking rate in the native speaker data again suggesting that %V reflects a different aspect of speaker competence. The negative correlation with %V may reflect the presence of more filled pauses in the lower scored speech, which tend to have an extended duration.

3.3. Native/Non-native Syllable level differences



(a) Mean (b) Standard Deviation

(c) Speaking rate (d) Consecutive syllable differences

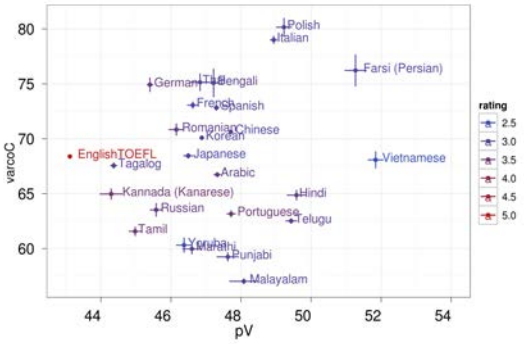
Figure 4: *Duration differences across corpora*

In the results reported above, syllable level measures had the highest correlation with pronunciation scores, particularly speaking rate and syllable to syllable variability. In order to test the robustness of these features across different data sets of non-native spontaneous speech, we compared timing data between the TPO and TAST data sets.

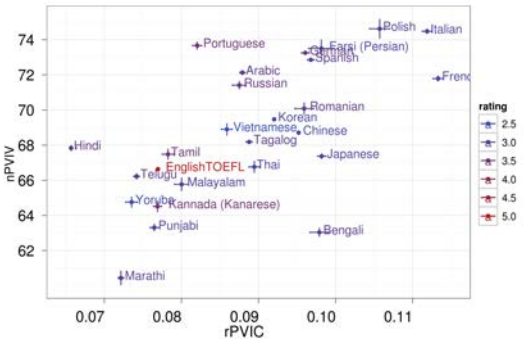
[2]In the figures, the symbol Δ in the rhythm metrics is represented by *d*, and % is represented by *p*; for example, dC represents ΔC and psp represents %sp.

Figure 4 shows mean syllable duration, standard deviations of syllable durations, speaking rate, and mean syllable-to-syllable differences, calculated for each speaker for the TAST, TPO and TOEFL-NS data sets. We see significant differences for these features between the native and non-native sets, but not between the non-native sets. As expected from the TPO results, the graphs show that non-native speakers speak slower, in terms of syllables per second, and have more variable syllable durations (excluding short pauses) in both the non-native speech corpora. So, it seems these relatively high level durational features are useful in the broad classification of native versus non-native speech.

3.4. The Effect of L1



(a) %V versus VarcoC



(b) rPVI-C versus nPVI-V

Figure 5: *L1 differences: Means and standard errors of the means in the TPO and TOEFL-NS data sets. The color scale represents mean proficiency scores for each language.*

Studies such as [1, 2] have suggested that rhythm metrics reflect consistent differences between L1 speech, as well as between L1 and L2 speech [6], based on read speech. To see how well these findings extend to spontaneous speech, we examined the distribution of rhythm measures values by L1. Figure 5 shows means for different L1 language groups in the TPO data set (for languages with more than 10 samples). Figure 5a mirrors the approach of [1] (%V versus VarcoC), while Figure 5b reflects that of [2] (nPVI-V versus rPVI-C). Neither graphs matches the previous studies in the terms of the ordering of languages. For example, English (typically categorized as a *stress timed* language) is expected to have a higher nPVI-V score than the Romance languages French and Spanish (which are categorized as *syllable timed*). Similarly, [6] find Spanish

speakers of English to have lower VarcoV than native speakers, which we do not find in our data. This casts doubt on the stability of these metrics on different corpora, e.g., when looking at read vs. spontaneous speech. Additionally, recent studies have called into question the notion that languages can be clearly differentiated based on rhythm metrics, since the degree of inter-speaker variability in these metrics for a single language is often similar in magnitude to the degree of variability between languages [4, 14].

Nevertheless, we do observe that the L2 speech roughly groups around language families: for example, French, Spanish, Portuguese and Romanian are relatively close. So, while these measures do not produce the same topology as the original native speaker studies, they still may be useful for characterising L1 transfer effects from the different language families.

4. Conclusions

In this study we investigated a variety of rhythm metrics in two corpora of non-native spontaneous speech and compared their distributions to a corpus of native speech. Several of the metrics resulted in large group-level differences between native and non-native speakers and also showed moderate correlations with holistic proficiency scores assigned to the non-native spoken responses. These two findings indicate that these types of metrics should be incorporated into applications that provide automated assessments of spontaneous spoken English, such as [15], in addition to the more commonly used fluency and pronunciation features.

In prior studies, duration based measures were shown to be useful for distinguishing native and non-native speech in terms of fluency, e.g. lower pause durations and higher speaking rate correlate with higher pronunciation scores [16, 17, 18]. This study replicated this finding by demonstrating that the *srate* feature had the highest correlation with proficiency scores among all of the rhythm metrics. However, this study also demonstrated that several additional segment-based rhythm metrics have significant correlations with proficiency scores with a magnitude close to the *srate* feature. This finding indicates that these rhythm metrics can be useful additional indicators of a non-native speaker's fluency in spontaneous speech.

Future research will integrate these rhythm features into a system for automated assessment of non-native speech in order to see how much of an impact these features can have on the prediction of holistic English proficiency scores. In addition, we will investigate the distributions of the rhythm features on a data set that contains both spontaneous and restricted speech in order to determine whether these two speaking styles have different rhythmic characteristics in non-native speech.

5. Acknowledgments

The authors would like to thank Su-Youn Yoon, Lei Chen, Anastassia Loukina, and three anonymous reviewers for their feedback about this paper.

6. References

[1] F. Ramus, M. Nespor, and J. Mehler, "Correlates of linguistic rhythm in the speech signal," *Cognition*, vol. 75, no. 1, 2000.

[2] E. Grabe and E. Low, "Durational variability in speech and the rhythm class hypothesis," *Papers in laboratory phonology*, vol. 7, no. 515—546, 2002.

[3] V. Dellwo, "Rhythm and speech rate: A variation coefficient for ΔC," *Language and language-processing*, pp. 231–241, 2006.

[4] A. Loukina, G. Kochanski, B. Rosner, and E. Keane, "Rhythm measures and dimensions of durational variation in speech," *Journal of the Acoustical Society of America*, vol. 129, no. 5, pp. 3258–3270, 2011.

[5] L. White, S. L. Mattys, and L. Wiget, "Language categorization by adults is based on sensitivity to durational cues, not rhythm class," *Journal of Memory and Language*, vol. 66, no. 4, 2012.

[6] L. White and S. Mattys, "Calibrating rhythm: First language and second language studies," *Journal of Phonetics*, vol. 35, no. 4, pp. 501–522, 2007.

[7] P. Mok and V. Dellwo, "Comparing native and non-native speech rhythm using acoustic rhythmic measures: Cantonese, Beijing Mandarin and English," in *Proceedings of Speech Prosody*, 2008.

[8] A. Tortel and D. Hirst, "Rhythm metrics and the production of English L1/L2," in *Proceedings of Speech Prosody*, 2010.

[9] J. Tepperman, T. Stanley, K. Hacioglu, and B. Pellom, "Testing suprasegmental English through parroting," in *Proceedings of Speech Prosody*, 2010.

[10] T.-Y. Jang, "Automatic assessment of non-native prosody using rhythm metrics: Focusing on Korean speakers English pronunciation," in *Proceedings of the 2nd International Conference on East Asian Linguistics*, 2009.

[11] E. Nava, J. Tepperman, L. Goldstein, M. Zubizarreta, and S. Narayanan, "Connecting rhythm and prominence in automatic ESL pronunciation scoring," in *Proceedings of Interspeech*, 2009.

[12] L. Chen and K. Zechner, "Applying rhythm features to automatically assess non-native speech," in *Proceedings of Interspeech*, 2011.

[13] J. Yuan and M. Liberman, "Speaker identification on the SCOTUS corpus," in *Proceedings of Acoustics '08*, 2008.

[14] A. Arvaniti, "The usefulness of metrics in the quantification of speech rhythm," *Journal of Phonetics*, vol. 40, pp. 351–373, 2012.

[15] K. Zechner, D. Higgins, X. Xi, and D. M. Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken English," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[16] J. Liscombe, "Prosody and Speaker State: Paralinguistics, Pragmatics, and Proficiency," Ph.D. dissertation, Columbia University, 2007.

[17] J. Yuan, Y. Jiang, and Z. Song, "Perception of foreign accent in spontaneous L2 English speech," in *Proceedings of Speech Prosody*, 2010.

[18] C. Cucchiarini and H. Strik, "Automatic assessment of second language learners' fluency," in *Proceedings of the 14th International Congress of Phonetic Sciences*, 1999.

Underdifferentiation of English Lexical Stress Contrasts by L2 Taiwan Speakers

Chiu-yu Tseng[1], Chao-yu Su[1] and Tanya Visceglia[2]

[1]Phonetics Lab, Institute of Linguistics, Academia Sinica, Taipei, Taiwan
[2]Education Center for Humanities and Social Sciences, National Yang Ming University, Taipei, Taiwan
cytling@sinica.edu.tw

Abstract

Learning the stress patterns of English words presents a challenge for L1 speakers from syllable-timed and/or tone languages. Realization of stress contrasts in previous studies has been measured in a variety of ways. This study adapts and extends Pairwise Variability Index (PVI), a method generally used to measure duration as a property of speech rhythm, to compare F0 and amplitude contrasts across L1 and L2 production of stressed and unstressed syllables in English multisyllabic words. L1 North American English and L1 Taiwan-Mandarin English speech data were extracted from the AESOP-ILAS corpus. Results of acoustic analysis show that overall, stress contrasts were realized most robustly by L1 English speakers. A general pattern of contrast underdifferentiation was found in L2 speakers with respect to F0, duration and intensity, with the most striking difference found in F0. These results corroborate our earlier findings on L1 Mandarin speakers' production of on-focus/post-focus contrasts in their realization of English narrow focus. Taken together, these results demonstrate that underdifferentiation of prosodic contrasts at both the lexical and phrase levels is a major prosodic feature of Taiwan English; future research will determine whether it can also be found in the L2 English of other syllable-timed or tone language speakers.

Index Terms: L2 English, lexical stress, Taiwan Mandarin, underdifferentiation

1. Introduction

While past studies of non-native accent have focused primarily on segmental variation between L1 and L2, a growing body of research has shown that differences between L1 and L2 production of prosody also make a significant contribution to the perception of non-native accent [1, 2]. One of the factors which have been demonstrated to affect intelligibility across a range of listener groups is misplacement or non-target realization of lexical stress. Field [3] required groups of native and non-native listeners to transcribe recorded material in which lexical stress had been acoustically manipulated. For both native and non-native groups, rightward stress shift and stress shift unaccompanied by a change in vowel quality were found to have the strongest effect on intelligibility. Tajima et al. [4] re-synthesized two-word utterances in Mandarin-accented English to match temporal characteristics of the same utterances recorded by native English speakers and temporally distorted the same utterances recorded by native English speakers to match the temporal characteristics of Mandarin-accented ones. Intelligibility of unmodified L1 English stimuli declined after temporal distortion from 94% to 83%. Intelligibility of unmodified L1 Mandarin English phrases was 39%, which increased to 58% after temporal correction.

Other specific features found in both Taiwan and Beijing L2 English include underdifferentiation of narrow-focus and on-focus/no focus contrasts [5, 6]. Similar underdifferentiation patterns were obtained in PVI analyses of Vietnamese Australian L2 English speech rhythm [7]. Both Taiwan Mandarin and Vietnamese are syllable-timed tone languages; thus, it is possible that such prosodic patterns are found in the L2 English of many other syllable-timed or tone language speakers.

The following study presents acoustic analysis of L1 English and L1 Taiwan-Mandarin English speech data extracted from the AESOP-ILAS corpus (Asian English Speech cOrpus Project, Institute of Linguistics, Academia Sinica) for the purpose of investigating differences in the realization of English lexical stress by L1 speakers of North American English and Taiwan Mandarin. This study focuses on contrast insufficiency at the lexical level and uses PVI to measure the production of stress contrasts in F0, duration and intensity. Traditionally, PVI has been used to measure differences in duration as a component of speech rhythm [8], but we have adapted and extended this method to measure F0 and intensity. PVI measures average difference in acoustic features between adjacent phonological units such as vowels, consonantal intervals or syllables. In present study, the syllable is adopted as the unit of feature extraction for representing stress-related contrasts. Our purpose in performing these analyses was to compare L1 English and Taiwan Mandarin speakers' realization of English lexical stress contrasts and to determine whether F0 and intensity patterns similar to those found in our comparison of L1 and L2 narrow focus would be obtained [5, 6]. Similar patterns would suggest that similar planning strategies are employed by L2 speakers at both the lexical and phrase levels. If different patterns emerge, our focus would shift to determining which acoustic correlates represent the most substantial source of difference between the L1 and L2 speaker groups, and discussing the implications of those differences.

2. Method

2.1. Recording Materials

The materials used in this study represent a subset of the core phonetic experimental tasks developed by AESOP (Asian English Speech cOrpus Project), a multinational collaboration established with the goal of building speech corpora to represent the varieties of English spoken in Asia [9] using the same recording set-up. This experiment uses Task 1, in which 1, 2-, 3- and 4-syllable target words of all possible stress patterns were embedded in a fixed, sentence-medial position; a total of 20 target words were selected (money, morning, white wine, hospital, apartment, department, tomorrow, video,

overnight, January, supermarket, elevator, available, Japanese, afternoon, misunderstand, information, experience, California, Vietnamese). Each of the experimental sentences contains one target word appearing in a broad-focused position two syllables removed from any phrase boundary.

2.2. Procedure

A total of 14 speakers: 7 L1 speakers (2 male and 5 female) and 7 Taiwan L2 speakers (3 male and 4 female) were recorded by trained proctors in quiet rooms directly into a laptop computer. Proctors used a recording platform developed specifically for the AESOP project with pre-loaded experimental sentences, each appearing individually on a computer screen. Participants wore head-mounted Sennheiser PC155 microphones positioned 2 cm away from their mouths; they were instructed to speak naturally at a normal rate and volume. All data were preprocessed automatically for segmental alignments using the HTK Toolkit then manually spot-checked by trained transcribers for accuracy of segmental alignment. Subsequent manual checking of F0 and intensity values was also performed to ensure extraction accuracy.

2.3. Data Analysis

The PVI index, i.e. the average difference in duration between adjacent phonological units such as vowels, consonantal intervals or syllables is among the most accepted methods to compare and represent rhythmic differences among languages [10]. Stress-timed languages are reported to exhibit higher PVI values than syllable-timed languages [8, 11]. Analyses of Japanese (mora-timed) and Estonian L2 English have shown that PVI is also a useful detector of non-native speech rhythm [11]. We began our analysis with the acoustic correlate duration and chose the syllable as the phonological unit of PVI analysis to more accurately reflect Mandarin speech rhythm. The syllable is also the phonological unit of Mandarin tone; thus, this choice also facilitates the inclusion of tone in future prosodic analyses. We then applied the same rationale to analyze average difference in the acoustic correlates F0 and intensity.

The data analysis procedure includes 2 steps: (1) calculating the difference between the current interval and the one that follows in terms of a particular acoustic parameter (2) computing the average of all differences. The PVI extraction equation appears below for duration d, in which k=syllable index and m= number of syllables in the target word:

$$\text{PVI} = \sum_{m-1}^{k-1} (|d_k - d_{k-1}| / (\text{m} - 1))$$

To facilitate comparison across speakers, duration values were subjected to Z-score normalization. F0 and intensity were normalized using the maximum and minimum values in each sentence. In addition, duration extraction was further refined to remove the effect of inherent segmental duration and boundary effects using the multi-layered normalization method that appears below [12], in which *factor1* represents information at the segmental level, *factor2* represents respective syllable position within the word (to remove word-final boundary lengthening effects), and ε_i represents all other unpredictable values. Extracted values μ_i thus represent duration values which have been normalized for inherent segmental duration and boundary effect:

$$x_i = \mu_i + factor_1 + factor_2 + ... + \varepsilon_i$$

3. Results

3.1. L1/L2 Production of Lexical Stress Contrasts

Average values for twenty English words across two speaker groups and three acoustic parameters are given in Table 1. Overall, results show between-group differences in all three acoustic parameters measured, with the most obvious difference appearing in F0, for which the degree of contrast produced by L1 English speakers is twice that of L2 speakers. In terms of duration and intensity, the degree of contrast produced by L1 speakers is only slightly higher, with a L1/L2 ratio of 1.281 and 1.003, respectively.

Table 1: *Average stress contrast for 20 English words by speaker group and acoustic parameter*

Speaker group / Prosodic attributes	L1	L2	L1/L2
F0	0.031	0.015	2.036
Duration	0.161	0.134	1.207
Intensity	0.1999	0.1994	1.003

3.2. PVI distribution by lexical item, speaker group and acoustic parameter

Figure 1 shows distribution of the L1/L2 PVI ratio across words and acoustic parameters, which was calculated in order to determine the most stable indicator across lexical items for distinguishing L1 and L2 speech. The dotted line in Figure1 represents equal L1/L2 PVI. Values above the dotted line indicate a higher level of stability as an indicator to distinguish L1 and L2. For F0, 19 out of 20 words exhibit higher than equal values; for intensity and duration, only 10 words exhibit higher than equal values.

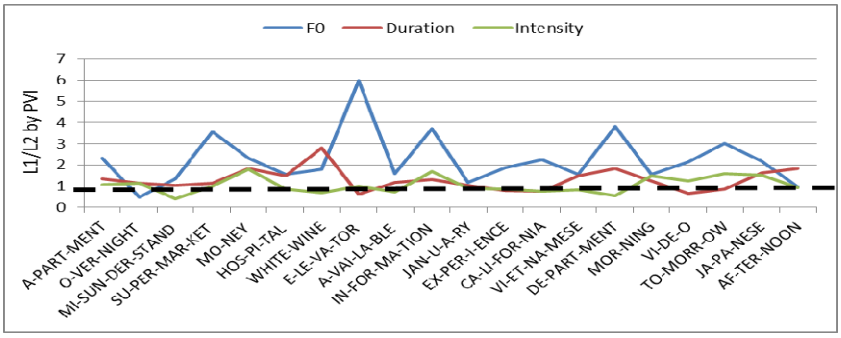


Figure1. L1/L2 ratio of PVI by acoustic parameter and word

3.3. Prosodic patterns by syllabicity, acoustic parameter and speaker group

In this section, we observe prosodic patterns occurring in 2-, 3- and 4-syllable words in order to determine whether production of stress contrasts could be related to the number of syllables or placement of stress in different lexical items. This analysis is illustrated in the figures below, containing the items "money" (2-syllable initial stress), "tomorrow" (3-syllable penultimate stress) and "California" (4-syllable penultimate primary stress, initial secondary stress) respectively. Each figure contains 3 sub-figures representing normalized F0, duration and intensity. Each sub-figure compares L1 and L2 English by individual acoustic parameter. The word "money", seen in Figure2, shows similar F0 and intensity patterns for both L1 and L2 English. F0 and intensity in the first syllable are higher than second; however, for both F0 and intensity a higher degree of contrast between syllables is found in L1 English. In terms of duration, the degree of contrast for both L1 and L2 English is minimal.

