Fast frequency discrimination and phoneme recognition using a biomimetic membrane coupled to a neural network

Bioinspiration & Biomimetics

OPEN ACCESS

RECEIVED
10 September 2020

ACCEPTED FOR PUBLICATION
6 November 2020

PUBLISHED
25 January 2021

PAPER

Fast frequency discrimination and phoneme recognition using a biomimetic membrane coupled to a neural network

Woo Seok Lee[1,2], **Hyunjae Kim**[1,3], **Andrew N Cleland**[4] and **Kang-Hun Ahn**[1,3,*]

[1] Department of Physics, Chungnam National University, Daejeon, 34134, Republic of Korea
[2] Center for Theoretical Physics of Complex Systems, Institute for Basic Science (IBS), Daejeon 34051, Republic of Korea
[3] Deep Hearing Ltd., Daejeon, 35220, Republic of Korea
[4] Pritzker School of Molecular Engineering, University of Chicago, Chicago, IL 60637, United States of America
[*] Author to whom any correspondence should be addressed.

E-mail: ahnkh@deep-hearing.com

Keywords: basilar membrane, cochlea, hearing, deep learning

Supplementary material for this article is available online

Abstract

In the human ear, the basilar membrane plays a central role in sound recognition. When excited by sound, this membrane responds with a frequency-dependent displacement pattern that is detected and identified by the auditory hair cells combined with the human neural system. Inspired by this structure, we designed and fabricated an artificial membrane that produces a spatial displacement pattern in response to an audible signal, which we used to train a convolutional neural network. When trained with single frequency tones, this system can unambiguously distinguish tones closely spaced in frequency. When instead trained to recognize spoken vowels, this system outperforms existing methods for phoneme recognition, including the discrete Fourier transform, zoom FFT and chirp z-transform, especially when tested in short time windows. This sound recognition scheme therefore promises significant benefits in fast and accurate sound identification compared to existing methods.

1. Introduction

The cochlea is the key organ for sound detection and recognition in mammals [1, 2], comprising a mechano-electrical transducer that converts sound pressure into neuronal electrical signals. The cochlea generates these signals in a frequency-selective manner, due to the varying stiffness of the basilar membrane (BM) along the length of the cochlea [3]. For a given pure-tone audio signal, the location of the maximum displacement of the BM depends on the frequency of the tone: the base region responds to high-frequency sounds, while the apex region responds to low-frequency sounds. The audible range of the human auditory system, as determined by the mechanical response of the BM, ranges over twenty octaves, from roughly 20 Hz to 20 kHz, with a wide dynamic range of about 120 dB [1].

Displacements of the BM are detected through the sympathetic motion of hair cells in the membrane, which generate electrical signals in the auditory nerves. The distinct spatial patterns of the BM's frequency-dependent response are encoded in the corresponding neural patterns, which are recognized in the brain. Using this auditory system, humans are able to accurately distinguish very short-duration sounds, while conventional frequency analysis faces significant challenges.

Building a biomimetic analog of the BM, combined with an analog of the brain's pattern recognition system, provides an interesting approach for the frequency analysis of sound. Deep-learning techniques involving neural networks with multiple layers are important components in modern automatic speech recognition systems [4], and have been used for feature representation [5], acoustic modeling [6], and language modeling [7]. In this study, we combine a physical model of the BM, yielding a frequency-dependent spatial response, with an artificial deep neural network for pattern recognition.

The question we pose is what, if any, are the benefits of this approach to signal processing applications? Our findings can be summarized as follows:

- A mechanical system with poor innate frequency selectivity can be significantly enhanced when combined with a trained neural network.

- The membrane together with the neural network provides enhanced pitch recognition for short duration signals, clearly distinguishing two frequencies f_1 and f_2 even when sampling for a time T shorter than $1/|f_1 - f_2|$.
- The signal processing ability of the membrane combined with the neural network outperforms the standard discrete Fourier transform (DFT), the zoom fast Fourier transform (ZFFT) and the chirp Z-transform (CZT), the latter comprising popular methods for improving frequency resolution [8–10].
- We demonstrate that our system is helpful in recognizing real speech phonemes, and additionally can distinguish very short duration speech signals.

2. Membrane fabrication and experiments

Our artificial basilar membrane (ABM) is a macroscopic structure mimicking von Békésy's model of the cochlea [11, 12]. Several research groups have previously developed similar approaches [13–20], where frequency-selective ABMs have been designed in which the membrane width and thickness are varied along the membrane length to allow the response of the ABMs to mimic the tonotopy of the cochlea. The microscopic mechanical membrane in these works requires a sound pressure level (SPL) of more than 80 dB to obtain a significant response amplitude, as their motional amplitudes for lower SPL levels are quite small, of order a few tens of nanometers. Low-volume sounds are therefore quite challenging to detect using these structures, due to the difficulty of detecting very small (nm-scale) displacements. The purpose of the present study is not to develop a practical ABM but rather to find a novel mechanism for auditory signal processing. We therefore designed a large-scale mechanical membrane, using soft silicon rubber as a model of the BM, to obtain larger mechanical responses than in previously-developed microstructures.

