Bio-inspired gesture recognition with baffled transducers using temporal and spectral features

Reference:
Laurijssen Dennis, Schenck Anthony, Schouten Girmi, Kerstens Robin, Aussems Sebastiaan, Paillet Eric, Gomez Randy, Nakamura Keisuke, Daems Walter, Steckel Jan.- Bio-inspired gesture recognition with baffled transducers using temporal and spectral features
Proceedings of IEEE Sensors - ISSN 1930-0395 - IEEE, 2023, p. 1-4
Full text (Publisher's DOI): https://doi.org/10.1109/SENSORS56945.2023.10324935
To cite this reference: https://hdl.handle.net/10067/2024360151162165141

Bio-inspired Gesture Recognition with Baffled Transducers using Temporal and Spectral Features

Dennis Laurijssen*†, Anthony Schenck*†, Girmi Schouten*†, Robin Kerstens*†,
Sebastiaan Aussems*†, Eric Paillet*, Randy Gomez‡, Keisuke Nakamura‡, Jan Steckel*†§
*FTI-CoSys Lab, University of Antwerp, Belgium
†Flanders Make Strategic Research Centre, Lommel, Belgium
‡Honda Research Institute Japan Co., Ltd., 8-1 Honcho, Wako-shi, Saitama, 351-0188, Japan
§jan.steckel@uantwerpen.be

Abstract—**Echolocating bats can provide engineers with tremendous inspiration for constructing active ultrasonic sensors for airborne applications. Previous research has demonstrated that the recognition of hand gestures can be facilitated by means of ultrasonic sensing, often relying on classical engineering principles. In this paper we merge the insights from research into biological echolocation systems with gesture recognition, and present a gesture recognition sensor inspired by the echolocation system of bats, using spatiospectral features induced by the geometric shape of baffled microphones. We use the spatiospectral features in combination with a support vector machine classifier as gesture classification mechanism. We show the efficacy of the proposed approach using experimental data gathered from twenty persons performing four different gestures.**

I. Introduction

Active ultrasonic sensing is a sensing modality which relies on the emission of an ultrasonic signal and recording the reflected echoes using one or multiple microphones. Airborne active ultrasonic sensing, often called SONAR (Sound Navigation and Ranging) is an established sensing modality in a wide range of sensing challenges. Indeed, applications of active sonar can be found in robotics [1], [2], the recognition and mapping and of vegetation types [3], [4], human presence detection [5], human pose estimation [6] and gait recognition [7]. The common denominator in all of these applications is that sonar is used as a low-cost and robust alternative sensing modality to the more widely spread optical sensing modalities such as 2D/3D cameras or LIDAR sensors. The sonar sensing modality is inherently low-cost due to the low speed of sound, which results in the fact that array-based sensors are quite straightforward to construct, as has been demonstrated in [8], [9], where a high-performance array based 3D imaging sensor was developed for a prototype cost of around 150 euro.

Ultrasonic sensing has also been widely applied in human computer interaction [10], and more specifically in the recognition of hand gestures. The literature presents a wide variety of system for hand gesture recognition based on single receivers or array-based receivers [11]–[16]. As most of the proposed systems operate in a narrowband regime (40kHz with a bandwidth often less than 1kHz), it is straightforward to use the motion-induced micro-doppler signatures as dominant features in the gesture recognition system [17]–[20].

Nature provides engineers with interesting insights into solving challenging engineering tasks. This bio-inspiration approach is especially true in the case of airborne ultrasound sensing. Bats are expert users of advanced sonar sensing, displaying unprecedented skill in a wide variety of tasks [21]–[23]. The ability of bats for highly adaptive and intelligent behaviour using their advanced sonar sensors has inspired a wide range of researchers to construct sonar sensors based on their biological counter part [24]–[28]. One of the main enabling features in bat echolocation is the interplay between the often large signal bandwidths used (ranging from 20kHz to 100kHz are common) and the intricate shapes of the bats outer ears (pinnae). The pinnae act as direction-dependent filters, forming a so-called Head-Related Transfer Function [29]–[31]. Analysing the spectral content of the received echoes allows the bats to infer the direction of the impinging echo [32].