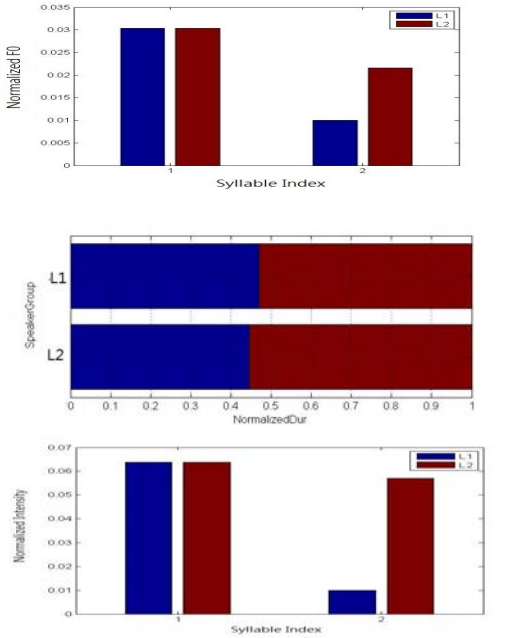


Figure2: *Acoustic patterns of "money" by acoustic parameter and speaker group*

The F0 and intensity pattern of "tomorrow", illustrated in Figure3, shows that L1 English speakers consistently produce the highest F0 and intensity on the second syllable, whereas L2 English speakers realize the same stress contrast using duration only.

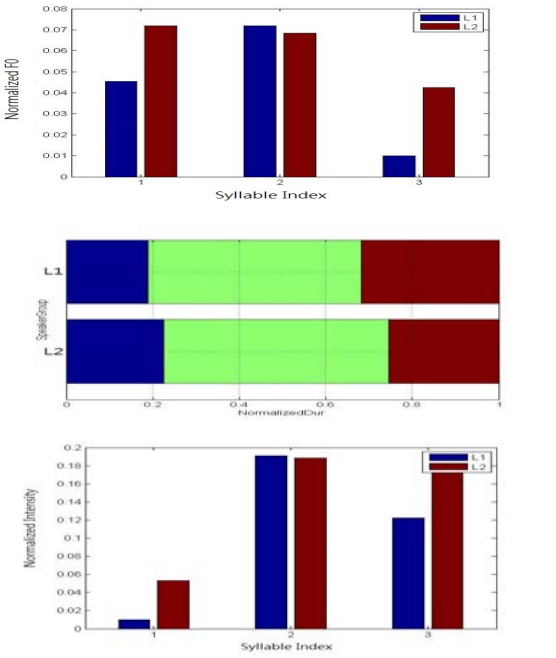
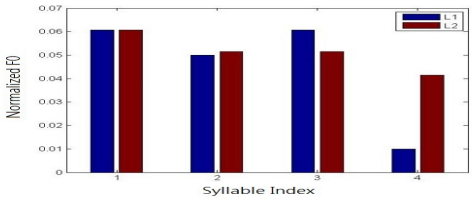


Figure 3: *Acoustic patterns of word "Tomorrow" by feature and speaker group*

In "California", L2 speakers produce a smaller intensity contrast than L1 speakers do. As for F0, only L1 English corresponds with the canonical stress pattern; for L2 speakers, the highest F0 value occurs in first syllable rather than the third. No clear stress contrast patterns in duration were found for either speaker group.



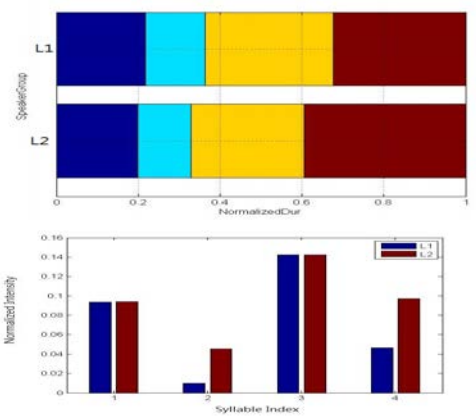


Figure 4: *Acoustic patterns of word "California" by feature and speaker group*

4. Discussion

Overall, results show between-group differences in all three acoustic parameters measured, with the most obvious difference appearing in F0, for which the degree of contrast produced by L1 English speakers is twice that of L2 speakers. Item-based analyses also confirm that F0 is a more stable indicator than duration and intensity to distinguish L1 from L2 speech across lexical items. Thus, underdifferentiation of F0 contrast in realization of lexical stress seems to make a more substantial contribution to Taiwan-accented English than either duration or intensity. Analyses based on syllabicity found that in L1 English, the highest F0 and intensity are always realized on primary-stress syllables, but the difference between primary and secondary stress syllables is not very distinct (e.g. the distinction between the first (secondary-stress) and the third (primary-stress) syllables in the word "California"). In contrast, L2 English patterns of F0 and intensity do not always correspond to canonical stress patterns, and in cases in which they do correspond, the degree of contrast produced is lower than that of L1 English. It is interesting to note that no clear duration patterns were observed for either group in this analysis.

The present study has obtained results similar to those of previous studies [7] in which the L2 English of syllable-timed L1 speakers exhibits substantially less duration contrast in realization of lexical stress, suggesting rhythmic difference is a major prosodic feature. However, PVI analysis of F0, duration and intensity has revealed that stress contrast is realized more robustly by means of F0 than by duration or intensity. Moreover, for the words recorded by the L1 speakers in our experiment, the highest F0 and intensity were always realized on stressed syllables, whereas duration either often exhibited no clear pattern or played a relatively smaller role. In contrast, the Taiwan English speakers' production of F0 and intensity did not always correspond to canonical stress patterns; when they did, however, the L2 speakers realized the contrast less robustly than native speakers did. Interestingly, this speaker group also exhibited no clear patterns with respect to duration.

5. Conclusion

Based on these results, it appears that L1 English speakers produce lexical stress contrasts more robustly than L2 Taiwan English speakers do. Moreover, their pattern of contrast underdifferentiation is realized in terms of F0 and intensity, echoing the pattern found in our study of narrow focus, which suggests that insufficient contrast is a feature of L2 prosody at both the lexical and phrase levels. Using PVI to measure F0 and intensity, in addition to duration, further revealed that F0 and intensity appear to play a larger role than duration in marking English stressed syllables. However, we must note here that that inconsistent stress assignment were found across three different dictionaries for 6 of the 20 words in our task (tomorrow, hospital, video, overnight, misunderstand and Vietnamese). Subsequent studies will investigate whether and how the inconsistencies are realized by both L1 and L2 speakers. Since F0 appears to be the most salient cue of underdifferentiation, future studies will include more refined, syllable-internal analysis of the same words embedded in a variety of intonation contexts in order to examine the effect of layering higher levels of prosodic information on their production. Future research will also investigate the question of whether similar patterns can be found in the L2 English of other syllable-timed and tone language speakers.

6. References

[1] Magen, H.S., "The perception of foreign-accented speech", Journal of Phonetics, vol. 26, 381-400, 1998.

[2] Anderson-Hisieh, J., Johnson, R. and Koehler, K. "The relationship between native speakers judgments of nonnative pronunciation and deviance in segmentals, prosody and syllable structure", Language Learning 42: 4 529-555, 1992.

[3] Field, J. (2005). Intelligibility and the listener: The role of lexical stress. *TESOL Quarterly*, 39(3), 399– 423.

[4] Tajima, K., Port, R., and Dalby, J. (1997). Effects of temporal correction on intelligibility of foreign-accented English. Journal of Phonetics, 25, 1-24.

[5] Visceglia, T., Tseng, C. Y., Su, Z. Y. and Huang, C. F. "Realization of English Narrow Focus by L1 English and L1 Taiwan Mandarin Speakers", The 7th International Congress of Phonetic Sciences. Hong Kong, China, 2011.

[6] Visceglia, T., Su, C. Y. and Tseng, C. Y. "Comparison of English Narrow Focus Production by L1 English, Beijing and Taiwan Mandarin Speakers", Oriental COCOSDA 2012 47-51. Macau, China, 2012

[7] Mixdorff, H. and Ingram, J. "Prosodic analysis of foreign-accented English", Proc. Interspeech 2009, 6-10 Sep. Brighton UK, 2009.

[8] Grabe, E. & Low, E. L. "Durational variability in speech and the rhythm class hypothesis ", In Gussenhoven, C. & Warner, N. (eds.) Papers in Laboratory Phonology 7, Berlin, Mouton de Gruyter, 515-546, 2002.

[9] Visceglia, T., Tseng, C. Y., Kondo, M., Meng, H and Sagisaki, Y. "Phonetic aspects of content design in AESOP (Asian English Speech cOrpus Project)", Oriental COCOSDA 2009. Beijing, China, 2009.

[10] Asu, E.L. & Nolan, F. "Estonian and English rhythm: a two-dimensional quantification based on syllables and feet", In Proceedings of Speech Prosody 2006, Dresden, Germany, OS1-5_0229, 2006.

[11] Kinoshita, N. and Sheppard, C., "Validating acoustic measures of speech rhythm for second language acquisition", ICPhS XVII, Hong Kong, 1086-1089, 2011.

[12] Tseng, C. Y. Su, Zhao-yu. "Dynamic Discourse Speech Tempo and Phonological Timing", The 7th International Congress of Phonetic Sciences. Hong Kong, China, 2011

Intermediate phonetic realizations in a Japanese accented L2 Spanish corpus

Mario Carranza

Departamento de Filología Española, Universitat Autònoma de Barcelona,
Edifici B, Campus UAB, Cerdanyola del Vallès, 08193 Bellaterra, Barcelona, Spain

mariocarranzadiez@gmail.com

Abstract

This paper addresses the issue of manual transcription of non native speech in an attempt to establish rule-based strategies for labelling intermediate realizations. The problems of transcribing non canonical realizations of L2 sounds which present shared features of the target (Spanish) and the source language (Japanese) will be considered. We introduce a Japanese accented non native L2 Spanish corpus, and exemplify the use of decision trees in manual transcriptions as a systematic method for dealing with ambiguous realizations. This approach could help a potential error detection system to detect both canonical and erroneous realizations, contributing to the development of CAPT tools.

Index Terms: non native speech transcription, ASR, CAPT, L2 Spanish, L1 Japanese, non native spoken corpus

Introduction

Adapting Automatic Speech Recognition (ASR) systems to non native speech raises several difficulties. Since canonical native speech is used during the training phase in general-purpose recognizers, the generated acoustic models cannot fit properly with the non native data and this leads to an increase in Word Error Rate (WER). Many proposals have been put forward to find methods of adapting the ASR systems to non native data ([1], [2]). The adaptation can be done at the HMM level or by means of knowledge-based rules, that can automatically generate variants which are incorporated in the pronunciation dictionary ([1]). A pronunciation error detection system can be developed by generating new acoustic models for the non native realizations of L2 phonemes and by the systematization of L1-based typical errors by means of rules ([1], [2]). In order to do so, phonetically transcribed non native spoken corpora are needed; however, manual transcription of non native speech is a time-consuming costly task, and current automatic transcription systems are not accurate enough to carry out a narrow phonetic transcription.

Moreover, non native speech shows two characteristics that make manual transcription a challenging task. First of all, a high degree of variability may cause a lack of consistency in the realizations of L2 phonemes. This variability is due not only to the simultaneous presence of L1 and L2 sounds, but also to the occurrence of sounds which do not belong to the L1 or L2 inventories ([3]). Second, the process of acquisition of the L2 phonetic system is dynamic; in other words, the process of creating new phonemic categories or adapting the categories of the source language to the new categories of the target language is not consistent and changes through time ([3]). Therefore, non native speech cannot be described in terms of a correct/incorrect dichotomy, but rather as an acoustic continuum from the non native realization of the targeted L2 sound to its canonical realization. Within this continuum there is a wide range of different realizations which cannot be considered either completely erroneous or canonical. This kind of realizations could possibly interfere in the recognition performance of ASR systems resulting in an increase in error rate.

The goal of our project is to compile a Japanese accented database of (semi)spontaneous L2 Spanish speech that could be used as a training database for the ASR module of a Computer Assisted Pronunciation Teaching (CAPT) system. Two transcription levels –canonical phonemic and narrow phonetic– are included with the aim of automatically deriving rule-based generated pronunciation variants for Japanese accented Spanish speech. Nevertheless, labelling problems arise when dealing with ambiguous phones which present mixed features belonging to the target and to the source languages. In this paper we will present a proposal for labelling intermediate categories based on systematic decision processes. In section 1 we will describe a non native spoken corpus compiled for this research. The difficulties of labelling and transcription of intermediate categories and the use of decision trees for disambiguation are presented in section 2, followed by a conclusion in section 3.

1. Database description

A Japanese accented L2 Spanish corpus was compiled, transcribed and manually annotated in view of its future adaptation as a training corpus for developing CAPT tools oriented to Japanese learners of Spanish. The data was obtained from the recordings of oral exams at the Tokyo University of Foreign Studies. Pronunciation errors were systematically encoded using a notation system suited to the automatic processing of errors and to the development of transformation rules.

1.1. Contents

The recorded data consists of 8.6 hours of spontaneous, semi-spontaneous and read speech. Each participant was recorded four times –once every 6 months– throughout the first two academic years at the university. The recordings took place from April 2010 to March 2012. Twenty students (10 male and 10 female) participated in the project. The participants were carefully selected to be representative of the population; therefore, students were chosen considering, among other criteria, their oral proficiency level: 6 students with a high oral proficiency level, 8 students with an intermediate level and 6 students with a low oral proficiency level.

The corpus contains different tasks designed according to the learning stage: semi-spontaneous speech (students were asked to speak about a topic given some weeks in advance) followed by a conversation with the examiner after 6 months and after 12 months of study; spontaneous speech (students were asked to act in a daily situation with no previous preparation), a conversation and a reading task after 18 months; finally, two samples of spontaneous speech and a reading task were required after 24 months.

The recordings were segmented into individual audio files that were transformed to WAV format and labelled using a code which represents speaker number, test period and task type. This will allow to automatically compute error rates according to proficiency level, learning stage and task. The total recording time sorted by task type and learning stage is shown in Table 1.

Task type/Stage	6	12	18	24	Total
semi-spontaneous	39	52	–	–	91
spontaneous	–	–	64,1	150	214,1
reading	–	–	5,4	4	9,4
conversation	57,3	78	66,3	–	201,6
Total	96,3	130	135,8	154	**516,1**

Table 1. *Total recording time (in minutes) by task type and learning stage (in months).*

1.2. Transcription levels and error annotation

The recordings were manually segmented and transcribed using Praat ([4]). The TEI conventions ([5]) and the EAGLES guidelines ([6]) for the orthographic transcription and encoding of spontaneous speech were followed. The text was annotated in XML format; some specific labels were added to cover non native speech phenomena. Since the corpus is aimed at the analysis of errors produced by non natives speakers, a canonical phonemic level representing the standard pronunciation of the word, and a narrow phonetic transcription level for the actual pronunciation by the speaker were included ([7]); the alignment of both levels is intended to allow the automatic generation of pronunciation variants and the retrieval of statistical information for all words in the corpus. Finally, an error tier to label and encode the pronunciation errors for their further processing plus two additional tiers for non-lexical phenomena were added. In total, the transcription consists of six levels of representation, as shown in Figure 1.

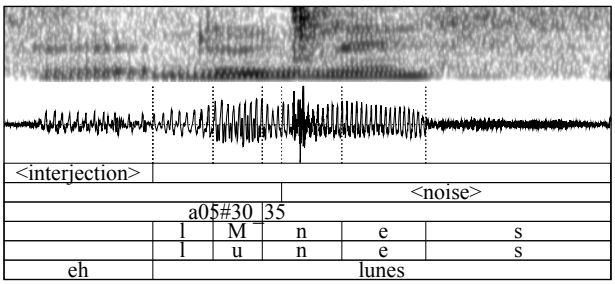


Figure 1: *Levels of transcription and annotation.*

1.2.1. Orthographic transcription

In the orthographic tier each word was transcribed in its standard orthographical form. No punctuation marks were used due to the difficulties of establishing sentence boundaries in spontaneous speech. The vocal content, such as filled pauses, was also orthographically transcribed in a standardized form and labelled in the vocal tier (see 1.2.4). XML labels were used to mark phenomena such as foreign words, non-existing words, repetitions, truncations, unclear and unintelligible utterances.

1.2.2. Canonical phonemic transcription

The canonical phonemic transcription level shows the phonemic representation of words as pronounced in isolation. Thus, coarticulatory phenomena taking place within words and at word boundaries are not considered. Northern Castilian Spanish was adopted as the standard form for the transcription; although some of the Japanese students received input from speakers of other Spanish dialects, Castilian Spanish was the main variety used in their courses. An adaptation of SAMPA to Spanish ([8]) was used for the transcription; the symbols corresponding to allophonic variants were not considered in this level. Only linguistic content was phonemically transcribed; the noise-corrupted cases were marked as <unclear> in the orthographic transcription and left with no phonemic transcription.

1.2.3. Narrow phonetic transcription

The actual pronunciation of words by the speakers is presented in the narrow phonetic transcription tier. In order to better reflect the Japanese pronunciation of L2 Spanish, 11 new symbols and 7 diacritics from X-SAMPA ([9]) were incorporated to the initial inventory. A preliminary version of the narrow phonetic transcription was automatically generated from the canonical phonemic tier. The resulting new tier was manually checked, realigned and corrected; the additional X-SAMPA symbols and diacritics were used to represent the Japanese pronunciation of Spanish sounds.

1.2.4. Other levels of representation

Vocalized but non lexical content (semi-lexical elements according to TEI [5]), such as hesitations (including filled pauses), laughters and interjections were marked in the vocal tier. Since the acoustic realization of these phenomena is usually very similar to that of linguistic sounds –such as long vowels for the hesitations– the aim of doing this was to explicitly identify the segments which should not be considered for the development of the acoustic models in the ASR training. Other phenomena such as coughs, breathing, ambient noise, as well as overlapping speech of the examiner (non lexical phenomena in the TEI guidelines) are marked in the incident tier.

1.2.5. Error annotation

Whenever the canonical phonemic tier and the narrow phonetic tier presented a dissimilarity (not derived from the Spanish coarticulation rules), this was considered an error to be included in the error tier. All phones in the transcriptions have a corresponding two digit code used for error annotation. The encoding procedure is similar to that proposed in [10] and consists on a string of six numeric characters separated by (#) and (_) symbols. The first two digits correspond to the code of the affected phone, separated by a # character; the following four represent the previous and following sounds (the phonologic context of the error) separated by a _ character; finally, a letter (a, b or c) was added at the beginning of the string to represent the type of error. The letter "a" corresponds to errors of substitution of one phone by another; "b" represents errors of insertion; and "c" stands for the deletion of one phone that should have appeared in the Spanish canonical form. Since this annotation system was originally created to cover all type of errors, it is not possible to encode the resulting phone in substitution cases; however, as the error label is aligned with the narrow phonetic and canonical phonemic transcription tiers in the corpus, this information can be automatically recovered. Figure 2 shows an example of the error encoding. In this case, a substitution of the target phone [e] (code: 02) by the phone [i] has occurred between the phones [β̞] (26) and [s] (23).

The error annotation system combined with the speaker and task type codification in the file name were used to automatically quantify error types considering also the affected phone and the phonological context. Frequency of occurrence and likelihood ratios for each error will be also available after the processing of the corpus, and can eventually become a useful source of statistical information in the generation of data-driven pronunciation rules.