Our ABM design successfully responds with large and frequency-selective displacements for sound frequencies in the range of 100 to 1300 Hz, with displacements of order 10^{-4} m even for sound intensities of 65 to 70 dB SPL, significantly smaller intensities than detectable with earlier designs. This large responsivity makes the detection of the sound-induced displacements relatively straightforward. The ABM we used in this study and the water-filled chamber it resides in were designed in reference to von Békésy's model [11, 12]; a schematic is shown in figure 1. Our ABM comprises a 0.1 mm thick sheet of silicone rubber, whose width varies along its 2 cm length from 1 mm to 5 mm. The membrane forms the mid-plane of a water-filled chamber, where the upper and lower parts of the chamber are analogs of the vestibular and tympanic canals, respectively. The upper and lower chambers are connected by a 5 mm diameter hole, analogous to the helicotrema in the cochlea. The upper part of the chamber is directly connected to an audio speaker with a plastic rod (4.5 cm), delivering an acoustic signal to the entire membrane. The SPL was measured by driving the speaker with the same power and placing a microphone immediately below the ABM and recording the value.

Acoustic signals emitted from the speaker produce spatial displacement patterns in the membrane, in synchrony with the audio signal, where the steady state response is reached in a time less than about 10 to 20 ms. The displacement pattern of the ABM is detected by reflecting beams of light from n separate lasers (650 nm wavelength, Laserlab LDC650-3.5-5) onto n one-dimensional position-sensitive detectors (Hamamatsu Photonics S3932). We used $n = 6$ for two-tone frequency discrimination and $n = 10$ for vowel recognition.

The pattern response of the ABM was measured by applying pure tone signals to the ABM with tone frequencies ranging from 100 Hz to 1300 Hz, at intervals of 100 Hz, with sound amplitudes of 65–70 dB SPL. The measured root-mean square (RMS) displacement response is shown in figure 2. Low-frequency tones yield large displacements toward the apex end of the membrane, with the maximum deflection shifting towards the base with increasing frequency, a result of the variation in membrane width from apex to base; this is similar to the cochlear response. Unlike the cochlea, we have motion along the entire length for each frequency tone, possibly because our membrane lacks the feedback mechanism provided by the outer hair cells in the cochlea. Nonetheless, we can use the measured frequency-dependent spatial response to identify frequency components in a more complex audio signal.

3. Fourier transformation and other numerical methods for the enhancement of frequency resolution

The standard numerical method for extracting frequency components $\tilde{s}(f_j)$ from a one-dimensional discrete times-series signal $s(t_k)$ is to use the DFT,

$$\tilde{s}(f_j) = \frac{1}{N}\sum_{k=0}^{N-1} s(t_k)\exp(\mathrm{i}2\pi jk/N), \qquad (1)$$

where $f_j = j\Delta f (j = 0 \ldots N-1)$ and $t_k = k\Delta t (k = 0 \ldots N-1)$. The frequency resolution is limited to $\Delta f = 1/(N\Delta t) = 1/T$, where $T = N\Delta t$ is the total duration of the time series.

Two methods that have been developed to improve the frequency resolution are known as