In this paper we present a gesture recognition sensor for hand gestures inspired by the echolocation system of bats. We constructed an array-based sensor with 3D printed plastic covers which introduce direction-dependent spectral filters, inducing additional cues which facilitate the gesture recognition process. We describe the hard- and software architecture of the sensor, and illustrate it's efficacy by an experiment recognizing four different gestures performed by twenty persons.

II. Hardware architecture

The hardware architecture of the sensor can be seen in figure 1. The sensor system was designed as a modular system for a multitude of applications in robotics and human machine interaction. The microcontroller board used in this sensor has an ARM Cortex M4 (STM32F429) at its core and features a great number of peripherals which have been made accessible by distributing the I/O pins of the chip to two 50 pin headers on either the top or the bottom for creating a stack of interconnecting PCBs. For interfacing to a computer a second custom board was used which uses a FTDI USB interface controller chip that translates the USB protocol to four independent UART communication channels. In order to use this system as an in-air sonar sensor an extra board was designed that implemented both the acoustic emitter and the receivers. As emitter a Prowave 328ST160 ceramic transducer was chosen because of its (relatively) broad bandwidth (30kHz-42KHz) and con-

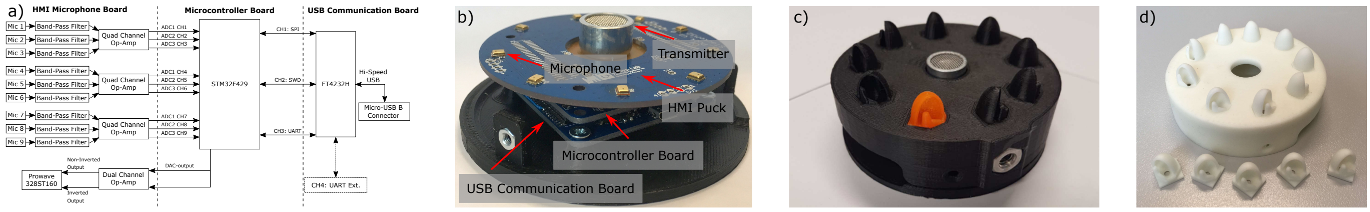

Fig. 1. Overview of the proposed system architecture. Panel a shows the overall hardware architecture of the gesture recognition sensor. Panel b) shows the implementation on three distinct printed circuit boards (USB, Microcontroller and microphone puck board). Panels c) and d) show the baffled transducers printed using two 3D printing mechanisms: the prototype in c) is printed using an FDM printer, while the prototype in d) is printed using the SLA production process.

venient electrical specifications for driving it (low voltages and currents). As the receivers Knowles SPU0410HR5H SMD microphones were chosen because of their frequency response in the ultrasonic spectrum, small footprint and low cost. The STM32F4 microcontroller has three ADCs which can each start quantizing a channel triggered by a single timer ensuring synchronous measurements. Dependent on the configuration of microphones (3, 6 or 9) the ADCs will sequentially measure their assigned microphone channels with a neglectable time difference.

We encased the microphones with 3D printed baffles, whose shape is inspired on the pinna shape of echolocating bats. As can be seen in figure 2, the baffle shape has an asymmetric tragus positioned inside the pinna, which has been shown to introduce spatiospectral cues in the received echoes. We designed the baffles to be modular and exchangeable, as can be seen in figure 1, panels c) and d). When measuring the directivity patterns of the baffled microphones, we have observed significantly varying directivity patterns across frequency as well as across the individual microphones. We hypothesize that the spectrum of the received echoes contain rich information about the performed gesture, which will be demonstrated in the subsequent section.

III. Feature extraction and Machine Learning

The bio-inspired gesture recognition system can be modelled and understood from a linear systems point of view. A signal $s_t(t)$ is emitted by the transducer, which is in our case a linear chirp from 36kHz to 42kHz in 1 millisecond. This signal is reflected by a set of point reflectors (N) originating from direction $\boldsymbol{\psi_n} = [\theta, \varphi]^T$, with θ the azimuth direction and φ the elevation direction. The receiver array consists of K microphones, and the k-th microphone signal can be written as:

$$s_r^k(t) = \sum_{n=1}^{N} h_{\boldsymbol{\psi}_n}(t) * s_t(t - \Delta t_n)$$

with $h_{\boldsymbol{\psi}_n}(t)$ the impulse response for the baffled microphone in direction $\boldsymbol{\psi}$, and $*$ denoting time-domain convolution. From these K microphone signals we extract various describing features to allow our machine learning algorithm to infer the gesture identity. We use hand-crafted features instead of the now-popular convolutional neural networks, due to the fact that these algorithms require huge datasets to converge, which would be impractical to obtain. Furthermore, hand-crafted features are to be preferred when the system has salient features which can be easily extracted from the data.