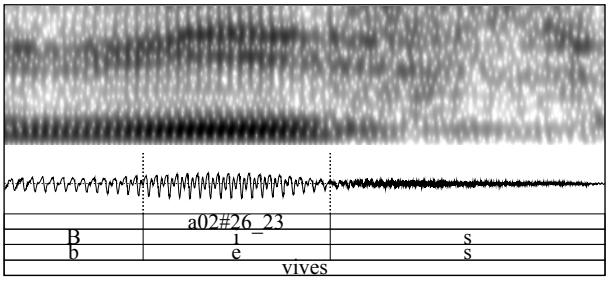


Figure 2: *Example of error annotation (phone substitution).*

2. Intermediate realizations

One of the difficulties in the manual transcription of the corpus is due to ambiguous realizations of L2 sounds that present features of both the target and the source language. In these cases transcribers should establish clear criteria for deciding whether the realization should be classified as correct or incorrect based on the acoustical information, since perceptual judgments might be biased and the level of disagreement might be high if more than one transcriber is involved ([6]).

During the transcription of the corpus three types of intermediate categories have been detected, depending on the phonetic or phonologic status of the related sounds.

2.1. Realizations between two L2 phonemes

In Spanish there is a phonological contrast between the alveolar lateral approximant /l/ and the alveolar rhotic tap /ɾ/. In Japanese, both phones are possible realizations of the alveolar tap phoneme, represented also by the symbol /ɾ/ ([11]). Although several studies have shown that there is a context-dependent preference for one realization or the other ([11], [12]), the realization of this phoneme seems to exhibit a relatively high degree of individual variation ([11], [12]). Furthermore, some cases of an intermediate category [l̺] (lateral flap) have been detected in the corpus (Figure 3). Intermediate realizations present formant continuity, typical of the lateral realization [l] and a closure release, more characteristic of the tap realization [ɾ].

In order to distinguish between [l] or [ɾ] realizations in intervocalic position, duration must be taken into account, as segmental duration seems to be longer in lateral than in rhotic segments; this tendency is also found in native Spanish ([13]). Therefore, a speaker-based duration threshold can be established as a way of disambiguating intermediate realizations.

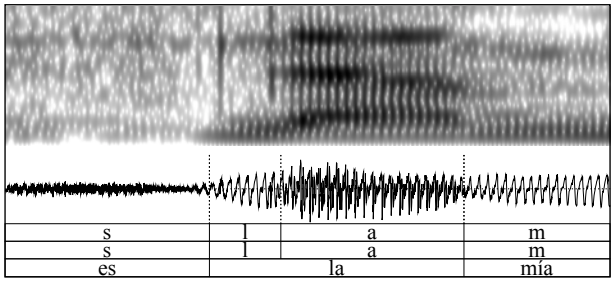


Figure 3: *Intermediate realization between [ɾ] and [l].*

In consonant cluster position some realizations that might be described as showing mixed features of both sounds have also been detected. Figure 4 shows a case of vowel epenthesis before /l/ in consonant cluster position; although, in Spanish vowel epenthesis are characteristic of rhotic segments in consonant clusters. This case was considered an error of phone insertion before the [l] phone.

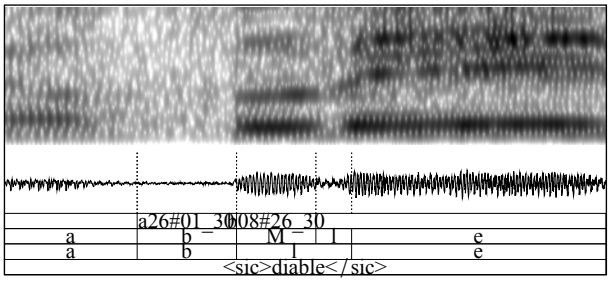


Figure 4 : *Lateral realization of [l] with vowel epenthesis*.

2.2. Realizations between two L2 allophones

The Spanish phonemes /b/, /d/ and /g/ present two allophonic variants: stops ([b], [d], [g]) after a pause or a nasal consonant ([d] can also appear after [l]), and approximants ([β̞], [ð̞], [ɣ̞]) in all other contexts ([14]). Approximant allophones for these phonemes do not exist in Japanese, so Japanese students of L2 Spanish tend to replace the approximant realizations by their stop counterparts. Nevertheless, intermediate realizations can show the formant continuity characteristic of approximant realizations as well as the closure release typical of stops (Figure 5). Since no clear difference in segment duration between the two allophones has been detected in the non native realizations, alternate ways of distinguishing between stop or approximant realizations should be considered.

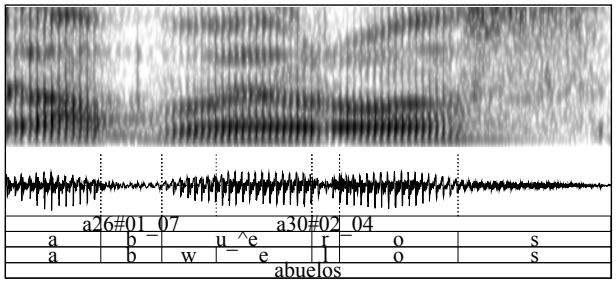


Figure 5: *Intermediate realization between [b] and [β̞].*

Figure 6 shows a decision tree for the disambiguation between the stop and the approximant realizations. First of all, a hierarchical set of acoustic features is established. Each node in the tree corresponds to a yes/no question regarding acoustic features, the spectrum, or the waveform of the segment. In the example shown in Figure 6, a maximum of three steps is needed to disambiguate. At the end of each leaf of the tree, one phone is preferred as candidate for the realization and the other is discarded. Although the decision tree shown in Figure 6 corresponds to the disambiguation between [b] and [β̞], it is also valid for the disambiguation of [d] / [ð̞], and [g] / [ɣ̞].

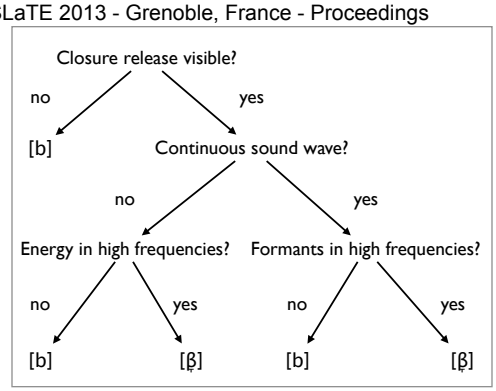


Figure 6: *Decision tree for the disambiguation between [b] and [β]*

.

2.3. Realizations between one L2 phoneme and one L1 phoneme

Intermediate categories can also be the result of the realization of new L2 phonemes which have no correspondence in the L1 inventory, but are similar to one or more L1 categories. For example, the Spanish unvoiced fricative velar phoneme /x/ is usually assimilated by Japanese speakers to the Japanese fricative phoneme /h/, which presents four allophones depending on the following vowel ([12]): voiceless/voiced glottal fricatives [h] / [ɦ], voiceless bilabial fricative [ɸ], and voiceless palatal fricative [ç]. Furthermore, some mixed realizations between these sounds have also been found in the corpus. Figure 7 represents an example of an intermediate realization between a velar fricative [x] sound and a palatal fricative [ç] sound. The decision tree shown in Figure 8 for the disambiguation of these sounds takes into account the following vowel as well as spectral features to disambiguate between various possible realizations of the targeted [x] sound.

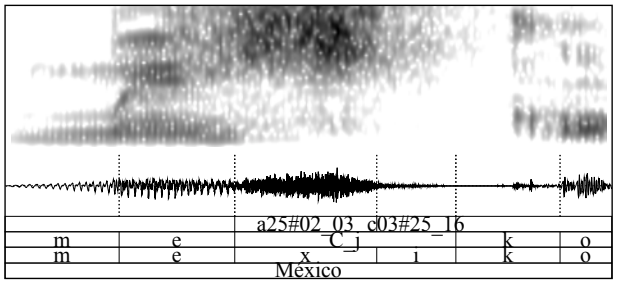


Figure 7: *Intermediate realization of [x] as [ç].*

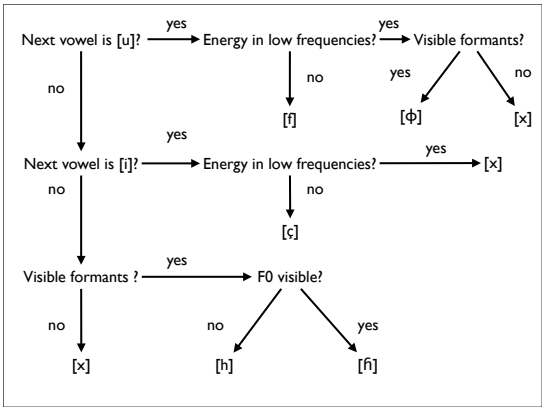


Figure 8: *Decision tree for the disambiguation of [x], [h], [ɦ], [ɸ] and [ç].*

.

3. Conclusion

The adaptation of ASR technology to non native speakers requires a considerable amount of phonetically transcribed non native data; moreover, if the technology is to be applied in computer assisted pedagogical applications, it should be able to successfully detect mispronunciations of targeted sounds. In order to achieve so, a detailed phonetic transcription of the data is required, but manual transcription faces some difficulties when dealing with non native speech.

This paper provides a description of a Japanese accented Spanish L2 spoken corpus, phonetically annotated considering its future application to the development of a Spanish CAPT system aimed at Japanese speakers. During the manual transcription of the corpus intermediate realizations of L2 phonemes were documented, showing the need for disambiguation strategies. We propose the adoption of a decision tree for each case of intermediate realization, based on acoustic criteria. This approach is aimed at helping transcribers in choosing the adequate phone label in narrow phonetic transcription of L2 speech. Furthermore, given that decision trees can be mathematically expressed as algorithms, once manually tested they can be implemented for automatic corpus processing. The advantage of decision trees relies in the fact that they can be systematically applied; since they are not based on perceptual criteria, their adoption could contribute to overcome some of the difficulties found in the transcription of non native speech.

4. References

[1] S. Goronzy, *Robust Adaptation to Non-Native Accents in Automatic Speech Recognition*. Berlin: Springer, 2002.

[2] R. E. Gruhn, W. Minker, and S. Nakamura, *Statistical Pronunciation Modeling for Non-Native Speech Processing*. Berlin: Springer, 2011.

[3] J. E. Flege, "Phonetic approximation in second language acquisition," *Language Learning*, vol. 30, no. 1, pp. 117–134, 1980.

[4] P. Boersma and D. Weenink, *Praat: doing phonetics by computer,* 2009. [Online]. Available: http://www.fon.hum.uva.nl/praat/. [Accessed: 31-Mar-2013]

[5] TEI Consortium, "Transcription of speech," *TEI P5: Guidelines for electronic text encoding and interchange*, 2013. [Online]. Available: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/TS.html. [Accessed: 31-Mar-2013].

[6] D. Gibbon, R. Moore, and R. Winski, *Spoken language system and corpus design*. Berlin: Mouton De Gruyter, 1998.

[7] A. Neri, C. Cucchiarini, and H. Strik, "Selecting segmental errors in non-native Dutch for optimal pronunciation training," *IRAL - International Review of Applied Linguistics in Language Teaching*, vol. 44, no. 4, pp. 357–404, 2006.

[8] J. Llisterri and J. B. Mariño, "Spanish adaptation of SAMPA and automatic phonetic transcription," ESPRIT PROJECT 6819 (SAM-A Speech Technology Assessment in Multilingual Applications), 1993.

[9] J. C. Wells, "Computer-coding the IPA: a proposed extension of SAMPA," *Speech, Hearing and Language, Work in Progress*, vol. 8, pp. 271–289, 1994.

[10] S. Detey, "Coding an L2 phonological corpus: from perceptual assessment to non-native speech models — an illustration with French nasal vowels," (in press).

[11] M. Kimura, "Phonetic Errors of the Spanish lateral sound /l/: Problems and correction," *Sophia Linguistica*, vol. 20/21, pp. 287–296, 1986.

[12] T. Akamatsu, *Japanese Phonetics*. München: Lincom Europa, 1997.

[13] S. Torner and L. del Barrio, "La duración consonántica en castellano," *Estudios de Lingüística*, vol. 13, pp. 9–36, 1999.

[14] J. I. Hualde, *The Sounds of Spanish*. Cambridge: Cambridge University Press, 2005.

Universal contrastive analysis as a learning principle in CAPT

Jacques Koreman, Preben Wik, Olaf Husby, Egil Albertsen

Department of Language and Communication Studies, NTNU, Trondheim, Norway
jacques.koreman@ntnu.no, prebenwik@gmail.com, olaf.husby@ntnu.no, egil.albertsen@ntnu.no

Abstract

Cross-linguistic comparison is a good starting point for computer-assisted pronunciation training (CAPT). A comparison between the segment inventories of a learner's mother tongue (L1) and the target language (L2) can be made on the basis of the IPA categories. Since these are claimed to reflect universal phonemic distinctions, mastering the contrasts in the target language in perception and production ensures communicative effectiveness for learners at the segmental level. The Computer-Assisted Listening and Speaking Tutor (CALST) implements contrastive analysis in two types of exercises for phonetic and abstract listening. In these exercises users can practice with word pairs/sets demonstrating unfamiliar sound contrasts of the target language to improve their perceptual discrimination. Since substitutions for unfamiliar sounds depend on L1, the selection of sound contrasts which are trained in the exercises should also depend on L1. We shall argue for a pragmatic approach to the selection of exercises.

Index Terms: computer-assisted pronunciation training, CAPT, CALST, sound contrasts, differential substitution

1. Introduction

Pronunciation training is often given limited attention in foreign language classes. One of the reasons is that users with different native languages often have very different challenges when it comes to acquiring a correct, or at least communicatively effective, pronunciation. These varied challenges are difficult to address in class, also because language teachers can at best only have in-depth knowledge of a limited number of foreign languages. We have therefore developed a CAPT system for Norwegian which adapts the learning trajectory to the user's native language (L1) on the basis of a contrastive analysis. The system is called the Computer-Assisted Listening and Speaking Tutor (CALST). It is used in Norwegian classes for foreign students and staff at NTNU and can be downloaded for free [1].

The system uses cross-linguistic comparison as a basis for pronunciation training. Although the Contrastive Analysis Hypothesis, especially its strong (predictive) version, has been criticized [2], it is nevertheless a good starting point in language teaching. Or as Ellis expressed it: "No theory of L2 acquisition is complete without an account of L1 transfer" [3, p. 341]. In classes where learners with the same language background learn a second language (L2), e.g. Norwegian students studying English in Norway, this principle can be implemented successfully. But in classes where the students have mixed native language (L1) backgrounds this is not possible, and L2 teachers often adopt a more constrained implementation of contrastive analysis, focusing on typical pronunciation problems which most foreign learners have when they acquire the L2 – often in combination with instruction to correct individual pronunciation errors in class.

With the increasing variation in language backgrounds of students in the Norwegian courses at NTNU (and presumably also in language courses in many other countries, given the increasing mobility across countries), variation in the students' L1 backgrounds becomes more and more of a challenge to standard classroom teaching. Computer-assisted pronunciation training (CAPT) systems can use a contrastive approach to guide language learners with varying L1's through relevant instruction and individual exercises, and may thus help to meet this challenge. This article describes a segmental CAPT implementation of the principle of contrastive analysis on a "universal" basis in Section 2.

At present, CALST users work through *all* the exercises for *unfamiliar* L2 sounds. Ideally, a narrower selection of exercises should be made: It is well known that differential substitutions occur across L1's, i.e. an L2 speech sound may be substituted with different sounds depending on the learner's L1. Differential substitutions occur even if the L1's have the same relevant (phonetically close) set of phonemes in their inventories: German and Dutch both have /t/ and /s/, but German speakers often substitute /s/, while Dutch speakers often substitute /t/ for English /θ,ð/ [4]. Since we aim to offer CAPT users an efficient learning trajectory, only exercises for substitutions which actually occur should be offered. The link between contrastive analysis and the selection of appropriate exercises (which would be different for German and Dutch learners of English because of the different substitutions) is discussed in Section 3. We describe possible linguistic approaches to this challenge and their inherent problems, and explain why and how we have decided to implement a pragmatic solution instead.

2. Learning sound contrasts

The main stress in the CALST system is on listening skills, since it is generally accepted that this is an important prerequisite for correct pronunciation [5], but also because pronunciation skills require automatic error detection, which at present still has clear limitations [6]. CALST has two special features [7]. The first is that for each user, the exercises which the learner is guided through are based on a comparison of the Norwegian segment inventory with the segment inventory of the user's native language. At present, CALST uses a database which contains over 500 languages. Although CALST has Norwegian as the target language, the contrastive analysis tool can be used as a basis for CAPT systems for any language. The contrastive analysis is further described in Section 2.1.

The second special feature of CALST, described in Section 2.2, is determined by a peculiarity of Norwegian: It has no accepted pronunciation standard. For this reason four main dialect variants, each represented by one male and one female speaker (with sub-dialectal variation), are available in CALST.

Instead of forcing the user to exclusively commit to one single dialect as the target dialect, the system offers the possibility of acquiring production skills in one (target) dialect, while the user can develop perception skills in all four dialects. The latter is necessary because learners will hear all dialects in their everyday interactions and pronunciation can vary strongly across dialects, with the sound inventories varying in size from 42 to 53 phonemes [8].

After defining the source and target language/dialect, the user is guided through exercises for sound segments that are likely to present a challenge to learning Norwegian pronunciation. The exercises use minimal pairs which demonstrate sound contrasts. Each exercise starts with a short articulatory explanation of the similarities and differences between the contrasted speech sounds (Section 2.3).

The user can then choose between two exercise types. The first exercise type enables the user to listen phonetically to the unfamiliar sound in contrast with another sound in a minimal pair. After listening to the word pair the user hears a new realization of one of the two words again, and has to decide which of the first two words it resembles. A more detailed description is given in Section 2.4.

The second exercise type, described in Section 2.5, requires more abstract listening skills. In these exercises, the listener only hears one word, and has to select a button on the screen which is labeled with that word.

To facilitate understanding, it is recommended that the reader first download and open CALST [1].

2.1. Multi-language contrastive analysis

The tool L1-L2*map* [9] was developed on the basis of the UCLA phonetic segment inventory database (UPSID, [10]) to enable contrastive analysis of any language pair. The UPSID database, which contains 451 languages, was extended with languages spoken by larger immigrant groups in Norway which are not available in the UPSID database, and we now have access to the segment inventories of over 500 languages. Since foreign speakers have problems with pronouncing known segments in unusual syllable positions [11,12], position in the syllable was also added as a descriptive feature and consonants can be marked for their appearance in syllable onset, nucleus and coda [9]. Since vowels always occur in the syllable nucleus, they are not marked for position.

The visualization of the contrastive analyses in L1-L2*map* is based on an easily interpretable colour coding in the consonant and vowel charts of the International Phonetic Association [13]. This makes it easy to interpret the analysis results, although typically the information in the charts is not shown to CAPT users. Normally, a CAPT system will use the results from the contrastive analysis to make a selection from all available pronunciation exercises dependent on the user's native language, without actually showing the analysis results. All that CALST requires of the user in order to perform a contrastive analysis is that (s)he specifies his/her native language and the target dialect when starting the program for the first time.

The use of IPA charts in L1-L2*map* makes it easy for language experts to define the segment inventory of his/her native language in L1-L2*map* (language expert privileges required). For each phoneme in the charts, a number of phonetic variants are shown when a cell in the charts (or phoneme symbol) is clicked, allowing the language expert to select the phonetic realization of the phoneme which best represents a "canonical" pronunciation in the language – for consonants usually the realization in the onset of a stressed syllable in an isolated word, and for vowels their pronunciation in isolation. L1-L2*map* is implemented as a wiki and can be incorporated into any CAPT system as a server-client system. The biggest practical challenge for the wiki is involving language experts for the many languages in the database to contribute with positional information for the consonants or to define the segment inventories for languages which are not yet available.