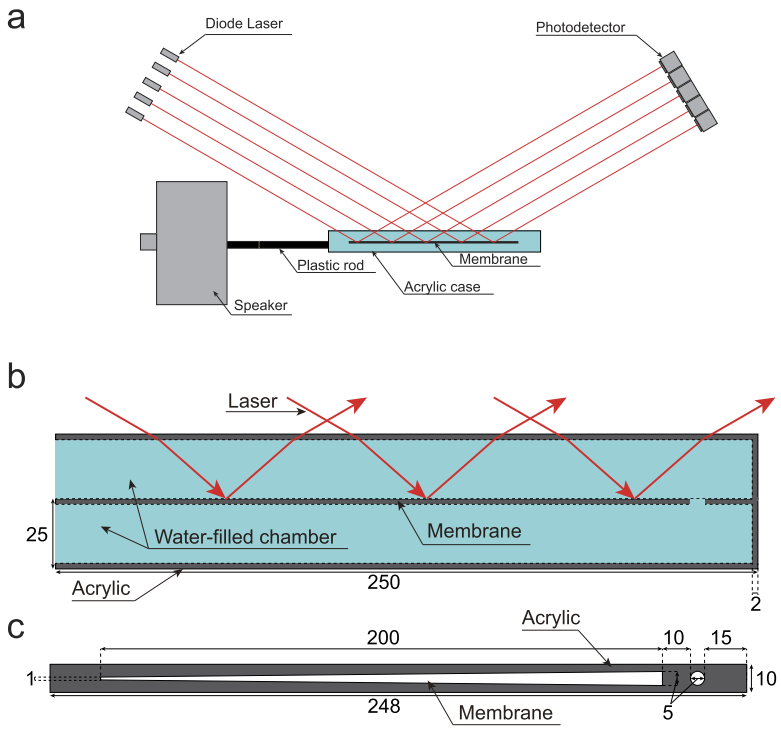

Figure 1. Experimental setup for membrane and motion detection. (a) The ABM is driven by a speaker and its displacement pattern monitored with multiple lasers paired with displacement sensors. Six lasers were used for two-tone discrimination experiments and ten lasers were used for vowel discrimination. (b) Side view of the ABM, positioned at the midplane of the upper and lower water-filled chambers, with the hole connecting the chambers to the right. (c) Top view of the ABM, with dimensions in mm.

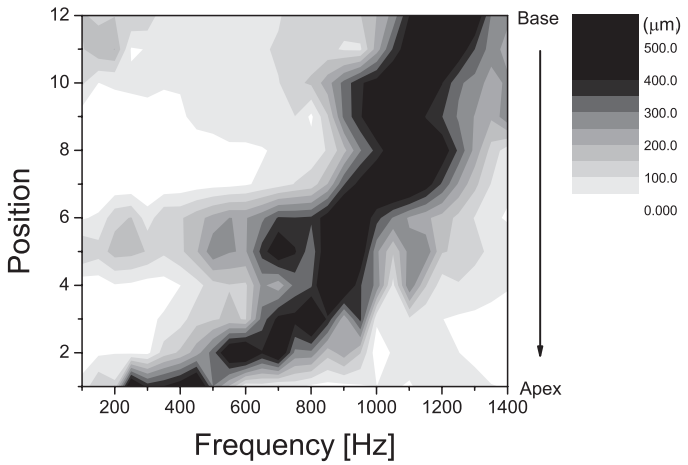

Figure 2. Displacement response of the ABM (grey scale) as a function of position along the ABM (vertical scale) and the frequency of a continuous pure audio tone (horizontal scale), with displacement amplitudes measured after the ABM displacement has reached steady state. For this measurement we monitored the RMS displacement amplitude by scanning 12 positions on the ABM using a single laser. Displacement position 1 is located 4 cm from the apex, and positions 2–12 are spaced evenly by 0.8 cm toward the base direction.

the ZFFT and CZT. The principal concept for the ZFFT is to use the same total sample duration T, but to increase the time step Δt, resulting in down-sampling of the data. Any resulting non-integer relation between T and Δt is compensated by zero-padding. This gives an increased frequency resolution Δf from the same data set, however accompanied by the loss of high-frequency components due to down-sampling, and the addition of spurious frequency components due to zero-padding. Increased down-sampling loses information about the original signal and reduces the quality of the transform.

The CZT is a generalization of the DFT. The DFT samples the transform plane at uniformly-spaced points along the unit circle, while the CZT samples along spiral arcs, effectively using a non-uniform frequency coverage. It is defined as

$$\tilde{s}(f_j) = \sum_{k=0}^{N-1} s(t_k) A^{-j} W^{jk} \quad (2)$$

where $A = e^{if_i/f_s}$ and $W = e^{-i2\pi(f_f - f_i)/M}$, f_i and f_f are the initial and final and frequencies, and M the number of samples in frequency. The CZT can significantly enhance the frequency resolution without increasing the number of sampling data points, but it generates significant spectral leakage. For wider frequency signals, the result of the CZT is inaccurate.

Unlike these and other frequency refinement methods where no additional information is added to the time-series signal, the use of the mechanical response of the ABM in our approach adds spatial information to the time-series signal through the tonotopy, resulting in better frequency resolution without loss of signal.

4. Training neural networks

The neural network model we use to analyze the spatial patterns of the ABM is the convolutional neural network (CNN) [21] with a fully connected network [22]. The required nonlinear response of the system is provided by the rectified linear unit [23]. For supervised learning, we use a quadratic cost function. An overview of the membrane-neural network system is shown in figure 3, and details of the implementation and results for the neural network training and testing are given in the SI.