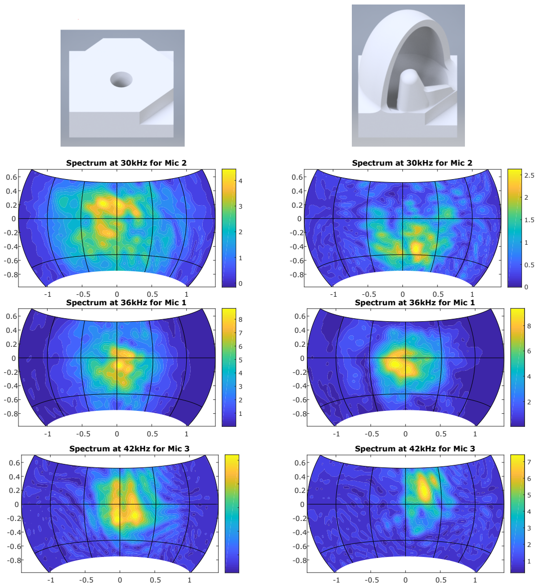

Fig. 2. Spatiospectral features introduced by the baffle shapes. The plot shows the spatiospectral response of the system for a point reflector positioned at various azimuth/elevation locations. We calculated the received spectra for a frequency range from 30kHz to 42kHz. The system response is shown for three distinct microphones.

As a first step, we perform a bandpass filter $h_{bp}(t)$ to remove unwanted noise in acoustical frequencies in which we did not ensonify the target. We use a sixth order Butterworth filter with cut-off frequencies of 35kHz and 45kHz. Next,we extract the envelope of the signal using full-wave rectification and subsequent low-pass filtering using a second order Butterworth lowpass filter $h_{lp}(t)$:

$$s_e^k(t) = h_{lp}(t) * \left| h_{bp}(t) * s_r^k(t) \right|$$

From these K envelope signals we extract the following two features: the maximum of the first peak (which is the main reflection of the hand), and the time at which this maximum occurs. The first feature encodes the strength e_t^k of the echo and the second one encodes the range r_t^k of the object in front of the sensor. The temporal feature vector for a single sonar emission at time τ is then equal to:

$$F_{TD}(\tau) = \left[e_t^1(\tau) \quad r_t^1(\tau) \quad \ldots \quad e_t^K(\tau) \quad r_t^K(\tau)\right]^T$$

with K equal to 9 in our proposed system (as our sensor has nine microphones). Each sonar measurement thus yields an $[18 \times 1]$ vector of temporal features. The second set of features are the spectral features, which we believe to convey significant information about the performed gesture due to the direction-dependent filtering which is introduced by the baffles around the microphones. These features are extracted by calculating the spectral content of the reflection signal. The spectrum $S_r^k(j\omega)$ of the k-th received signal is calculated using the DFT. From these spectra, we extract the frequency $f^k(\tau)$ on which the spectrum reaches its maximum value $e_f^k(\tau)$. These features are then combined to yield the spectral feature vector of size $[18 \times 1]$ at sonar emission at time step τ:

$$F_{TD}(\tau) = \left[f^1(\tau) \quad e_f^1(\tau) \quad \ldots \quad f^K(\tau) \quad e_f^K(\tau)\right]^T$$