2.2. The pronunciation of Norwegian

Many languages have a pronunciation standard, although they may vary in the rigidity with which speakers use that standard. The choice of using a dialectal versus a more standard pronunciation will often depend on the formality of the situation. In contrast, Norwegian speakers use their dialect largely independently of social status and context, so that even the Prime Minister, for example, speaks to the nation in his own dialect. As a result of migration, different dialects are often spoken within the same area. This means that foreigners need to be able to deal with a variety of dialectal pronunciations.

To select one dialect as the target variant for production, i.e. as a role model, the user specifies the target dialect when starting CALST the first time. Since any learner of Norwegian will have to become familiar with the sometimes quite different pronunciation variants of *all* the different dialects in order to become communicatively effective, perception exercises in CALST offer the possibility of switching between dialects. The dialects are implemented as different talking heads on the right-hand side of the CALST window. For legibility of the text in the figures, these are not shown in the figures in this article.

2.3. Articulatory explanations

As a first step in the selection of the sound contrast exercises, the CALST user chooses a phoneme which is unfamiliar from his/her L1, e.g. the voiced palatal nasal consonant /ɲ/. After that, a sound contrast is selected, e.g. /ɲ-ŋ/, i.e. the contrast between the voiced palatal nasal and its velar counterpart. The list of sound contrasts is set up on the basis of experience collected from classroom practice.

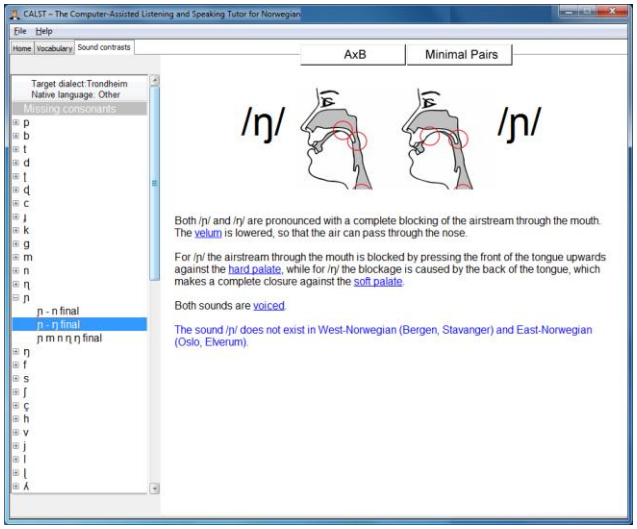


Figure 1: Explanatory text for the articulatory /ɲ-ŋ/ contrast

Drawn sagittal cross-sections for the contrasted sounds are shown at the top of the CALST window with circles to indicate place of articulation (here: palatal versus velar), manner of articulation (oral versus nasal, i.e. raised or lowered velum) and voicing (circle around the glottis), cf. Figure 1. A short articulatory description of the sounds is given for these three dimensions, always following the same pattern: first, the commonalities between the two sounds are described, and then the difference(s). If the sounds do not occur in all dialects of Norwegian, this is commented on at the end of the description. For vowels, the point of maximum constriction is indicated by a red circle.

Like L1-L2*map*, the articulatory descriptions follow IPA standards: For consonants, the dimensions manner and place of articulation and voicing are used, while vowels are described on the basis of degree of opening, the front-to-back dimension and lip rounding as well as length. Although Norwegian has long consonants in stressed syllables, the length is not phonemic and depends on the length of the preceding vowel for syllable-final consonants: Long vowels are followed by short consonants, and short vowels are followed by longer consonants. Long and short consonants thus occur in a complementary distribution. The length feature is not distinctive for consonants and therefore not important for communicative effectiveness.

2.4. Phonetic listening

CALST users will normally start with phonetic listening exercises to become familiar with a sound contrast (see Figure 2 for a single trial of a phonetic listening exercise). These exercises are implemented as ABX exercises, i.e. the user first listens to one word (e.g. [k^hɑtː], while the button for <katt> on screen is highlighted) and then to the other word in the minimal pair ([k^hɑʈː], button for <kart> highlighted). Then a third word is spoken by the tutor (implemented as a talking face, not shown in the figure) while the middle button on the screen is highlighted, and the user is required to click on a button with the text label for the word spoken, i.e. <katt> or <kart>. Note that the exercise is somewhat misleadingly called AXB on account of the visualization on the screen, where the middle button represents the *third* word (in real AXB the second word is the test word).

To prevent users from focusing on irrelevant acoustic properties to distinguish between the words, two recordings are available for each word, and the same recording is never used twice in a single trial. Irrelevant acoustic properties can be non-distinctive differences in fundamental frequency or even coincidental background noises, although great care was taken to prevent such differences in the recording of the words.

2.5. Abstract listening

After each ABX exercise, the learner is expected to take a minimal pair exercise. This exercise differs from the ABX exercises in that only one word is spoken by the tutor, and the user is required to click on one of two buttons on the screen, labeled with the words in the pair. Since the user only hears a single word and cannot rely on an acoustic comparison, (s)he has to use an internalized representation of the phonemes to decide which of the buttons to click. This exercise is therefore more advanced than ABX exercises.

After taking several related minimal pair exercises, for example contrasting sets of consonants which may be confused across L1's, such as the voiced and voiceless dental and retroflex plosives, the user can take *sound set* exercises. As in the minimal pair exercises, the user only hears a single word, and clicks on the corresponding button on the screen. In sound set exercises, the buttons are not labeled with words, but with sound symbols with the text "as in <WORD>" underneath (see Figure 3). The reason for labeling the buttons with sound symbols is that it is usually not possible to find minimal sets, i.e. word sets that are only distinguished by the phonemes which are the focus of the exercise. This also requires that the language learner listen to the sounds at an even higher level of abstraction, since the sounds occur in different contexts.

Segmental differences can thus be implemented in CAPT systems in a relatively straightforward manner on the basis of a "universal" contrastive analysis. The direct comparison of any L1 with the segment inventory of L2 is the basis for selecting exercises which are directly relevant for the user. It may be possible to extend this approach to prosodic-phonetic and other linguistic properties.

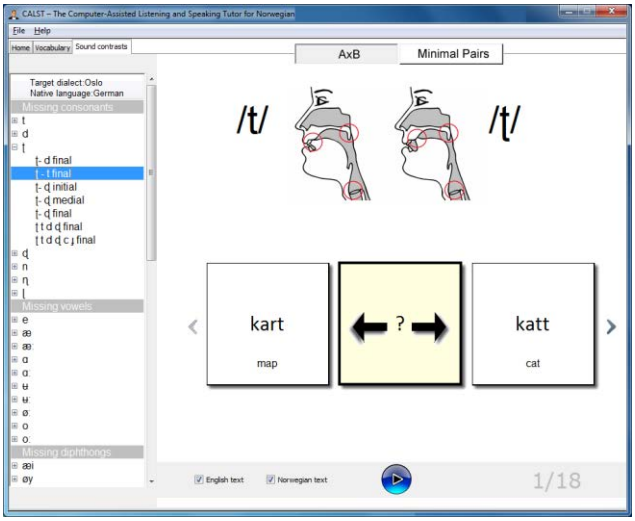


Figure 2: Single trial for an ABX exercise for the /t-ʈ/ contrast

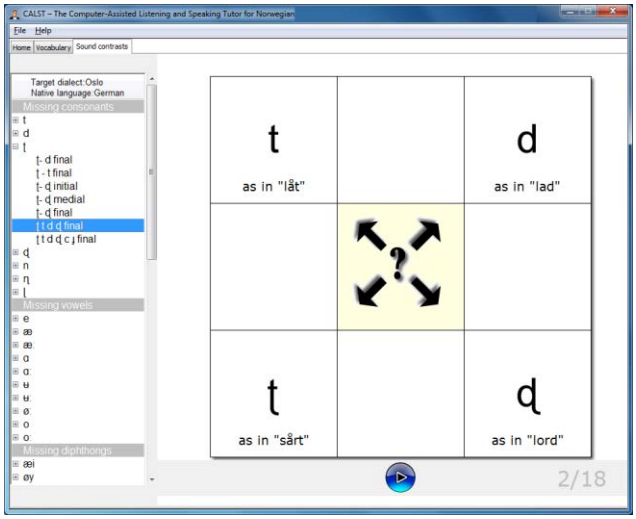


Figure 3: Single trial from a sound set exercise for /t-d-ʈ-ɖ/

3. Exercise selection

Presently, each unfamiliar phoneme in L2 gives access to a set of exercises in CALST in which that particular phoneme is contrasted with *several* phonetically similar phonemes in different syllable positions (in onset and coda, for consonants). When L2 learners learn a new sound, however, the errors they make depend on their L1, and learners with different L1's may substitute different sounds. This is called differential substitution. Since our aim is to offer an efficient learning trajectory, the coupling between the contrastive analysis and the selection of exercises should take this into consideration. Exercises for unfamiliar phonemes should only contrast these with the phoneme(s) learners are likely to use as a substitute.

Two different approaches can be used for selecting exercises on the basis of actually occurring (differential) substitutions. The first is based on a linguistic mapping of the phonemes in L2 onto the L1 system on the basis of the phonological features of the L1. In section 3.1, we shall discuss these phonological approaches, analyzing their usefulness for our purpose and presenting some partly speculative ideas which will have to be verified or falsified in future research. The second approach is based on observations of actual errors and is described in section 3.2. This latter approach is now being implemented in CALST as part of a user logging system.

3.1. Linguistic selection

Differential substitution is a well-known phenomenon in L2 learning. The differential substitutions for English /θ,ð/ have been investigated particularly extensively, and will be used as a basis for discussion. It is well known that these consonants, which are unfamiliar for many foreign learners of English, may be replaced by /t,d/, /s,z/ or /f,v/ depending on the L2 learner's native language (e.g. in Russian, German and Finnish, respectively). Several explanations have been offered in the literature as to why L2 learners make different substitutions even though they have similar phoneme inventories (at least with respect to the substituted phonemes). Particularly underspecification theory [14] and optimality theory [15] offer good explanations of the phenomenon. Both however must rely on external evidence for the underspecification patterns or the constraint rankings which explain the differential substitutions. This evidence can come from phonological processes in the L1 (including the phonological adaptations of loanwords to L1 phonology) or from substitution errors in foreign language learning. The latter makes the problem cyclic, since we want to use the underspecification pattern for *predicting* phoneme substitutions in the L2. Evidence from L1 phonology may solve the problem, but [14] suggests that such evidence is often absent. This would create a learning problem, since it is not clear how a child can learn the underspecification patterns for its native language, and by implication the substitutions in an L2 cannot be predicted on the basis of the L1 underspecification patterns.

In the explanations for differential substitutions which were mentioned above, we must have access to detailed knowledge of the L1 phonology in order to predict the possible misperceptions by L2 learners and their repair strategies when dealing with an unfamiliar sound in L2. As Weinberger writes, the segmental phonologies of languages do not always "contain the requisite rules to direct us in the construction of an optimal underspecified matrix" [14]. Therefore, "…while [the theory of underspecification] is fundamentally correct insofar as it simplifies the task of the first language learner, it has the problem of generating ambiguous matrices. That is, without native language evidence, multiple underspecified formulations are logically possible," and the ambiguous matrices can therefore not always predict which differential substitutions actually occur. With that reservation, incomplete or ambiguous underspecification matrices can still be used to predict the set of *possible* or *likely* L1-dependent substitutions. This can narrow down the possible substitutions, and thus constrains the number of exercises which an L2-learner with a given L1 needs to take. Clearly, generating underspecification matrices or OT constraint rankings for many languages is a laborious task and exceeds the scope of our present project.

Another method for the selection of sound contrast exercises may be based on an approach to generating underspecification matrices which makes claims to universality. In similarity with other underspecification- and OT-based explanations, the fully underspecified lexicon (FUL) approach to speech perception assumes different feature specifications for different languages [16]. However, beyond a basic set of contrasts (features) which is present in all languages, FUL claims that "[a]ll other features depend on the phonological systems of individual languages. The assignment does not depend on whether any feature is active in a phonological rule, but only if it is necessary to establish a phonemic contrast." This means that the set of underspecified features for any language can be defined solely on the basis of the sound contrasts occurring in the language. Since this would enable the generation of an underspecification matrix for each language solely on the basis of the phoneme inventories which CALST already uses (see Section 2.1), this makes the FUL model very compatible with the multi-language approach of CALST. This approach will be investigated further in the future.

Phonological explanations on the basis of the phoneme inventories of L2 and L1 are faced with yet another problem: In some languages the substitutions are known to be position-dependent, as are the substitutions for English /θ,ð/ by Dutch learners [17,3]. All substituted phonemes /t,d,s,z,f/ (and possibly /v/) can occur in all positions in Dutch, with the exception that only voiceless phonemes occur in final position due to final devoicing in Dutch. One must therefore either assume that there is a position-dependent phonological specification or assume a phonetic explanation.

A phonetic explanation of different substitutions could possibly be that some speakers of Dutch are so familiar with English that they are aware of phonetic variation of the phonemes in English. English dental fricatives can start with an occlusion in word-initial position, whereas they are generally continuant in other positions. This may affect the substitutions applied by L2 learners, although we are aware that this statement is very speculative. It is not unreasonable to assume familiarity of Dutch learners with the acoustic quality of English, because for example television series in English are not usually dubbed, but subtitled. This is not the case for instance in German, where we should therefore expect less variation in the observed substitutions.

To summarize, we point out that the problems intrinsic to ambiguous underspecification matrices apply equally to the incompleteness of an OT description of languages with respect to a ranking of all operating constraints. Although a linguistic approach to predicting differential substitutions is to be preferred because it allows us to select contrastive exercises solely on the basis of phoneme inventories, this is presumably not feasible within our CAPT approach, since it requires a complete

underspecified feature matrix or a complete OT constraint set for each L1 in the L1-L2*map* database which forms the basis for the CALST system. To be used for exercise selection within the multi-lingual approach of CALST, a linguistic solution would have to be universal. Since no such solution as yet exists, we opt for a more pragmatic approach.

3.2. Pragmatic selection

Projects like the Speech Accent Archive [17], in which foreign speakers are recorded and their substitutions are transcribed and categorized, represent a descriptive basis for differential substitutions in a given L2. Such databases also give interesting insights into the completeness of the substitutions. Hanulíková and Weber point out, for example, that the segments substituted for English /θ/ by Dutch and German speakers differ acoustically from the perceptually closest phoneme, although they are target-like and accepted as good exemplars of the specific categories of the substituted phoneme [4]. For the purpose of CALST, where we aim for (at least) communicatively effective pronunciation and perception, a phonemic representation of the substitutions suffices. In a small pilot project, we have therefore started collecting and transcribing foreigners' pronunciation of a short Norwegian text in which all Norwegian phonemes are represented in several contexts.

On the basis of observational data on actually occurring substitutions in L1-L2 pairs, it will be possible select exercises which are useful for learners with a specific L1 to train unfamiliar sounds from an L2, Norwegian in our case. Such a pragmatic approach to the selection of sound contrast exercises for L2 depending on L1 can make use of the CALST system itself. At the moment, CALST is being converted from a downloadable program into a web-based system where users' progress will be logged in a database. This is useful for the learner, since the system keeps track of the satisfactorily completed exercises and those (s)he still needs to take, and also (in the case where the system is used in combination with classroom teaching) for the teacher, who can monitor students' progress and errors, and use that as a basis for pronunciation teaching in class.

Information which is stored in the database will also reflect the differential substitutions which actually occur. With sufficient data for a given L1, only those exercises can be selected which train the user to hear distinctions which are difficult for an L2-learner to perceive. ABX or minimal pair exercises with no or few mistakes for a given L1 can be discarded. This will help to make the system more efficient for future L2-learners.

In CAPT systems for other languages than Norwegian the same strategy can be used, as long as the system has a similar structure to CALST.

4. Discussion

One can ask oneself whether it is pedagogically optimal to only offer training for challenges which foreign language learners are confronted with, since including exercises which are easy (e.g. sound contrasts which occur in both L2 and L1) may help to motivate the learner. Nevertheless, we have focused on a maximally efficient learning trajectory with sound contrast exercises for unfamiliar sounds from L2.

CALST users are directed to ABX (phonetic listening) and minimal pair/sound set exercises (abstract listening) for those speech sounds that are predicted to be problematical on the basis of a contrastive analysis (although each user does have access to the complete exercise set for all phonemes in the target language). These problematical speech sounds are practiced in exercises for all contrasts that can be relevant for learners with *any* L1 background. In order to limit the number of exercises a learner takes, a pragmatic solution will be adopted to the selection of exercises which only considers actually occurring differential substitutions. These are obtained from logged exercise results for learners with the same L1. Exercises where learners with the same L1 made no or few mistakes can be taken off the list of exercises for all learners with that L1.

Over time, the collection of data for each L1 will also allow us to set up an underspecification matrix for that specific L1, at least in as far as it is relevant for substitutions in Norwegian. We hope such matrices will also be relevant for the selection of exercises in other languages, where they can be used for a linguistic prediction of (some) substitutions. In this way, the approach may also contribute to a "universal", or at least multi-lingual, approach to exercise selection.

5. Acknowledgements

This work was financed in part by Norgesuniversitetet, project number P54/2009 and in part by VOX project 2010/59.

6. References

[1] The Computer-Assisted Listening and Speaking Tutor (CALST), http://calst.hf.ntnu.no, 2012.

[2] Odlin, T., Language Transfer. Cambridge University Press, 1989.

[3] Ellis, R., The Study of Second Language Acquisition. Oxford University Press, 1994.

[4] Hanulíková, A. and Weber, A., "Production of English interdental fricatives by Dutch, German, and English speakers", in K. Dziubalska-Kołaczyk, M. Wrembel, and M. Kul (eds.), Proceedings of the 6th International Symposium on the Acquisition of Second Language Speech, New Sounds 2010, Poznań (Poland): 173-178, 2010.

[5] McAllister, R., "Perceptual foreign accent: L2 users' comprehension ability", in A. James and J. Leather (eds.), Second-language speech: Structure and process, 119-132, 1997.

[6] Wik, P., The Virtual Language Teacher. Ph.D. thesis, KTH School of Computer Science and Communication, 2011.

[7] Husby, O., & Øvregaard, Å., Wik, P., Bech, Ø., Albertsen, E., Nefzaoui, S., Skarpnes, E. & Koreman, J., "Dealing with L1 background and L2 dialects in Norwegian CAPT", Proc. of the workshop on Speech and Language Technology in Education (SLaTE2011), Venice (Italy), 2011.

[8] Husby, O., Høyte, T. Nefzaoui, S., Nordli, I., Robbins, S. and Øvregaard, Å, An introduction to Norwegian dialects. Tapir Akademisk Forlag, 2008.

[9] Koreman, J., Bech, Ø., Husby, O. & Wik, P., "L1-L2*map:* a tool for multi-lingual contrastive analysis", Proc. 17th Int. Congress of Phonetic Sciences (ICPhS2011), Hong Kong, 2011.

[10] Maddieson, I., UPSID: UCLA phonological segment inventory database. Phonetics Laboratory, Department of Linguistics, 1980.

[11] J. Flege, "Second Language Speech Learning Theory, Findings, and Problems", in W. Strange (ed.), Speech Perception and Linguistic Experience: Issues in Cross-Language Research. Timonium, MD: York Press, 1995.

[12] Setter, J. and Jenkins, J., "State-of-the-Art Review Article", Language Teaching, 38(1): 1-17, 2005

[13] Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet. Cambridge: Cambridge University Press, 1999.