Finally, we combine the two feature vectors in to a single feature vector at time step τ:

$$F(\tau) = \left[F_{TD}(\tau)^T \quad F_{TD}(\tau)^T\right]^T$$

which has a size of $[36 \times 1]$. Each gesture recognition sequence consists of 15 sonar measurements gathered at a rate of approximately 10Hz. So each sonar gesture recording event spans approximately 1.5 seconds. For each of these sonar measurements we calculate the feature vector $F(\tau)$ which are then combined into the overall feature vector G:

$$G = \left[F(1)^T \quad F(2)^T \quad \ldots \quad F(15)^T\right]^T$$

which has a size of $[15 \cdot 2 \cdot 18 \times 1] = [540 \times 1]$. All subsequent operations are performed on this feature vector G, which we hypothesize to contain sufficient information to perform gesture recognition through supervised learning by support vector machines [33]. To validate this approach we generate a labelled training dataset consisting of four gestures: swipe left, swipe right, swipe up and swipe down, performed by 20 test subjects. Each gesture is repeated 20 times in randomized order, yielding 80 gestures per person, resulting in 1600 performed gestures. We divide this dataset into 1000 training samples and 600 test samples. We concatenate the training and test data into two matrices, D_{tr} for training data ($[540 \times 1000]$) and D_{te} for the test data ($[540 \times 600]$). We then calculate a PCA basis from the training data to perform dimensionality reduction [34]. We keep the 40 most prominent eigenvectors, and construct a projection basis V_{PCA} of size $[40 \times 540]$. The dimensionality-reduced feature vectors can then be found by the matrix-product between the data-matrices and the PCA-matrix:

$$D_{tr}^V = V_{PCA} \cdot D_{tr} \quad D_{te}^V = V_{PCA} \cdot D_{te}$$

These dimensionality-reduced feature vectors are then classified using a multi-class support vector machine [33]. We use support vector machines with polynomial kernels of order three and automatic kernel scaling. Using this approach, we can show that the sensor is able to distinguish between the different performed gestures with an average precision of 92% (see figure 3, left panel). To verify the importance of the spectral features introduces by the baffled microphones, we retrain the machine learning algorithm using only the temporal features F_{TD}, more specifically, only the r_k components of F_{TD}. This removes all amplitude information from the data, which is induced mainly by the baffled transducers. As expected, the performance of the gesture recognition drops significantly, which can be seen in figure 3, right panel. While the up and down gestures can still be recognized appropriately, there is significant left-to-right confusion. This can be explained that the up and down gestures have a large impact on the time-domain information due to large variations in range during these gestures. During the left and right gestures the variability of the range is much less pronounced, which is therefore poorly encoded by the temporal features. The amplitude and spectral features encode these variations much more saliently due to the baffled transducers, explaining the improved performance of the gesture recognition system.

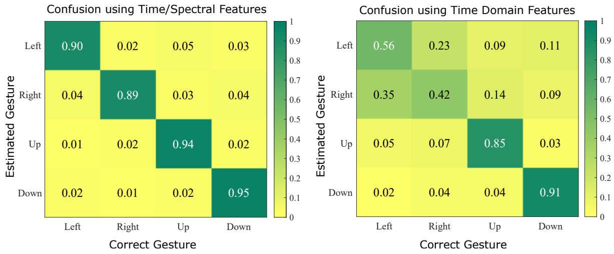

Fig. 3. Confusion matrices for the four-gesture recognition task. The right matrix shows the confusion for recognition using only temporal features while the left matrix shows the confusion for the recognition using temporal+spectral features. The machine learning algorithm was trained with 1000 gestures, performed by 20 subjects, each performing the four different gesture classes.

IV. Conclusions and Future Work

In this paper we presented a novel ultrasonic gesture recognition sensor based around baffled microphones. We inspired the design of the microphone baffles on the outer ear structures of echolocating bats. We show the spatiospectral cues introduced into the echoes by the microphone baffles, and explained the necessary signal processing steps for feature extraction. We demonstrate the efficacy of the sensor in a gesture recognition task with twenty subjects, each performing four gestures with twenty repetitions. The system using spectral and temporal features is capable of recognizing the gestures with high accuracy, while the performance drops when only using temporal features. In the future, we will expand the range of gestures that the system can recognize, and increase the experimental dataset, to allow more accurate gesture recognition due to improved generalization of the machine learning algorithm.

References

[1] J. Steckel, A. Boen, and H. Peremans, "A sonar system using a sparse broadband 3d array for robotic applications," 2012.