[14] Weinberger, S., “Minimal segments in second language phonology”, in J. Leather and A. James (eds.), New Sounds 90, Proceedings of the 1990 Amsterdam Symposium on the Acquisition of Second-Language Speech. Amsterdam: University of Amsterdam.
[15] Lombardi, L., “Second language data and constraints on Manner: explaining substitutions for the English interdentals”, Second Language Research 19(3): 225–250, 2003.
[16] Lahiri, A. and Reetz, H., “Distinctive Features: Phonological Underspecification in Processing”, J. of Phonetics, 38, S. 44-59.
[17] Weinberger, S. "Speech accent archive", http://accent.gmu.edu/. George Mason University, 2003.

Automatic Recognition of Vowel Length in Japanese for a CALL System motivated by Perceptual Experiments

Greg Short[1], *Keikichi Hirose*[1], *Nobuaki Minematsu*[2]

[1]Graduate School of Information Science and Technology, University of Tokyo, Japan
[2]Graduate School of Engineering, University of Tokyo, Japan
{short,hirose,mine}@gavo.t.u-tokyo.ac.jp

Abstract

Acquisition of the Japanese vowel length contrast can be problematic for non-native speakers. For these speakers, a CALL system which can automatically recognize vowel length could be of great benefit for pointing out their errors and issuing corrective feedback. However, a method that can adequately do this has not been proposed yet. Vowel length recognition is made difficult because the vowel length distinction is dependent on the surrounding vowel durations which vary due to speaking rate and other factors. Hidden Markov Models (HMMs), the standard way of recognizing this distinction, do not make use of this information. Methods have been proposed to recognize this in the past, but they do not appear viable unless knowledge about the durations of other vowels is present. Thus, we carry out perceptual experiments to gain more knowledge about the vowel length contrast. From this analysis, we develop an automatic recognition algorithm for vowel length that uses support vector machines (SVMs). We tested this method on a speaking rate corpus, native speech, and non-native speech. The method produced recognition results that are overall superior to HMMs and also more robust against speaking rate differences with an average of a 0.83 correct recognition rate for the 3 datasets. The error and non-error classification rates on non-native speech for this are 0.86 and 0.84 respectively.

Index Terms: Japanese, vowel length, recognition, speaking rate, perception, CALL

1. Introduction

There are two phonemic lengths for each vowel in Japanese: long and short. This distinctive feature is important for human word recognition and mistakes in it can reduce intelligibility and naturalness [1]. This contrast can be difficult for learners of Japanese to acquire [2], though, so there is a need to develop a CALL system with an algorithm that can automatically recognize this contrast.

In speech recognition, the typical way of recognizing this contrast is through the use of HMMs. The phonemic vowel lengths are recognized by using different HMMs for short vowels and their long counterparts. Despite being the method commonly used in automatic speech recognition, HMMs are not an ideal way to automatically differentiate between the two lengths. One reason for this is that HMMs are said not to be good at differentiating between items that are temporally different but spectrally similar [3].

Another problem, is that the decision boundary (perceptual boundary) between long and short vowels varies due to speaking rate. As speaking rate goes from fast to slow the durations of short vowels and long vowels increase and, thus, the vowel duration at the perceptual boundary increases as well. The frame-by-frame processing that HMMs employ will not take this information into account when classifying. Without integrating this into the model, if the learner does not speak at the same rate as the speech used for training the recognition model, there will likely be many vowels that are not errors misclassified as errors and vice-versa.

Thus, for a CALL system, a recognition method that can take these factors into account is desirable. Such a CALL system could have a flow like the one in Fig. 1. In this flow the phonemes for the text that the learner reads and the mic input are forced aligned, and from the phoneme alignments the vowel lengths are recognized. Based on these alignments, error classification and feedback generation can be carried out.

In this paper, we will focus on the vowel length recognition/error classification stages of this flow. Previous research for developing methods to carry out recognition for a CALL system has attempted to factor such features in, but the results do not appear satisfactory and cannot be applied in a simple manner. This will be further discussed in the Section 2. Because of these issues, more knowledge about perception of vowel length is necessary.

Taking the above into consideration, we have conducted perceptual experiments in order to better understand what vowel length classification depends on. After gaining more understanding of the mechanisms of human perception, we have developed an algorithm that is motivated by these perceptual experiments and makes use of SVMs with features motivated by the perceptual experiments.

In this paper, we discuss that algorithm and the basis for its development. We then test that algorithm on a speaking rate corpus using nonsense words and a corpus made up of native Japanese speech data, and non-native speech data. In Section 2, previous methods proposed for automatic recognition of vowel length will be discussed. In Section 3, the perceptual experiments we carried out to develop the method will be overviewed. In Section 4, the proposed method based on these listening tests will be overviewed. In Section 5 the recognition experiments will be discussed. In Section 6, the conclusion will be given.

2. Research on Automatic Classification of Vowel Length

There have been several methods proposed for automatically classifying vowels as short or long. The most common such method is the use of two HMMs to represent the two vowel lengths: a short vowel HMM and a long vowel HMM. This is the simplest and the most widely used method. The problems with this method, however, are that HMMs are not good at

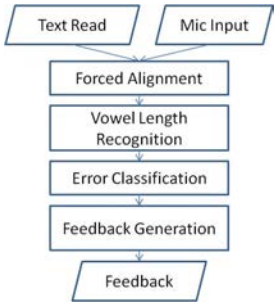


Figure 1: This figure shows the system flow to recognize vowel length and generate feedback for a CALL system.

distinguishing between phonemes that differ mainly in temporal structure and that it does not take into account effects from speaking rate.

In Kawai et al. [4], they employed listening tests for developing a method. In these tests, they used minimal pairs differentiated by the length of one vowel such as /to:ru/ (to pass) and /toru/ (to take). The stimulus sets were created by lengthening and shortening the vowel length that distinguishes the words of the minimal pair. Thus, in this case a continuum of stimuli from /toru/ to /to:ru/ was created. Then, they played each of these stimuli to native speakers having them select which word of the minimal pair the stimulus was. After obtaining the results from these tests, a logistic equation was obtained for each word by fitting the vowel duration/selection rate graph and this was used for classification. This method, however, did not take speaking rate into account when classifying.

Two other methods that attempted to take this into account were then proposed. One of these methods was carried out by Yamamoto et al [5]. In their research, they conducted listening tests like in Kawai's method. The difference was that they resynthesized the word length as well as the vowel length. First, the entire word duration was resynthesized to different durations. Then for each word duration, the target vowel was resynthesized to different durations to create continua from short to long for various word durations. From the perceptual experiments they conducted with this data, they derived an equation that predicted how the perceptual boundary of the sound would change due to the duration of the word and used this equation to recognize vowel length.

This can lead to problems, though. For one, it is not apparent if word duration is how humans carry out speaking rate normalization. Another problem is that it assumes that other vowels in the word were correctly produced. It is possible that other vowel lengths may have been incorrectly pronounced. For example, assume learner uttered /oji:saN/ (grandfather). For determining whether the /i:/ was mispronounced as /i/ the word duration is used. If the learner mispronounced the /o/ as /o:/, though, the word duration will be longer than usual and thus misclassifications can arise due to this so the word duration normalization calculation must take this into account. This is not simple, though, because all of the vowels could potentially be pronounced with the incorrect length. Thus, this does not appear to be a viable solution when all of the vowel lengths are unknown.

The other method that has been developed to handle speaking rate was proposed by Ishi et al [6]. They looked into using the inverse speaking rate (ISR) calculated by dividing the number of seconds by the number of morae as a means to automatically classify vowel length. However, this method has the problem that in order to calculate the ISR it is necessary to determine how many morae there are. In order to determine how many morae there are, though, it is necessary to classify the lengths of all the vowels. Thus, the inputs of the function require the outputs and this method is not easy to carry out.

3. Perceptual Experiments

3.1. Overview

In the previous section, we introduced four methods for automatically recognizing vowel length, three of which were for CALL systems. Two of the methods do not take into account variations due to speaking rate. The other two do not appear easy to apply if all vowel lengths are unknown, unless changes are made to the algorithms. Thus, a new approach to automatic recognition of vowel length is necessary.

For such an approach, better understanding of the perception of vowel length is needed. For this, listening tests like the ones in Kawai et al [4] and Yamamoto et al's [5] works can be used to further investigate the mechanism of perception.

Thus, we have carried out such perceptual experiments. In these experiments we have investigated how the duration at the perceptual boundary of a target vowel changes due to the durations of surrounding vowels. First, we recorded several nonsense words of the syllable structure CVCVCV in a soundproof room. Then, the durations of one or two context vowels or a context consonant is chosen to be manipulated. The context vowel(s) we manipulate is(are) manipulated to M durations creating M subsets. For the cases where two context vowels are manipulated, both of the context vowels have a different duration for each subset. Then for each duration of the context vowel(s) (each subset), the target vowel was manipulated N times, creating M x N stimuli.

After this listening test, the selection rates for the N stimuli of each of the M subsets are fit to a logistic curve to get M fits

$$P(VowelLength = Long) = \frac{1}{1 + e^{\alpha(td-\beta)}} \quad (1)$$

where td is the target vowel duration, β is the target vowel duration at the perceptual boundary and α is the slope at the perceptual boundary. By analyzing the change in β due to changes in the context vowel durations and preceding consonant duration, it can be understood which sound durations are important for a classification algorithm.

An online program was used to carry out the experiments. When this program initiates, a sample is played to a subject at random and two buttons are displayed. The target sound is shown in the top button with only the katakana character for the mora with the target vowel and on the bottom button with the katakana character for that mora along with a dash to indicate that it is a long vowel. All of the phonemes except for the target vowel are masked with '*' and displayed in both the top and bottom buttons. The subject was instructed to click the bottom button if he or she perceived the vowel as long and the top button if he or she perceived it as being short. A total of 90 native Japanese speakers participated in the experiments.

3.2. Stimulus Sets

In this paper, we discuss three different groups to see the effects of the durations of the surrounding vowels and preceding consonant. For the first group, we chose the middle vowel

of a three syllable word to be the target and the surrounding two vowels to be the manipulated context sounds to create the subsets. Both of these context sounds were lengthened for the creation of each subset and for each of the subsets the target vowel was manipulated 13 times. This was to see how manipulating the durations of two context vowels would affect perception. In the second group, one of the context sounds was manipulated and the other was held constant. This was to see how perception would change if only one surrounding vowel was manipulated for comparison with the case where two were manipulated. Lastly, in the third group, the preceding consonant was manipulated to various durations and the other sounds were held constant to determine if the consonant duration affected vowel length perception. For the second and third groups, for each length of the context sound, the target sound was manipulated 9 times. The details concerning the manipulations are given in Table 1. The context sounds that were not manipulated were set to roughly 100ms in duration for vowels and 80ms for consonants except for the first consonant which was set to be 40ms.

Table 1: *Manipulations for words for different sets.*

word	target	manipulated context	nonmanipulated vowel dur (s)
Bepisa	V2	V1 & V3	N/A
Takese	V2	V1 & V3	N/A
Bepisa2	V2	V1	0.2s
Zatogi	V2	V3	0.2s
BepisaC	V2	C2	0.15s

3.3. Perceptual Experiment Results

The results for the sets where the surrounding vowels were manipulated are shown in Fig 2. The top graph shows the sets for which both context sounds were manipulated. For this graph, both the left and right contexts are plotted. The set names suffixed by 'L' indicate that that plot is for the duration of the left context vowel (V1) duration. The names for the plots appended by 'R' indicate that that the x-values for that plot are for the right context (V3) duration. For the case of /bepisa/ the right context sound was shorter and in the case of /takese/ the left context sound was shorter. Despite this, the plots for the shorter of the two context sounds and the plots for the longer of the two context sounds overlap.

In the bottom graph, the duration of the shorter of the two context sounds for both of the nonsense words used in the previous test are plotted with the sets with only one manipulated context vowel. For those two words the other surrounding vowel was set to be 200ms since we wanted this vowel to remain longer than the other surrounding vowel for most subsets. In this case as well, there is an overlap for the target vowel length at the perceptual boundaries of the four different sets. For the two sets in which one of the surrounding vowels was not manipulated, it appears there is a peak in the vowel duration at the perceptual boundary as the manipulated context duration gets longer than the non-manipulated context sound. This would also agree with the idea that the shorter vowel is what is being used for classification, since it should reach a peak as the manipulated surrounding vowel approaches the duration of the other surrounding vowel. We have conducted more experiments related to vowel length perception in [7] which also showed that

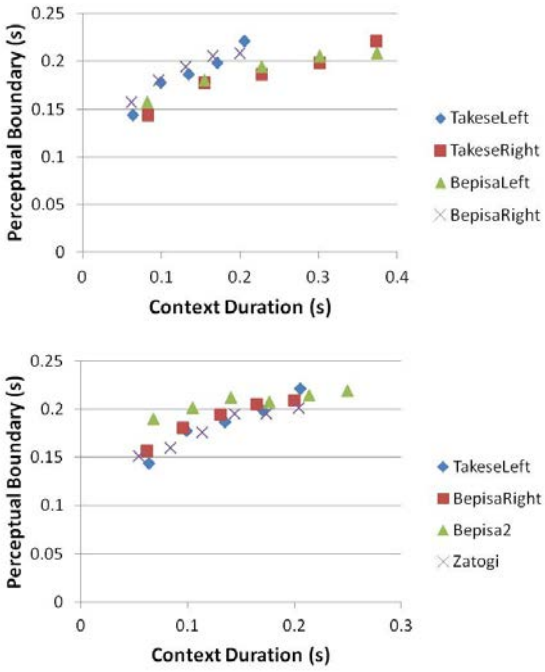


Figure 2: Top: Results for the vowel length experiment. Two different sets are compared in this graph, 'Takese' and 'Bepisa'. The durations of the vowel to the left of the target vowel are listed under the plots appended by 'left', and the ones to the right of the target vowel are appended by 'right.' Bottom: These results compare four different sets including 'Takese' and 'Bepisa' from the top graph. For 'Takese' and 'Bepisa' the plot corresponding to the shorter of the manipulated context sounds is given. For 'Bepisa2' and 'Zatogi', the manipulated plot for the change in the length of the target vowel at the perceptual boundary with the manipulated context sound duration is given.

as the manipulated context vowel becomes longer than the one that was not manipulated, the target vowel duration at the perceptual boundary peaked.

The results from the set where the consonant was the manipulated context sound can be seen in Fig. 3. In this figure, the x-axis indicates the duration of the consonant and the y-axis the duration of the vowel at the perceptual boundary. We carried out a linear fit on this data and found the coefficient of determination (r^2) between the consonant duration and the duration of the vowel at the perceptual boundary to be 0.06. This means that perhaps the consonant duration is not essential for classification.

4. Vowel Length Recognition Proposal

4.1. Background of Proposal

In the results of the previous section, it appeared as though the duration of the shorter of the two context vowels is more important in determining whether or not a vowel is perceived as long or short. Also, it did not appear as though the preceding consonant played a role in its perception. Based on those results, we present a vowel length recognition algorithm in this section along with experiments conducted using it.

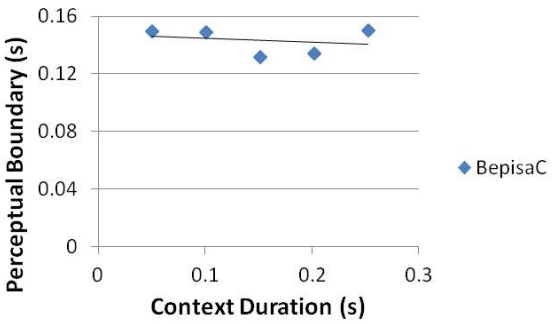


Figure 3: This figure shows the relationship between the length of the preceding consonant and the perceptual boundary. In these results, only a very small correlation of determination was observed by performing a line fit.

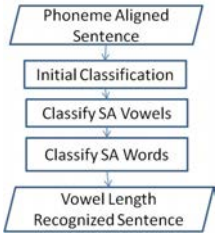


Figure 4: This flow chart shows the process flow for vowel length recognition. The inputs are the results from forced alignment shown in the system flow in Fig. 1

4.2. Details of Proposal

For considering how to develop an algorithm, the important thing to note is that it appeared as though the shorter of the two surrounding vowels was causally related with the changes in the perceptual boundary. We propose to perform recognition centered around this idea. The recognition flow we propose is given in Fig. 4. This takes the phoneme alignments for a sentence as the inputs and outputs the sentence with the recognized vowel lengths. The output is passed along to the error classification stage given in Fig. 1.

The first stage of this algorithm classifies all of the vowels as being longer than, shorter than, or the same length as the surrounding vowel it is compared to. In the second stage, all of the vowels classified as being the same length are reclassified as long or short based on how the surrounding vowels are classified. If all of the vowels for a particular word are classified as being the same length, in the third stage these vowel lengths will be classified based on the durations of the vowels of the surrounding words.

4.2.1. Initial Classification

In the first stage, each vowel is recognized as shorter than (S), longer than (L), or the same length as (Sa) the surrounding sounds used in classification. At this stage, these are classified with 3-class SVMs. Since vowel length is a lexical feature, this is performed at the word level so if the vowel to the left or right is a separate word, it is not used at this stage. Ideally, the vowels for the sentence 'watashiwa/isogashi:' (I am busy) will be recognized as 'Sa-Sa-Sa-Sa/Sa-Sa-S-L' at this stage.

4.2.2. SA Vowel Classification

The above stage leaves us with three relative vowel lengths. In the second stage the vowel lengths labeled as Sa are classified as S or L for the words where one or more of the vowels were classified as S or L. If a word does not contain a vowel that has been labeled L or S, this stage is skipped for that word. In this stage, each vowel labeled as Sa next to a vowel that is not labeled as Sa is relabeled to be the same length as the shorter of the surrounding vowels. This process is repeated until all vowel lengths have been relabeled. After completing this process the vowels of 'watashiwa/isogashi:' should ideally be labeled as 'Sa-Sa-Sa-Sa/S-S-S-L'.

4.2.3. SA Word Classification

The above stage leaves the vowels for words which only have Sa length vowels unchanged. At the third stage, words that contain only Sa vowels are relabeled. The vowels for the first word next to a word that does not contain any Sa vowels are recognized. If the word comes after the word without any Sa vowels (non-Sa word), the vowel duration of the last vowel of the non-Sa word is used as a feature to classify the length of the first vowel of the Sa-word and if it comes before, the duration of the first vowel of the non-Sa word is used to classify the length of the last vowel as in the first stage. After classifying, the length of the classified vowel will be Sa, L, or S. If it is Sa, it is relabeled to be the same length as the vowel that was used to classify it. Then, the lengths for the rest of the vowels for that word can be classified as in stage two. This process is repeated until there are no Sa words. After completing this stage, the vowel lengths should ideally be 'S-S-S-S/S-S-S-L' for 'watashiwa/isogashi:'. If all of the vowels for all of the words are Sa, the number of L vowels and S vowels for the word in the canonical spelling are counted. The vowels of the utterance are chosen to have the length which is more prevalent in the canonical spelling. The reason for doing this is because the perceptual vowel length is relative. Since this method uses a relative approach to recognition, differences in speaking rate should be accounted for.

4.2.4. Features Used

For the features, we try a variety of feature sets in addition to simply using the duration of the shorter surrounding vowel. These features are listed in 2. 'Target' indicates the vowel to recognize. The nasal duration used is the duration of the moraic nasal /N/ that sometimes follows vowels in Japanese. If it is not present, a duration of 0 is given. For 'BothSides' and 'BothSides2', if one of the vowel sounds is not present, a value of -10 was given for the duration.

Table 2: *Feature sets used for recognition. If not otherwise specified, the vowel/nasal features are the duration of the syllables vowel/nasal.*

Feature Set	Features Used
Shorter	shorter vowel, target
ShorterNorm	shorter normalized by target
ShorterNasal	shorter, target, shorter and target nasals
ShorterFormant	shorter, target typical shorter and target formant values
BothSides	left vowel, right vowel, target
BothSides2	2 vowels to left, left vowel, right vowel target

4.3. Training Data

Since we would like to build a system that is fairly robust against speaking rate, for training and testing we have constructed a small corpus consisting of 5 native Japanese speakers. This corpus consists of a variety of nonsense words read at three different speaking rates: slow, medium, and fast. We created a program to automatically generate these nonsense words. With this program, first, the number of syllables for the word was chosen. Then, for each syllable, a syllable type was randomly chosen from the syllable types that exist in Japanese (i.e. V, CV, VN, VQ, CVN, CVQ and their counterparts with long vowels except for CV:Q). Then, for each sound in the syllable, a phoneme was randomly chosen. We created words with 2,3,4, and 5 syllables. Also, for half of the data we placed syllable type constraints allowing only CV, V, CV:, and V: type syllables. The recording was conducted in a soundproof room.

5. Vowel Length Recognition Experiment

5.1. Experiment Conditions

For preprocessing, we conducted forced alignment on all of the utterances in the corpora using Julius [8]. For the initial recognition stage, we used SVMs trained with libsvm [9]. We used RBF kernels with parameters selected with 5-fold cross validation.

For our recognition experiments, we tested the method on three speech datasets. First we conducted speaker/vocabulary-item open tests training with 4 of the speakers from the speaking rate data and testing on the remaining one. The speaking rate corpus consists of 2044 randomly generated isolated words and 522 carrier sentences ('kokoga___da' - 'this place is ___') containing a subset of those randomly generated words. For the speaking rate corpus, 1644 of the isolated words and 450 of the carrier sentences were used for testing (a total of 4 speakers). The remaining speaker's data was solely used for training. Next, we conducted tests on the non-native Japanese database that we have constructed [10]. This database consists of the speech of 4 native speakers and 27 non-native speakers who speak 19 different languages as L1s (first languages). For recognition of the vowel lengths for the corpus, we trained SVMs using data from all 5 speakers of the speaking rate database. For the non-native speech, the vowel lengths for a total of 182 isolated words and 46 sentences were recognized and for the native speech, 832 isolated words and 29 sentences.

5.2. HMM-based Recognition for Comparison

For comparison, we also carried out HMM-based recognition. To do this we performed semi-forced alignment. For semi-forced alignment, all of the phonemes are fixed as in normal forced alignment, but the vowel lengths are not fixed. Thus, all vowels are permitted to be long or short in conducting alignment. We do not compare the proposed method to the other previous methods we introduced since it is not clear how to use them if all vowel lengths are unknown.

5.3. Labeling

For the speaking rate and native datasets we assumed that the vowel lengths produced by natives would be perceived at the length the speaker intended them to be. Thus, for alignment the transcript given for the native speaker to read was used for automatic alignment. For non-native speakers, there are errors in the pronunciation, though, so we had 7 native speakers of Japanese label each vowel length in every utterance as short or long. The vowel length label that made up the majority was chosen as the vowel length for that particular vowel.

5.4. Results and Discussion

The results for the experiment are given in Table 3. From these results it can be seen that the methods based on 'Shorter' overall perform better for all cases with 'Shorter' performing the best. This agrees with our results from the perceptual experiments from which it appeared that the shorter surrounding vowel duration is important for human classification. In all cases, the methods using the lengths of both sides performed lower. This can be thought to be due to it using information that is not pertinent to the recognition process. Also, the 'ShorterNorm' method did not perform as well as the method that used the actual lengths. This is probably either due to a) linear normalization not being appropriate, or b) errors in alignment.

The HMM-based method did not perform well on the speaking rate corpus and its performance was noticeably lower than the 'Shorter' based methods for the other two datasets. This can be thought to be HMMs performing recognition in a more absolute manner and not taking into account the duration information of the surrounding vowels. This means that if the speaking rate is near to what the average speaking rate was for corpus, this method should perform fairly well, but it will break down otherwise.

In Fig. 5 the results for the 'Shorter' method and HMM-based method are given at various speaking rates for the speaking rate corpus at different percentiles of the Gaussian distribution. To calculate the Gaussian distribution, the morae/second speaking rate was calculated for each utterance. The x-axis indicates the speaking rate percentiles for the sample Gaussian distribution, the different methods, and vowel lengths. The y-axis gives the recognition rate. From these results it can be seen that the proposed method overall performs better than the HMM-based method for a variety of speaking rates. The performance of the proposed method drops, however, for long vowel recognition at the fastest speaking rate, 81-100%. The short vowel recognition rate at slow speaking rates does not have this degradation, though. For the HMM-based method, it can be observed that as the speaking rate gets slower the recognition rate drops for short vowels and as it gets faster it drops for long vowels. This shows a higher-level of robustness for the proposed method.

We also analyzed the capabilities of 'Shorter' for error classification on the non-native corpus to ensure that errors and non-errors were being adequately classified so that it could be used for a CALL system. These results are given in Table 4. From these results, it can be seen that correct classification of errors is roughly equal to that of non-errors and both are fairly high, showing sufficient results for a CALL system.

This method recognizes short vowels robustly at all speaking rates, but this method does not perform as well for long vowels at a very fast speaking rate. At very fast speaking rates, alignment errors in time will be a larger percentage of the vowel duration and thus make automatic vowel length recognition more difficult. Thus, it is likely that mistakes in automatic alignment are a large contributing factor to the less robust performance at fast speaking rates for this method. It is also possible that other features should be accounted for as well such as intensity and pitch, which has been said to have some effects on vowel length perception [11].

Table 3: *Recognition rates for vowel length experiment. The columns represent the different feature sets/algorithms that were used for conducting the recognition with SVMs plus SpeechRec, the semi-forced alignment results. Rates that are shown in bold indicate that the average of the S/L pair for that Feature Set/Method was the highest.*

		methods						
		Shorter	Shorter Norm	Shorter Nasal	Shorter Formant	Both Sides	Both Sides2	HMM
SR	S	**0.89**	0.77	**0.89**	**0.89**	0.9	0.90	0.68
	L	**0.78**	0.83	**0.78**	**0.78**	0.73	0.74	0.72
Native	S	**0.91**	0.65	0.89	0.91	0.92	0.90	0.75
	L	**0.79**	0.85	0.8	0.76	0.72	0.75	0.77
Nonnative	S	**0.86**	0.56	0.84	0.84	0.77	0.81	0.73
	L	**0.76**	0.89	0.75	0.76	0.67	0.70	0.78
Average	S&L	**0.83**	0.76	**0.83**	0.82	0.79	0.8	0.74

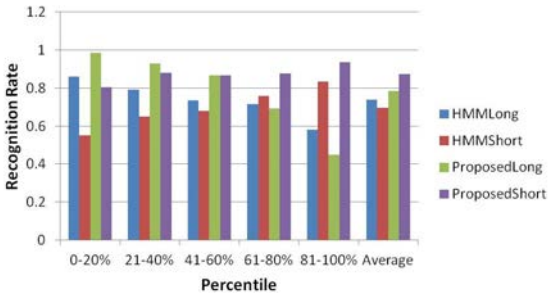


Figure 5: This figure shows the vowel length recognition for different speaking rates. The x-axis indicates the different different methods used, the vowel length, and the speaking rate percentile.

Table 4: *Confusion matrix for Non-error/Error Classification on non-native speech for proposed method*

	Non-error	Error
Non-error	0.84	0.16
Error	0.14	0.86

6. Conclusion

Japanese language learners have a difficult time in acquiring the vowel length contrast in Japanese. To assist them, we have developed an algorithm that automatically recognizes vowel length. Because the perceptual boundary changes due to the length of the surrounding vowels, this is not a straightforward problem. Thus, in order to solve this problem we conducted perceptual experiments. In those experiments it appeared as though the shorter of the surrounding vowels was important for recognition. Motivated by this, we created a novel algorithm to automatically recognize vowel length which outperformed a standard HMM-based method on three different databases and showed a higher degree of robustness against speaking rate. The error/non-error classification capabilities for non-native speech were also good. It appears sufficient for use in a CALL system.

While recognition of short vowels was fairly robust across speaking rates, recognition of long vowels was not very robust at fast speaking rates. In the future we would like to find ways to remedy this problem with perceptual experiments and by improving alignment. We would also like to explore using features such as pitch. We also plan to integrate this method into a CALL System with a feedback generation interface.

7. References

[1] C. Tsurutani, "Foreign accent matters most when timing is wrong," *In INTERSPEECH*, pp. 1854–1857, 2010.

[2] T. Toda, "Perceptual categorizationn of the durational contrasts by japanese learners," *Tsukuba University Linguistics Repository*, vol. Vol. 33, pp. 65–82, 1998.

[3] H. Strik, K. Truong, F. de Wet, and C. Cucchiarinia, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, pp. 845–852, 2009.

[4] G. Kawai, "Spoken language processing applied to nonnative language pronunciation learning," *PhD dissertation*, vol. 1, 02 1999.

[5] M. Yamamoto and J. Miwa, "Computer assisted learning system for japanese special mora and its evaluation," *The Acoustical Society of Japan*, pp. 1–8, 2000.

[6] C. T. Ishi, K. Fujimoto, and K. Hirose, "Identification of Japanese "tokushuhaku" regarding the influence of speaking rate." *Technical Report of IEICE*, vol. Vol. 100, no. No. 97, pp. 17–24, 2000.

[7] G. Short, "Perceptually-motivated Automatic Error Classification for Japanese Lexical Prosody in Non-native Speech," *PhD dissertation*, 2013.

[8] A. Lee, T. Kawahara, and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine." *In Proc. European Conference on Speech Communication and Technology (EUROSPEECH)*, pp. 1691–1694, 2001.

[9] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[10] G. Short, "Non-native corpus," 2012. http://www.gavo.t.u-tokyo.ac.jp/ short/corpus/.

[11] I. Takiguchi, H. Takeyasu, and M. Giriko, "Effects of a dynamic f0 on the perceived vowel duration in japanese," *Speech Prosody*, pp. 1–4, 2010.

Speaker-based Accented English Clustering Using a World English Archive

H.-P. Shen[1,2], N. Minematsu[2], T. Makino[3], S. H. Weinberger[4], T. Pongkittiphan[2], C.-H. Wu[1]

[1] National Cheng Kung University, Tainan, Taiwan
[2] The University of Tokyo, Tokyo, Japan
[3] Chuo University, Tokyo, Japan
[4] George Mason University, Virginia, USA
[2]{happy,mine,bank}@gavo.t.u-tokyo.ac.jp, [3]mackinaw@tamacc.chuo-u.ac.jp,
[4]weinberg@gmu.edu, [1]chwu@csie.ncku.edu.tw

Abstract

English is the only language available for global communication. Due to the influence of speakers' mother tongue, however, those from different regions often have different accents in their pronunciation of English. The ultimate goal of our project is automatic creation of a global pronunciation map of World Englishes on an individual basis, for speakers to use to locate similar English pronunciations. Creating the map mathematically requires a matrix of pronunciation distances among all the speakers considered. Our previous study proposed a good algorithm for that purpose [1], where, using phonetic reference pronunciation distances calculated from labeled data, a pronunciation distance predictor was trained and built for unlabeled data. Due to space limit in [1], the procedure for calculating the reference distances was not described in detail. Then in this paper, detailed descriptions are given and 498 world-wide native and non-native speakers in the Speech Accent Archive [2] are clustered using the phonetic reference distances. Results show high validity of using the calculated distances as reference distances for training a distance predictor.

Index Terms: World Englishes, IPA transcription, DTW, Speech Accent Archive, phonetic pronunciation clustering

1. Introduction

English is the only language available for global communication and it is true that English communication is done quite often between non-native speakers in international occasions. Due to the influence of the speakers' mother tongue, those from different regions inevitably have different accents in their pronunciation. Recently, more and more users of English accept the concept of World Englishes [3,4,5,6] and they regard US and UK pronunciations as just two major examples of accented English. Diversity of World Englishes is found in various aspects of speech acts such as dialogue, syntax, pragmatics, lexical choice, pronunciation etc. Among these kinds of diversity, this paper focuses on pronunciation. If one takes the philosophy of World Englishes as it is, he can claim that every kind of accented English is equally correct and incorrect. In this situation, there will be a great interest in how one type of pronunciation is *different* from another, not in how that type of pronunciation is *incorrect* compared to US or UK pronunciation. As shown in [7], the intelligibility of spoken English depends on the nature of the listeners as well as that of the speaker and the spoken content, and foreign accented English can indeed be more intelligible than native English. Generally speaking, speech intelligibility tends to be enhanced among speakers of similarly accented pronunciation.

The ultimate goal of our project is automatic creation of a global map of World Englishes on an individual basis, for a speaker to use to locate similar Englishes and to find where his pronunciation is located in the diversity of English pronunciations. If the speaker is a learner, he can then find the best and easiest-to-communicate English conversation partner. A learner can also know how his pronunciation compares to other varieties. If he is too distant from these other varieties, he may have to correct his pronunciation for the first time to achieve smoother communication with these others. In real-world application, the global but individual pronunciation map may be popularized to the world of international business. Here, people often encounter new types of accented English pronunciation, some of which may be very problematic and cause some miscommunication. With this map, however, one can know in advance how his pronunciation is different from his new business partner's. He may find his colleague whose pronunciation is similar to that partner's and ask the colleague for help.

For our project, however, we have two major problems. One is collecting data and labeling a part of them, and the other is creating a good algorithm of automatically drawing the global map for a huge amount of unlabeled data. Luckily enough, for the first problem, the fourth author has made a good effort in systematically collecting World Englishes from more than a thousand speakers from all over the world. This corpus is called the Speech Accent Archive (SAA) [2], which provides speech samples of a common elicitation paragraph with their narrow IPA transcriptions. The technical challenge in the second problem is that we need an algorithm that can focus exclusively on pronunciation differences between speakers by ignoring irrelevant differences such as those in age, gender, vocal tract length, etc. In our previous study [1], by using reference pronunciation distances calculated based on the IPA transcriptions, we built a pronunciation distance predictor using invariant pronunciation structure analysis. The invariant structure analysis was proposed in [8][9] inspired by Jakobson's structural phonology [10] and it can extract very robust features. The structural features were already introduced to various tasks such as pronunciation scoring [11], pronunciation error detection [12], language learners clustering [13], dialect analysis [14], automatic speech recognition [15,16], and speech synthesis [17]. In our previous study [1], our pronunciation distance predictor outperformed by far a baseline system that was built with a conventional HMM-based phoneme recognizer. Due to space limit in [1], however, the procedure for calculating reference distances was not described in detail. In this paper, detailed descriptions are given and 498 world-wide speakers in the Speech Accent Archive are clustered using the phonetic reference distances. For comparison between two IPA transcriptions, we adopt the DTW algorithm and the obtained alignment gives us a phonetic distance between them. For DTW, a phone-to-phone distance matrix is required and this is obtained through acoustic analysis of an expert phoneti-

cian's productions of all the IPA phones with/without a diacritic mark. It should be noted that pronunciation diversity of World Englishes is found in both segmental and prosodic aspects. In our previous study [1], reference distance was obtained by calculating distance between a pair of IPA transcriptions. This means that the reference distance in [1] ignored the prosodic diversity because IPA transcription gives us only phonetic information of a given utterance. We do not claim that the prosodic diversity is minor but, as will be shown in the current paper, it seems that the clustering of English users only based on the segmental aspect can still present visually and validly how World Englishes are diverse in terms of pronunciation.

This paper is organized as follows. In the following two sections, we describe the SAA corpus and how to estimate phone-to-phone distance information by acoustic analysis. In section 4, we explain how to estimate inter-speaker distances by using the DTW algorithm. Some results of speaker-based pronunciation clustering are presented in section 5. In section 6, a distance predictor constructed using the above inter-speaker distances is briefly introduced and its performance of inter-speaker distance prediction is shown. In section 7, this paper is concluded and future directions are also presented.

"Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station."

[pʰl̩iːz kʰɑːl stɛlʌ æsk hɚ ɾə bɹɪ̃ŋ ðiːz θɪ̃ŋz wɪθ hɛɹ fɹʌ̃m ðə stoːɹ sɪks spu̟ːnz əɣ fɹɛʃ snoʊ pʰiːz̩ faːɣ θɪkˋ slæːbz ə blu̟ː ʧiːz̩ ɛ̃n meɪbi ə snæ̰kˋ foɹ hɚ bɹʌðə bɑb wi ɑlso nu̟ːd ə smɑːl pʰl̩æstɪk sneɪkˋ æ̃n ə bɪgˋ tʰɔɪ fɹʊg fʊɹ ðə kʰɪɪdz̩ ʃi kʰɪ̃n sku̟ːp ðiːz̩ θɪ̃ŋz̩ ɪ̃ntʰu̟ θɹi ɹɛ̟d bæːgz æ̃n wi wɪl miːt hɛɹ wɛ̃ntsdi æt ðə tɹeɪ̃n steɪʃɪ̃n]

Fig. 1 The elicitation paragraph used in the SAA and an example of detailed IPA transcription with diacritic marks

2. Speech Accent Archive

The corpus is composed of read speech samples of more than 1,700 speakers and their corresponding narrow IPA transcriptions. The speakers are from different countries around the world and they read a common elicitation paragraph, shown in Fig. 1. It contains 69 words and can be divided into 221 phonemes by referring to the CMU dictionary [18]. Each sample has its narrow IPA transcription, which was provided by trained phoneticians, and an example is also shown in Fig. 1. The transcriptions will be used to calculate reference inter-speaker phonetic distances. Use of read speech for clustering is considered to reduce the pronunciation diversity because read speech will show us only "controlled" diversity. In [19] however, English sentences read by 200 Japanese university students showed a very large diversity in terms of pronunciation and [20] showed that the intelligibility of the individual utterances to American listeners covered a very wide range. Considering these facts, we considered that clustering of read speech samples can still capture well how diverse World Englishes are in their pronunciation. In the current study, only the data with no word-level insertion or deletion were used. The speakers' files that had exactly 69 words were automatically selected as candidate files and then, 515 files were obtained. However, the word order in some files were found to be wrong and we manually removed them. At the end of the day, 498 speakers' data were obtained and used in our study.

3. Phone-to-Phone Distance Estimation using Acoustic Analysis

In this study, the DTW algorithm is applied to compare two speakers' IPA transcriptions. Since the algorithm needs a distance matrix among all the existing IPA phones in the archive, we prepared the distance matrix firstly. In this paper, phone-to-phone distance was calculated through comparing acoustic characteristics of the two phones, which were produced by an expert phonetician. Before recording, we calculated frequency of each of the IPA phones, many of which were with a diacritical mark, and extracted the kinds of IPA phones that covered 95% of all the phones found in the archive. The resulting number of the kinds of the phones with/without a diacritical mark was 153. Table 1 shows the 153 phones. One expert phonetician, the third author, was asked to pronounce each of these phones twenty times. Here, he was asked to pay good attention to diacritical difference within the same kind of IPA phone. In the recording, the phonetician pronounced each vowel twenty times. For consonants, a consonant was succeeded and preceded at the same time by vowel [a]. For example, in order to collect data of phone [p], the phonetician spoke [apa] twenty times. In this way, each consonant was recorded.

Using the wav files and its phonetic transcription, a three-state HMM was built for each phone, where each state s_i (i ϵ1, 2, and 3) contained a single Gaussian distribution with mean vector M_{Si} and covariance matrix P_{Si}. Here, MFCC(1-12) and its derivatives were used as acoustic features.