[2] G. Orchard and R. Etienne-Cummings, “Discriminating multiple nearby targets using single-ping ultrasonic scene mapping,” *Circuits and Systems I: ...*, vol. 57, pp. 2915–2924, 2010. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs$_{a}ll.jsp?arnumber = 5491286$

[3] Y. Yovel, P. Stilz, M. O. Franz, A. Boonman, and H.-U. Schnitzler, “What a plant sounds like: the statistics of vegetation echoes as received by echolocating bats.” *PLoS computational biology*, vol. 5, p. e1000429, 7 2009.

[4] I. Eliakim, Z. Cohen, G. Kosa, and Y. Yovel, “A fully autonomous terrestrial bat-like acoustic robot,” *PLOS Computational Biology*, vol. 14, p. e1006406, 9 2018. [Online]. Available: http://dx.plos.org/10.1371/journal.pcbi.1006406

[5] I. Google and G. Dublon, “A survey of human-sensing: Methods for detecting presence, count, location, track, and identity antipodal staged processing in role-adaptive embedded systems view project sensing view project thiago teixeira a survey of human-sensing: Methods for detectin,” 2017. [Online]. Available: https://www.researchgate.net/publication/319791520

[6] D. Laurijssen, S. Truijen, W. Saeys, and J. Steckel, “Three sources, three receivers, six degrees of freedom: An ultrasonic sensor for pose estimation amp; motion capture.” IEEE, 11 2015, pp. 1–4. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7370689

[7] K. Kalgaonkar and B. Raj, “Acoustic doppler sonar for gait recogination.” IEEE, 9 2007, pp. 27–32. [Online]. Available: http://ieeexplore.ieee.org/document/4425281/

[8] R. Kerstens and J. Steckel, “Live demonstration: 3d sonar sensing using low-cost mems arrays.” IEEE, 10 2017, pp. 1–1. [Online]. Available: http://ieeexplore.ieee.org/document/8234021/

[9] R. Kerstens, D. Laurijssen, and J. Steckel, “Low-cost one-bit mems microphone arrays for in-air acoustic imaging using fpga’s.” IEEE, 10 2017, pp. 1–3. [Online]. Available: http://ieeexplore.ieee.org/document/8234087/

[10] T. Dahl, J. L. Ealo, H. J. Bang, S. Holm, and P. Khuri-Yakub, “Applications of airborne ultrasound in human–computer interaction,” *Ultrasonics*, vol. 54, pp. 1912–1921, 9 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0041624X14000973

[11] S. Gupta, D. Morris, S. Patel, and D. Tan, “Soundwave.” ACM Press, 2012, p. 1911. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2207676.2208331

[12] Y. Qifan, T. Hao, Z. Xuebing, L. Yin, and Z. Sanfeng, “Dolphin: Ultrasonic-based gesture recognition on smartphone platform.” IEEE, 12 2014, pp. 1461–1468. [Online]. Available: http://ieeexplore.ieee.org/document/7023784/

[13] A. Das, I. Tashev, and S. Mohammed, “Ultrasound based gesture recognition.” IEEE, 3 2017, pp. 406–410. [Online]. Available: http://ieeexplore.ieee.org/document/7952187/

[14] I. Dokmanic and I. Tashev, “Hardware and algorithms for ultrasonic depth imaging.” IEEE, 5 2014, pp. 6702–6706. [Online]. Available: http://ieeexplore.ieee.org/document/6854897/

[15] W. Ruan, Q. Z. Sheng, L. Yang, T. Gu, P. Xu, and L. Shangguan, “Audiogest.” ACM Press, 2016, pp. 474–485. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2971648.2971736

[16] S. Gupta, P. Molchanov, X. Yang, K. Kim, S. Tyree, and J. Kautz, “Towards selecting robust hand gestures for automotive interfaces.” IEEE, 6 2016, pp. 1350–1357. [Online]. Available: http://ieeexplore.ieee.org/document/7535566/

[17] R. J. G. van Sloun, S. Srinivasan, A. Pandharipande, and P. C. W. Sommen, “Ultrasonic array doppler sensing for human movement classification,” *IEEE Sensors Journal*, vol. 14, pp. 2782–2791, 8 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6785965/