After training an HMM for each kind of the phones, the Bhattacharyya distance (BD) was calculated between two corresponding states of every phone pair. The equation of the BD between s_i of phone x and s_i of phone y is denoted below.

$$D_B(p_{Si}^x, p_{Si}^y) = \frac{1}{8}\left(M_{Si}^x - M_{Si}^y\right)^T P^{-1}\left(M_{Si}^x - M_{Si}^y\right) + \frac{1}{2}\ln\left(\frac{\det P}{\sqrt{\det P_{Si}^x \det P_{Si}^y}}\right) \quad (1)$$

where M_{Si}^x and M_{Si}^y are mean vectors and P_{Si}^x and P_{Si}^y are covariance matrices of state s_i of x and state s_i of y, respectively. Note that $P = (P_{Si}^1 + P_{Si}^2)/2$.

For each phone pair, three Bhattacharyya distances were calculated, each corresponding to a state-to-state distance. By accumulating the distances and averaging them, we defined the acoustic distance between the phone pair. Equation 2 shows distance definition between two phones x and y.

$$d_{p_x p_y} = \sqrt{\frac{D_B(p_{S1}^x, p_{S1}^y) + D_B(p_{S2}^x, p_{S2}^y) + D_B(p_{S3}^x, p_{S3}^y)}{3}} \quad (2)$$

We note here that, since the HMMs were trained in a speaker-dependent way, all the distances were calculated in the same and matched condition.

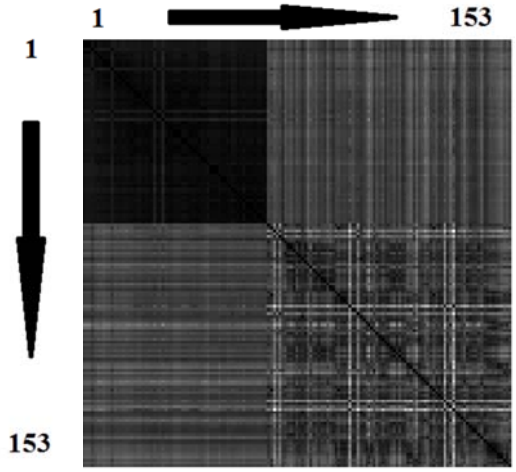


Fig. 2 The phone-to-phone distance matrix in gray-level image

The other 5% phones, which were not pronounced by the phonetician, were all with a diacritical mark and they were of very low frequency. For these phones, we substituted the HMMs of the same phones with no diacritical mark. With this substitution policy, the inter-phone distances among all the existing kinds of phones in the archive can be finally estimated. We converted the 153x153 phone-based distance matrix to its tree diagram. Although we do not show the diagram in this paper due to limit of space, the diagram confirmed us that we obtained a phonetically valid distance matrix. Fig. 2 shows a gray-level image of the 153x153 distance matrix instead.

In Fig. 2, X-axis and Y-axis denote the ID of phones (See table 1 in Appendix). Each pixel represents the distance between two phones. The first 66 phones are vowels and the others are consonants. The darker a pixel is, the more similar the phone pair are. From this figure, it can be seen that the distances between vowels are smaller than those between consonants or between a vowel and a consonant. We can also know that the distances between consonants have higher variance than those between vowels and those between a vowel and a consonant. Some small squares aligned in the diagonal line can be found in the figure because phones of the same kind with different diacritical marks are aligned together. Elements on this phone-to-phone distance matrix will be used as local distance or penalty to calculate the inter-speaker distance through the DTW alignment of two IPA transcriptions.

4. Dynamic Time Warping using Phone-to-Phone Distance Information

In this section, the DTW is done to compare every two IPA transcriptions in a word-by-word manner by using the distance matrix obtained above. The obtained DTW alignment gives us an accumulated distortion score, which will be used as reference pronunciation distance between the two speakers. This speaker-to-speaker phonetic distance can be used in automatically clustering speakers in terms of pronunciation. Since all the transcriptions contain exactly 69 words, word-level alignment is easy and we only have to deal with phone-level insertions, deletions, and substitutions between a word and its counterpart in the two transcriptions. The local and allowable path of the DTW used in this section is shown as Fig. 3.

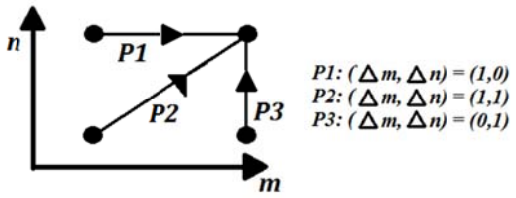


Fig. 3 Allowable paths of the DTW

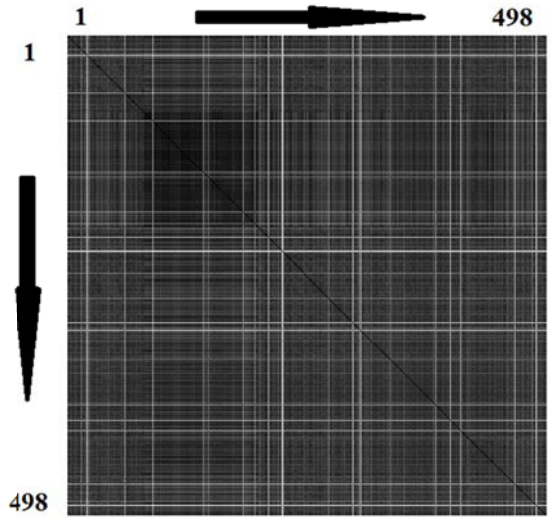


Fig. 4 The inter-speaker distance information matrix in gray-level image

P1, *P2* and *P3* are allowable paths of insertion, match and deletion. Path selection is done based on equation 3.

$$DTW[m, n] := minimum(\ DTW[m-1,n] + phone\_dist[m,n],\ DTW[m-1,n-1] + 2*phone\_dist[m,n],\ DTW[m,n-1] + phone\_dist[m,n]\ ) \quad (3)$$

DTW[m, n] is the current accumulated cost at position *(m,n)* and *phone_dist[m,n]* is a distance between the phone of time *m* and the phone of time *n*. Out of P1, P2, and P3, the path of which the accumulated cost at *(m,n)* is the minimum is selected. After normalizing this score by the total number of times of distortion accumulation, we can get a word-based distortion score. The 69 word-based scores are summed to be the final score for two given IPA transcriptions (speakers).

5. Speaker-based Pronunciation Clustering

After obtaining the inter-speaker distances, all the speakers can be clustered using Ward's method, one of the hierarchical clustering methods. Since the clustering result of the 498 speakers is too complicated, we firstly show the gray-level image of the distance matrix of the 498 speakers in Fig.4.

In the gray-level image, X-axis and Y-axis denote the ID of speakers. The IDs of speaker are assigned based on the alphabetical order of their countries' name. Each pixel represents the distance between two speakers. We can find a darker square in the top left. The distances in this region are from between native speakers of American English and this means that they have similar and stable English pronunciations. For non-native speakers, larger distances tend to be found to native

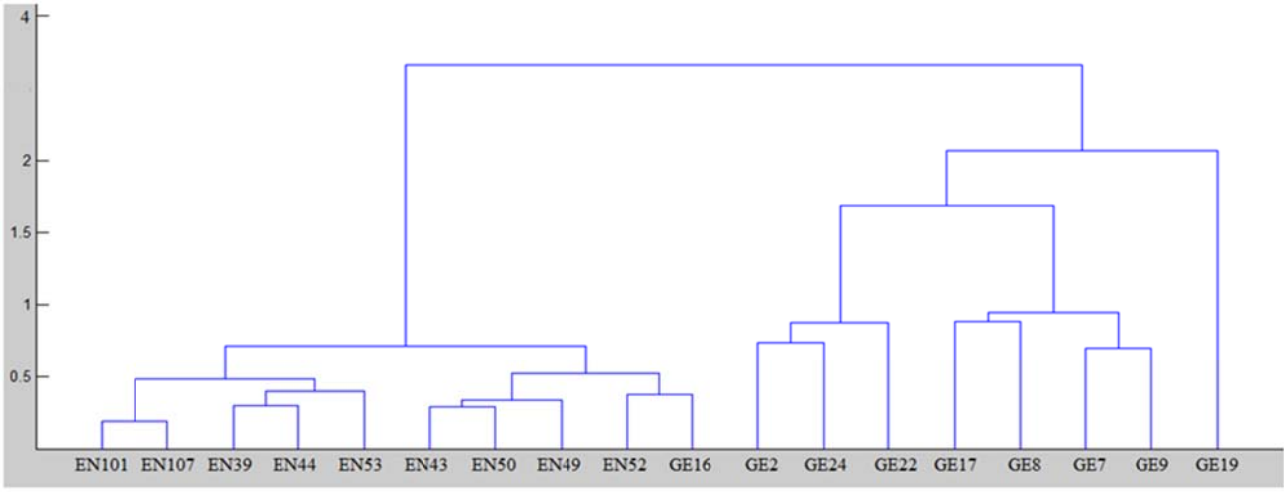


Fig.5 The clustering result of 18 selected speakers

speakers and to other non-native speakers. Non-native pronunciations can be affected by their mother tongue in different ways and to different degrees. In Fig.5, the clustering result of 18 selected speakers is shown. We picked up German speakers who were born in Germany in the archive, the number of whom was 9, and 9 native American English speakers were randomly selected. "EN" and "GE" denote American and German, respectively. The numbers succeeding "EN" or "GE" in the figure are speaker IDs. From Fig. 5, it can be seen that the all American speakers are clustered into one sub-tree and eight German speakers are clustered into the other sub-tree. Although GE16 is clustered into the same sub-tree with American speakers, by inspecting his biography included in the SAA, it is found that he had lived in USA for 4 years. It seems that his pronunciations has been reasonably affected by and adapted to American accent. On the other hand, most of the other German speakers live in America within or less than 1 year and this is supposed to be the reason why they are clustered into the other sub-tree. The 9 American and the 9 German samples will be included in the CD-ROM as media files. Interested readers should listen to those samples.

6. Use of the Reference Distances to Built a Distance Predictor for New Data

Using the inter-speaker distances calculated in the previous section as reference distances, a distance predictor for new speakers was trained and built in [1]. Here, only speech data of the new speakers were used and their IPA transcriptions were not. Invariant pronunciation analysis was adopted for pronunciation representation and Support Vector Regression was used for prediction. The correlation between manually prepared IPA-based distances and automatically predicted distances was 0.77. For comparison, an HMM-based phoneme recognizer was tested with word-based network grammar to convert new speakers' utterances into phoneme sequences, not phone sequences. Here the network grammar was built in order to cover word-based pronunciation variations found in the SAA. Then, two generated phoneme sequences were aligned through the DTW by using the HMM-based phoneme-to-phoneme distance matrix. Since almost all the data were non-native and the recording environment varied from sample to sample, the phoneme recognition performance was so low as 46 % and the resulting correlation between the IPA-based reference inter-speaker distances and the HMM-based distances was 0.043. The proposed predictor outperformed by far the HMM-based baseline system. Interested readers should refer to [1].

7. Conclusions

With the ultimate goal of drawing a global map of World Englishes on an individual basis, we're developing a method of predicting the pronunciation distance between any pair of speakers [1]. For this project, the reference pronunciation distances are required and this paper describes how to prepare these distances in detail. Since the SAA archive provides a narrow IPA transcription for each accented utterance of the fixed elicitation paragraph, the DTW was applied to those IPA transcriptions with a phone-to-phone distance matrix obtained from recordings by an expert phonetician. Using the obtained distances, speaker clustering was done. Results showed that speaker clustering was effectively and validly performed only in terms of pronunciation. Although we're focusing on only the segmental aspect of pronunciation, the obtained clustering result indicates that clustering only based on the segmental aspect can still capture how diverse World Englishes are in their pronunciation rather well. In future work, we are planning to collect a more data using social network infrastructure and incorporate the prosodic diversity into pronunciation distance calculation. Pedagogical application of the World and individual English map will also be considered in collaboration with language teachers.

8. Acknowledgements

The authors would like to thank National Science Council of Taiwan for their financial support. This work was supported in part by the National Science Council of Taiwan under the Grants NSC101-2917-I-006-011.

9. References

[1] H.-P. Shen, N. Minematsu, S. H. Weinberger, T. Makino, J. Novak, T. Pongkittiphan, C.-H. Wu,"Speaker-based pronunciation clustering of World Englishes based on pronunciation structure analysis,"*IEICE Technical Report*, SP2012-116, pp.7-12 (2013-2)

[2] S. H. Weinberger, Speech Accent Archive, George Mason University, http://accent.gmu.edu
[3] D. Crystal, *English as a global language*, Cambridge University Press, New York, 1995.
[4] J. Jenkins, *World Englishes: a resource book for students*, Routledge, 2009.
[5] B. Kachru, Y. Kachru, and C. Nelson, *The handbook of World Englishes*, Wiley-Blackwell, 2009.
[6] A. Kirkpatrick, *The Rougledge handbook of World Englishes*, Routledge, 2012.
[7] M. Pinet, Paul Iverson, Mark Huckvale, "Second-language experience and speech-in-noise recognition: the role of L2 experience in the talker-listener accent interaction", *Proc. of SLaTE*, CD-ROM, 2010.
[8] N. Minematsu, "Mathematical evidence of the acoustic universal structure in speech," in *Proc. ICASSP*, pp.889-892, 2005.
[9] N. Minematsu, Y. Qiao, S. Asakawa, M. Suzuki, "Speech structure and its application to robust speech processing", *Journal of New Generation Computing*, 28, 3, pp. 299-319, 2010.
[10] R. Jakobson and L. R. Waugh, *Sound shape of language*, Branch Line, 1979.
[11] M. Suzuki, Y. Qiao, N. Minematsu, and K. Hirose, "Pronunciation proficiency estimation based on multilayer regression analysis using speaker-independent strucural features," in *Proc. SLaTE*, CD-ROM, 2010.
[12] T. Zhao, A. Hoshino, M. Suzuki, N. Minematsu, K. Hirose, "Automatic Chinese pronunciation error detection using SVM with structural features," in *Proc. Spoken Language Technology*, pp.473-476, 2012.
[13] X. Ma, R. Xu, N. Minematsu, Y. Qiao, K. Hirose, A. Li, "Dialect-based speaker classification using speaker invariant dialect features", in *Proc. of Int. Symposium on Chinese Spoken Language Processing*, pp.171-176, 2010.
[14] N. Minematsu, K. Kamata, S. Asakawa, T. Makino, and K. Hirose, "Structural representation of the pronunciation and its use for clustering Japanese learners of English," in *Proc. SLaTE*, CD-ROM, 2007.
[15] Y. Qiao, N. Minematsu, "A study on invariance of f-divergence and its application to speech recognition," in *IEEE Trans. on Signal Processing*, vol.58, no.7, pp.3884-3890, 2010.
[16] M. Suzuki, G. Kurata, M. Nishimura, N. Minematsu, "Discriminative reranking for LVCSR leveraging invariant structure," in *Proc. INTERSPEECH*, CD-ROM, 2012.
[17] D. Saito, S. Asakawa, N. Minematsu, and K. Hirose, "Structure to speech -- speech generation based on infant-like vocal imitation --," in *Proc. INTERSPEECH*, pp1837-1840, 2008.
[18] The CMU pronunciation dictionary, http://www.speech.cs.cmu.edu/cgi-bin/cmudict
[19] N. Minematsu, Y. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "Development of English speech database read by Japanese to support CALL research," in *Proc. ICA*, pp.557-560, 2004.
[20] N. Minematsu, K. Okabe, K. Ogaki, K. Hirose, "Measurement of objective intelligibility of Japanese accented English using ERJ (English Read by Japanese) database," in *Proc. INTERSPEECH*, pp.1481-1484, 2011.

10. Appendix

Table 1: 153 phones used in acoustic analysis

Vowels and Consonants used in Acoustic Analysis					
1. i	2. ĭ	3. iː	4. i̯	5. ï	6. ĩ
7. y	8. ɪ̯	9. ɪ	10. ɪː	11. ɪ̯	12. ĩ
13. e	14. ë	15. ẽ	16. ɛ	17. ɛ̈	18. ɛ̃
19. æ̯	20. æ	21. æː	22. æ̃	23. a	24. ã
25. i̯	26. i̯	27. ˜	28. ʉ	29. ʉ̯	30. ɚ
31. ɜ	32. ɝ	33. ɐ	34. ɐ̆	35. ɐ̃	36. ɵ
37. ɵ̃	38. ə	39. ə̆	40. ə̯	41. ə̃	42. ə̯
43. ɯ	44. ɯ̈	45. ɯ̃	46. u	47. ŭ	48. uː
49. ü	50. ũ	51. ũː	52. ʊ	53. ɤ	54. o
55. ö	56. õ	57. ʌ	58. ʌ̃	59. ɔ	60. ɔː
61. ɔ̈	62. ɔ̃	63. ɑ	64. ɑː	65. ɑ̈	66. ɑ̃
67. p	68. pʰ	69. p⃗	70. b	71. b⃗	72. b̯
73. ɸ	74. β	75. β̯	76. β̃	77. f	78.v
79. y̯	80. ʋ	81.m	82. m̯	83. m̩	84.n
85. n̯	86. ɲ	87. ɳ	88. n̩	89. ŋ	90. ɴ
91.t	92. tʰ	93. t̪	94. t̯	95. t'	96. t⃗
97. d	98. d̪	99. d⃗	100. d̯	101. s	102. s̪
103. sʲ	104. z	105. z̯	106. ɹ	107. ɹ̯	108. ɻ
109. r	110. ɾ	111. ɾ̯	112. l	113. l̥	114. lⱽ
115. θ	116. ð	117. ɕ	118. ʑ	119. ʑ̯	120. ʃ
121. ʒ	122. ç	123. j̯	124. j	125. k	126. kʰ
127. k̯	128. k'	129. k̯ʰ	130. k⃗	131. g	132. g
133. g⃗	134. g̊	135. x	136. ɣ	137. ɣ	138. ɯ̯
139. ʔ	140. h	141. ɦ	142. w	143. ɥ	144. pɸ
145. tθ	146. dð	147. ts	148. dz	149. tɕ	150. dʑ
151. tʃ	152. dʒ	153. kx			

A Corpus-Based Analysis of Korean Segments Produced by Japanese Learners

Hyejin Hong[1], *Sunhee Kim*[2], *Minhwa Chung*[1]

[1]Department of Linguistics, Seoul National University, Seoul, Republic of Korea
[2]Speech Synthesis Research Lab, NHN Corporation, Seoul, Republic of Korea
souble1@snu.ac.kr, sunhkim@nhn.com, mchung@snu.ac.kr

Abstract

This paper examines variations of Korean segments produced by Japanese learners of Korean. For corpus-based statistical analysis, we have used Korean read speech corpus produced by Japanese learners. Contrastive analysis of the target language and the source language is performed to provide information for interpreting the results of corpus analysis. Segmental variations are analyzed by aligning canonical phonetic transcriptions with auditory phonetic transcriptions of the corpus. The results show that (1) Japanese learners tend to demonstrate substitutions due to differences in the phonemic systems of the two languages; and (2) they are likely to omit a consonant or insert a vowel to deal with the different syllable structures. These results with detailed statistical data are useful for designing a computer-assisted pronunciation training and assessment system for Japanese learners of Korean.

Index Terms: Korean language education, Japanese learners, segmental variations, computer-assisted language learning

1. Introduction

Recent years have seen that computer-assisted language learning (CALL) systems have been developed in line with advances in spoken language technology [1]. Due to a large variability observed in non-native speech, it is difficult to automatically process non-native learners' speech and to provide corrective feedback [2]. Therefore, analyzing variations in non-native learners' speech is essential to lay the groundwork for developing a CALL system.

In general, a contrastive analysis [3], which compares learners' native language with their target foreign language, is used to predict the general patterns of variations in non-native speech. In previous studies [4][5], the general patterns of Korean speech produced by Japanese learners are described based on the contrastive analysis. However, there are some drawbacks in the contrastive analysis such as (1) it is difficult to predict all possible variations, (2) it is uncertain that the predicted variations are totally matched with the ones which are found in learners' real speech, and (3) it is difficult to quantify which variations are more frequently produced by non-native learners. Statistical analysis of a spoken corpus produced by non-native learners is needed to compensate for drawbacks of the contrastive analysis.

In order to develop a computer-assisted pronunciation training and assessment system for learners of Korean as their foreign language, as a preliminary study, we have examined segmental variations of Korean produced by non-native speakers using both contrastive analysis and corpus-based statistical analysis. Considering that interest in learning Korean has been growing with the spread of Korean popular culture [6], CALL systems for teaching Korean can provide a useful learning environment to many foreign learners. In this paper, focusing on Japanese learners of Korean, the salient segmental variations produced by Japanese learners are presented and the factors which influence the variations are examined, based on a systematic analysis of speech data from 34 Japanese learners.

The remaining part of this paper is organized as follows. Section 2 presents a contrastive analysis of Korean and Japanese, which provides a ground for interpreting the results of corpus analysis. Section 3 describes details of materials and methods for analyzing Korean segments produced by Japanese learners. Analysis results are presented in Section 4, which is followed by conclusions in Section 5.

2. Contrastive Analysis

Contrastive analysis assumes that learners' speech production is mainly influenced by their native language, and difficulties learners encounter can be predicted by comparison of their native language and the target language. A more recent model of foreign language learning, the Speech Learning Model (SLM), claims that the phonetic system of the native language and that of the target language are mutually influenced [7]. Considering that learners' native language influences the foreign language speech production to a large extent, it is helpful to compare the native language and the target language.

Differences in the Korean and Japanese phonemic systems which are expected to be related to substitutional variations are discussed in this section, and differences in syllable structures which are likely to lead to insertions or deletions of segments are presented as well.

2.1. Phonemic systems

The focus of the analysis is more on phonemes than on phones, since our corpus analysis is performed at a level close to the phonemic level. The phonemic consonantal inventory of Korean and that of Japanese are presented in Table 1 and Table 2, respectively.

Table 1. *Korean consonants. Adapted from* [8].

	Bilabial	Dental/ Alveolar	Palatal	Velar	Glottal
Stop	b̥ pʰ p͈	d̥ tʰ t͈		g̊ kʰ k͈	
Affricate			ʤ̥ ʧʰ ʧ͈		
Fricative		s s͈			h
Nasal	m	n		ŋ	
Liquid		l			
Semi-vowel	w*		j	w* ɯ̹	

* classified as both bilabial and velar due to double articulation

Table 2. *Japanese consonants. Based on* [9].

	Bilabial	Dental/ Alveolar	Post-alveolar	Palatal	Velar	Uvular	Glottal
Stop	p b	t d			k g		
Affricate		ʦ	ʧ ʤ				
Fricative	ɸ	s z	ʃ				h
Nasal	m	n				ɴ*	
Liquid		ɾ					
Semi-vowel	w			j			

* moraic nasal

Korean has a three-way distinction of stops and affricates: lenis, aspirated and fortis. The lenis stops /p̥, t̥, k̥/ and the lenis affricate /ʧ̥/ are realized as slightly aspirated, the aspirated stops /pʰ, tʰ, kʰ/ and the affricate /ʧʰ/ as heavily aspirated, and the fortis stops /p͈, t͈, k͈/ and the affricate /ʧ͈/ as laryngealized and unaspirated [10]. They are all voiceless. In contrast, Japanese shows voicing contrasts in stops and affricates. Korean shows a two-way distinction between the lenis fricative /s/, the fortis fricative /s͈/, whereas there is a two-way voicing contrast in Japanese fricatives. For nasals, Korean has the velar nasal /ŋ/ which is only permitted as the final consonant, while Japanese has a moraic nasal 'N' which has various phonetic realizations including [m, n, ɲ, ŋ, ɴ]. The Korean velar semi-vowel /ɯ̯/ is a phoneme which does not exist in the Japanese phonemic system.

Regarding vowels, Korean has more phonemes than Japanese: the Korean vowel system has 8 vowels (excluding the two vowels [ʏ], [ø] since they are mostly pronounced as diphthongs rather than monophthongs), whereas there are 5 vowels in the Japanese vowel system.

Table 3. *Korean vowels. Adapted from* [8].

	Front	Central	Back
Close	i		ɯ u
Close-mid	e		o
Open-mid	ɛ		
Open		a	ʌ

Table 4. *Japanese vowels. Adapted from* [9].

	Front	Central	Back
Close	i		u
Mid	e		o
Open		a	

Although there are predicted detailed phonetic differences, the Korean vowel system has counterparts for all the Japanese vowels. However, the Japanese vowel system lacks the three vowels /ɛ/, /ʌ/, /ɯ/, which exist in the Korean vowel system.

The comparative study shows that the two languages differ in both consonants and vowels. These differences in the phonemic systems can lead to the prediction that Japanese learners show variations when the target phoneme is missing in Japanese or shows different distribution.

2.2. Syllable structures

A syllable in Korean is composed of an optional consonant, an optional semi-vowel and a vowel followed by an optional final consonant: (C)(j/w/ɯ̯)V(C) [8]. Japanese allows a syllable composed of an optional onset such as a consonant or a consonant and a semi-vowel /j/, and a nucleus such as a vowel or a vowel followed by the moraic nasal 'N' or the first half of a geminate consonant 'Q': (C)(j)V(N/Q) [11]. Basically, the canonical Japanese syllables are open syllables except when either a moraic nasal or a geminate consonant occurs as a coda. In contrast, Korean permits both open syllables and closed syllables. Differences in their syllable structures can lead to the prediction that Japanese learners tend to simplify Korean syllable structure in order to make it similar to the structure permitted in Japanese.

The comparative study shows that Korean and Japanese significantly differ in both phonemic systems and syllable structures, which poses problems in Japanese learners' Korean segmental production.

3. Method

In order to examine Korean segments produced by Japanese learners, we adopt a corpus-based statistical analysis using a Korean read speech corpus produced by Japanese learners.

3.1. Speech material

The speech material for Korean speech produced by Japanese learners is taken from the Korean read speech corpus uttered by 100 learners with various native languages such as Japanese, Chinese, English, Russian, Turkish and Vietnamese.

The data for the analysis consists of read speech produced by a total of 34 Japanese adult learners of Korean, aged from 22 to 52. Their proficiency in Korean language is distributed from novice to advanced levels. About 200 sentences are produced by each speaker. The speech data contains 6,877 sentences from textbooks for teaching Korean as a foreign language. Each sentence has 6.1 words on average. A total of 7,331 word types appear in the speech corpus.

3.2. Transcriptions

For the Korean speech corpus uttered by Japanese learners, at first, orthographic transcriptions at word level are manually created by 4 native transcribers who major in Korean linguistics and literature. The corresponding canonical phonetic transcriptions are automatically generated by using a Grapheme-to-Phoneme converter [12] and orthographic word transcriptions. Using canonical phonetic transcriptions as reference, each utterance is manually transcribed to get auditory phonetic transcriptions. Both canonical and auditory phonetic transcriptions include [ɾ] as an allophone of /l/ and five allophones of syllable-final consonants [p̚, t̚, k̚, m̚, n̚] besides the Korean phonemic segments presented in Table 1 and Table 3. According to the canonical transcriptions, a total of 275,536 target segments (152,991 consonants and 122,545 vowels) are obtained.

3.3. Analysis of segmental productions

In order to generate variation matrices which can tabularize the relationship between the target segments and their actual realizations in the Japanese learners' speech production, alignment of the auditory phonetic transcriptions with the

canonical phonetic transcriptions is performed. Based on a dynamic programming algorithm, weighted distance measures according to phonetic feature-based distance such as in [13] are introduced to acquire more accurate and consistent alignment. Vowels are not aligned with consonants. Results of alignment are manually checked when the alignments are suspicious.

Variation matrices are generated based on the alignment. In each variation matrix, target segments and their realizations in the learners' speech are provided. A 'variation' occurs when a target segment and its realized segment are not identical. For the analysis of segmental variations which are expected to be salient patterns of Japanese learners, only the target segments for which the learners show below-average segment correctness are selected. In this paper, among all the variations found in the analysis, the variations which reach more than 2% are considered as salient variations produced by Japanese learners. More discussions will be given in the next section for only these variations.

4. Results and Discussion

4.1. Results

The Korean segments which do not reach the average segment correctness of 96.13%, and their salient variations produced by Japanese learners are listed in Table 5. The resultant variations include 30 consonantal variations. Since no salient vocalic variations occurred, only consonantal variations are discussed in this paper.

4.2. Discussion

Most of the variations are the ones reported as the major pronunciation problems of Japanese learners in previous research [4][5]. However, variations such as deletion of /h/ and substitution of /w/ for /ɯ/ are not captured in the previous knowledge-based studies; they are newly introduced in this paper.

A closer look at the variations reveals that the target segments are the ones which do not exist in the Japanese phonemic system or the variations are the results of the learners' strategies to deal with unacceptable syllable structures in Japanese. Among the 30 variations, 22 variations are related to differences in phonemic systems. These variations will be discussed in Section 4.2.1. There are 8 variations which are due to differences in syllable structures. As expected, these variations include the cases where the target segment is deleted or an epenthetic vowel is inserted. The variations due to differences in syllable structures will be discussed in Section 4.2.2.

4.2.1. Variations due to differences in phonemic systems

Japanese learners tend to realize the aspirated stops /pʰ, tʰ, kʰ/ and the aspirated affricate /tɕʰ/ as their fortis counterparts /p͈, t͈, k͈/ and /tɕ͈/, respectively. Substitutions of the lenis segments /b̥, d̥, g̊, dʑ̥/ for the aspirated segments follow. In the case of the Korean fortis segments /p͈, t͈, k͈, tɕ͈/, Japanese learners tend to produce them more often as the lenis segments /b̥, d̥, g̊, dʑ̥/ than as the aspirated segments /pʰ, tʰ, kʰ, tɕʰ/. For the lenis segments, only two segments – /b̥, dʑ̥/ are selected as the ones to show the salient variations. For the bilabial lenis stop /b̥/, Japanese learners produce it more often as the aspirated stop /pʰ/ than as the fortis stop /p͈/. Producing the fortis affricate /tɕ͈/ for the lenis affricate /dʑ̥/ is selected as the salient variation, while the aspirated affricate /tɕʰ/ is less likely to be produced as a substitution by Japanese learners. For the alveolar fricatives which have the lenis and fortis contrast, producing the alveolar lenis fricative /s/ as the alveolar fortis fricative /s͈/ is found to be salient, however, vice versa is not selected as a salient variation. The variations related to the three-way contrast in the Korean stops, affricates and fricatives are illustrated in Figure 1.

Two variations are related to the Korean final nasals. The Korean syllable-final velar nasal [ŋ] is confused by Japanese learners as [n]. Japanese learners' difficulty in distinguishing [ŋ] from [n] in syllable-final position is in line with the study on Japanese learners' perception of English final nasals, which reports that Japanese listeners have difficulties in distinguishing [ŋ] from [n] [14]. In our results, the Korean syllable-final alveolar nasal [n] is confused by Japanese learners as [ŋ] as well.

As expected, Japanese learners produce the target segments that do not exist in the phonemic system of their native language as incorrect ones.

Table 5. *Korean segmental productions by Japanese learners of Korean.*

Target segment	Freq.	%Corr	Realized Segment (%)		Error type
b̥	4210	95.13	pʰ	2.38	Phonemic system
			p͈	2.14	Phonemic system
pʰ	1053	84.05	p͈	11.02	Phonemic system
			b̥	2.66	Phonemic system
p͈	659	83.46	b̥	11.23	Phonemic system
			pʰ	5.31	Phonemic system
tˀ	755	75.63	-	21.56	Syllable structure
tʰ	900	71.89	t͈	22.67	Phonemic system
			d̥	2.11	Phonemic system
t͈	2184	89.74	d̥	7.60	Phonemic system
			tʰ	2.47	Phonemic system
kˀ	1687	87.91	-	7.11	Syllable structure
kʰ	1282	73.24	k͈	20.75	Phonemic system
			g̊	2.50	Phonemic system
k͈	3169	88.51	g̊	8.46	Phonemic system
			kʰ	2.90	Phonemic system
dʑ̥	7671	95.80	tɕ͈	2.45	Phonemic system
tɕʰ	2524	87.80	tɕ͈	8.56	Phonemic system
			dʑ̥	2.42	Phonemic system
tɕ͈	1076	91.36	dʑ̥	6.51	Phonemic system
s	10022	93.33	s͈	6.50	Phonemic system
h	6434	93.43	-	6.31	Syllable structure
mˀ	6891	87.67	mV	9.48	Syllable structure
nˀ	13351	88.90	ŋ	7.59	Phonemic system
ŋ	5630	77.25	nˀ	15.67	Phonemic system
l	8240	91.80	ɾV	5.44	Syllable structure
			-	2.12	Syllable structure
w	2937	95.34	-	4.66	Syllable structure
ɯ	84	65.48	-	19.05	Syllable structure
			w	15.48	Phonemic system

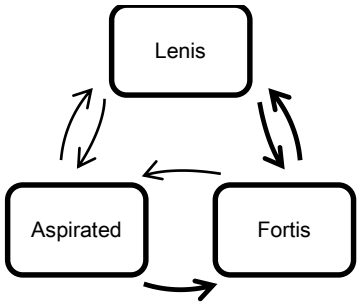


Figure 1: *Japanese learners' pattern of variations related to the three-way contrast in Korean. The bold line represents that the corresponding variation is more often found in the learners' speech than the variation marked with the solid line.*

4.2.2. Variations due to differences in syllable structures

The results of the analysis on Korean segments in Japanese learners' speech reveal that there are variations which are caused by differences in syllable structures. The prediction that Japanese learners encounter difficulties when Korean syllables have a final consonant is plausible because Japanese does not allow a coda except for the moraic nasal and the geminate consonant as presented in Section 2.2.

Japanese learners use two different strategies to deal with differences in syllable structures: either consonantal deletion or epenthetic vowel insertion. There are 1,839 occurrences of consonantal deletions (0.67%) and 1,196 occurrences of vocalic insertions (0.43%). Japanese learners omit the final Korean phonemic consonants /t/ and /k/ to avoid a syllable structure which is not permitted in Japanese. Japanese learners insert a vowel after the final consonant as well. This vowel insertion leads to re-syllabification making a new syllable as the case of the final consonant of /m/. For /l/, both consonantal deletions and vowel insertions are found, and Japanese learners tend to insert a vowel after /l/ more often than delete /l/.

For the deletion of /h/, 93.35% of /h/ is deleted when the preceding segment is a syllable-final consonant, especially a sonorant such as /m/, /n/ and /l/. This means that Japanese learners fail to produce /h/ followed by its preceding final consonant; instead, they omit the syllable-initial consonant /h/ and then make a new syllable with the preceding syllable-final consonant as its onset. Note that /h/ is voiced when it is between two sonorants in Korean like this case; however, totally deleting /h/ between two sonorants is not acceptable as a standard pronunciation [8].

Korean permits a sequence of a consonant and a semi-vowel before a vowel; however, this is very limited in Japanese. Japanese learners omit /w/ or /ɯ/ especially when there is a consonant which precedes them. These variations are the results of Japanese learners' strategy to avoid a sequence of two consonants.

As expected, Japanese learners employ consonantal deletion or vowel insertion strategies to avoid syllable structure which is not allowed in their native language.

5. Conclusions

This paper examines Japanese learners' Korean segmental variations by quantifying the patterns of learners' variations occurring in the learners' speech corpus. Based on contrastive analysis of differences in Korean and Japanese in terms of phonemic systems and syllable structures, the results of the corpus-based analysis on Korean segments produced by Japanese learners show that their segmental variations are related to differences in both phonemic systems and syllable structures. Firstly, Japanese learners are likely to show substitutional variations for Korean segments. Secondly, Japanese learners tend to omit a consonant or insert a vowel to meet the structural constraints of the native language's syllable. The results confirm that learners' segmental production in a foreign language is affected to a large extent by their native language as confirmed in a large amount of literature on foreign language acquisition.

The results of the analysis can provide background information which can be used when a computer-assisted Korean pronunciation training and assessment system is designed. In our future work, analysis of more detailed phonetic cues and effects of phonetic contexts will be investigated based on the preliminary results of this paper. In addition, analysis of Korean segments produced by learners with other languages will be examined.

6. Acknowledgements

This work was supported by the Industrial Strategic Technology Development Program, 10035252, development of dialog-based spontaneous speech interface technology on mobile platform, funded by the Ministry of Knowledge Economy of Rep. of Korea. The authors wish to thank the anonymous reviewers for their valuable comments and suggestions on an earlier version of this paper.

7. References

[1] Eskenazi, M., "An overview of spoken language technology for education", Speech Communication, 51(10):832-844, 2009.

[2] Van Compernolle, D., "Recognizing speech of goats, wolves, sheep and … non-natives", Speech Communication, 35(1-2):71-79, 2001.

[3] Lado, R., Linguistics across Cultures: Applied Linguistics for Language Teachers, Ann Arbor: University of Michigan Press, 1957.

[4] Cho, S., "An analysis for comprehensive education of pronunciation: focusing on Japanese learners", The Review of Korean Cultural Studies, 6:229-249, 2000. (in Korean)

[5] Han, J., Choi, J., Lee, H. -Y., Park, J., Lee, K., Cho, H., Cui, J., and Lee, S., Teaching Korean Pronunciation, Seoul: Hollym, 2003. (in Korean)

[6] Kim, Y., "The rising East Asian 'Wave': Korean media go global", in D. K. Thussu [Ed], Media on the Move: Global flow and contra-flow, 233-277, NY: Routledge, 2007.

[7] Flege, J. E., "Second language speech learning: Theory, findings, and problems", in W. Strange [Ed], Speech Perception and Linguistic Experience: Issues in cross-language research, 233-277, MD: York Press, 1995.

[8] Lee, H. -Y., Korean Phonetics, Seoul: Taehaksa, 1996. (in Korean)

[9] Vance, T., An Introduction to Japanese Phonology, Albany, NY: State University of New York Press, 1987.

[10] Cho, T., Jun, S.-A., and Ladefoged, P., "Acoustic and aerodynamic correlates of Korean stops and fricatives", Journal of Phonetics, 30:193-228, 2002.

[11] Kaye, J. and Yoshida S., "A government-based analysis of the 'mora' in Japanese", Phonology, 7:331-351, 1990.

[12] Lee, K. N. and Chung, M., "Morpheme-based modeling of pronunciation variation for large vocabulary continuous speech recognition in Korean", IEICE Transactions on Information and Systems, 90(7):1063-1072, 2007.

[13] Cucchiarini, C., "Assessing transcription agreement: methodological aspects", Clinical Linguistics and Phonetics, 102(2):131-155, 1996.

[14] Aoyama, K., "Perception of syllable-initial and syllable-final nasals in English by Korean and Japanese speakers", Second Language Research, 19(3):251-265, 2003.