[18] B. Raj, K. Kalgaonkar, C. Harrison, and P. Dietz, “Ultrasonic doppler sensing in hci,” *IEEE Pervasive Computing*, vol. 11, pp. 24–29, 2 2012. [Online]. Available: http://ieeexplore.ieee.org/document/6133264/

[19] G. Ogris, T. Stiefmeier, H. Junker, P. Lukowicz, and G. Troster, “Using ultrasonic hand tracking to augment motion analysis based recognition of manipulative gestures.” IEEE, pp. 152–159. [Online]. Available: http://ieeexplore.ieee.org/document/1550800/

[20] K. Kalgaonkar and B. Raj, “One-handed gesture recognition using ultrasonic doppler sonar.” IEEE, 4 2009, pp. 1889–1892. [Online]. Available: http://ieeexplore.ieee.org/document/4959977/

[21] H.-U. Schnitzler, C. F. Moss, and A. Denzinger, “From spatial orientation to food acquisition in echolocating bats,” *Trends in Ecology Evolution*, vol. 18, pp. 386–394, 8 2003. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S016953470300185X

[22] C. F. Moss and A. Surlykke, “Probing the natural scene by echolocation in bats.” *Frontiers in behavioral neuroscience*, vol. 4, pp. 1–16, 1 2010.

[23] D. S. Jacobs and A. Bastian, *Bat Echolocation: Adaptations for Prey Detection and Capture*. Springer, Cham, 2016, pp. 13–30. [Online]. Available: http://link.springer.com/10.1007/978-3-319-32492-0$_2$

[24] V. a Walker, H. Peremans, and J. C. Hallam, “One tone, two ears, three dimensions: a robotic investigation of pinnae movements used by rhinolophid and hipposiderid bats.” *The Journal of the Acoustical Society of America*, vol. 104, pp. 569–79, 7 1998. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/9670547

[25] F. Schillebeeckx, F. D. Mey, D. Vanderelst, and H. Peremans, “Biomimetic sonar: Binaural 3d localization using artificial bat pinnae,” *The International Journal of Robotics Research*, vol. 30, pp. 975–987, 9 2010. [Online]. Available: http://ijr.sagepub.com/cgi/doi/10.1177/0278364910380474

[26] J. Steckel and H. Peremans, “A novel biomimetic sonarhead using beamforming technology to mimic bat echolocation,” *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 59, 2012.

[27] C. J. Baker, G. E. Smith, A. Balleri, M. Holderied, and H. D. Griffiths, “Biomimetic echolocation with application to radar and sonar sensing,” *Proceedings of the IEEE*, pp. 1–12, 2014. [Online]. Available: http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=6766229

[28] A. B. Balleri, H. G. Griffiths, and C. B. Baker, Eds., *Biologically-Inspired Radar and Sonar: Lessons from nature*. Institution of Engineering and Technology, 7 2017. [Online]. Available: https://digital-library.theiet.org/content/books/ra/sbra514e

[29] J. Steckel and J. Reijniers, “Biomimetic target localisation using an emfi based array,” *17th International Symposium on Applications of Ferroelectrics(ISAF2008)*, vol. 3, pp. 1–2, 2008. [Online]. Available: http://ieeexplore.ieee.org/xpls/abs$_{a}ll.jsp?arnumber = 4693876$

[30] R. Müller, “Numerical analysis of biosonar beamforming mechanisms and strategies in bats.” *The Journal of the Acoustical Society of America*, vol. 128, pp. 1414–25, 9 2010. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/20815475

[31] M. Aytekin, E. Grassi, M. Sahota, and C. F. Moss, “The bat head-related transfer function reveals binaural cues for sound localization in azimuth and elevation,” *The Journal of the Acoustical Society of America*, vol. 116, p. 3594, 2004. [Online]. Available: http://link.aip.org/link/JASMAN/v116/i6/p3594/s1Agg=doi

[32] J. Reijniers, D. Vanderelst, and H. Peremans, “Morphology-induced information transfer in bat sonar,” *Physical review letters*, vol. 105, 2010. [Online]. Available: http://prl.aps.org/abstract/PRL/v105/i14/e148701

[33] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[34] L. V. der Maaten, “Dimensionality reduction: A comparative review,” *Journal of Machine ...*, vol. 10, pp. 1–41, 2009.