We trained the neural network using membrane patterns generated by combinations of two pure tones, using displacement data corresponding to time segments ranging from 20 ms to 200 ms in duration, with amplitudes of 65–70 dB SPL. The membrane displacement shows frequency selectivity for signals from 100 to 1300 Hz (figure 2). We used two pure tones, each with frequency in the range 600 to 695 Hz, with a step size and minimum frequency difference of 5 Hz, giving 190 different frequency combinations (cases). The network was trained so that each output node gives a binary output for each frequency component; training using two-tone signals, where two distinct nodes should yield TRUE outputs, therefore can discriminate 190 distinct frequency combinations, matching the number of cases. We trained the network with 200 data sets per case, and tested with 20 data sets per case, with no overlap between the training and test data. After training, the neural network yielded high test accuracy (over 95%) for the time window size of 30–50 ms.

The training errors are used to optimize the network parameters, using a back-propagation algorithm [24]. A test error can occur when the trained network is operated with data outside the training set. We compared the training and test error rates when training the network with data from just one displacement channel to that using six displacement channels. With just one channel, the learning speed of the network was very slow, and with poor final outcomes, with a highly-trained error rate of about 20%. Using six displacement channels, however, yielded much faster training as well as significantly fewer training errors, as shown in the SI.

5. Results

We find that multi-channel signal processing using the ABM and the neural network provides a frequency resolution that is superior to existing methods, especially when using small sampling time windows. In figure 4, we compare the frequency resolution for 50 ms-duration, two-tone signals, achieved using three numerical methods (DFT, ZFFT, and CZT) with that achieved with our trained network (2D network5 in the SI). The test data used to generate the neural network results in figure 4(d) are distinct from the data used to train the network. Our biomimetic system clearly resolves the two frequency tones, even for frequency differences as small as 5 Hz. By comparison, the resolution Δf of the DFT is $\Delta f = 1/T = 1/50$ ms $= 20$ Hz for a 50 ms time window, and the resulting poor resolution is shown in panel (a). Frequency components such as 605 Hz and 610 Hz cannot be resolved. The frequency resolution is improved with the CZT and ZFFT methods, as shown in figures 4(b) and (c), with the frequency resolution Δf reduced to 5 Hz. However, in this case, about 30% of CZT and ZFFT results have the wrong frequency peak values, with an error of about 5 Hz. In contrast, when performing frequency analysis with the membrane-based pattern recognition, it is possible to resolve frequencies separated by only 5 Hz, smaller by a factor of four from the DFT result, with a frequency accuracy close to 100 percent.

In addition to the good frequency identification delivered by the ABM, we find that fast recognition of complex sounds is feasible using the same system, a problem that is significantly more complex than resolving a two-tone signal. We test the performance of the ABM combined with a trained neural network, with the CNN trained to recognize five different Korean vowels ([a], [e], [i], [o], [u]) with different sampling time windows (0.5, 1, 5, 10, 20 ms). The data sets are from reference [25], in which 75 Korean students from Chungnam National University, Republic of Korea, were recorded saying the five vowels with recording durations of 440 ms. We used the vowel sounds from 60 students as training data and from the remaining 15 students as test data (figure 5(a)). In general, it is known that humans recognize speech depending on how the first (F1) and second (F2) formants are formed [26, 27]. The first formant (F1) of this database ranged from 200 to 900 Hz, and the second formant (F2) ranged from 500 to 3000 Hz [25]. While the F1 of all vowels can be covered by the frequency response range of our ABM, the ABM does

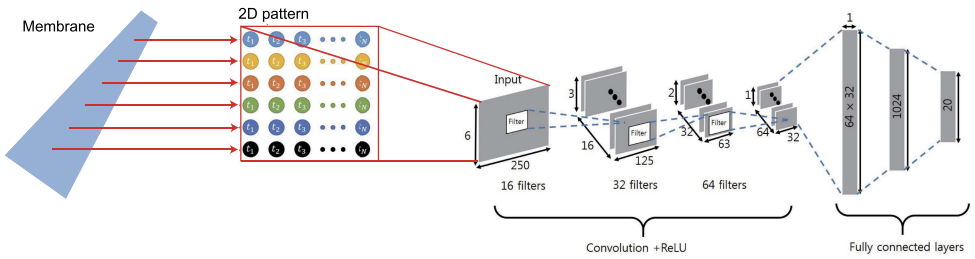

Figure 3. Overview of the membrane-neural network system. Sound is applied to the artificial membrane (left) and the resulting movements are detected and converted into a 2D pattern (center), with rows representing membrane position and columns representing time series data. This pattern is used to train a CNN (right). The network structure depicted is 2D Network 1 in supplementary information (SI) (https://stacks.iop.org/BB/16/26012/mmedia).

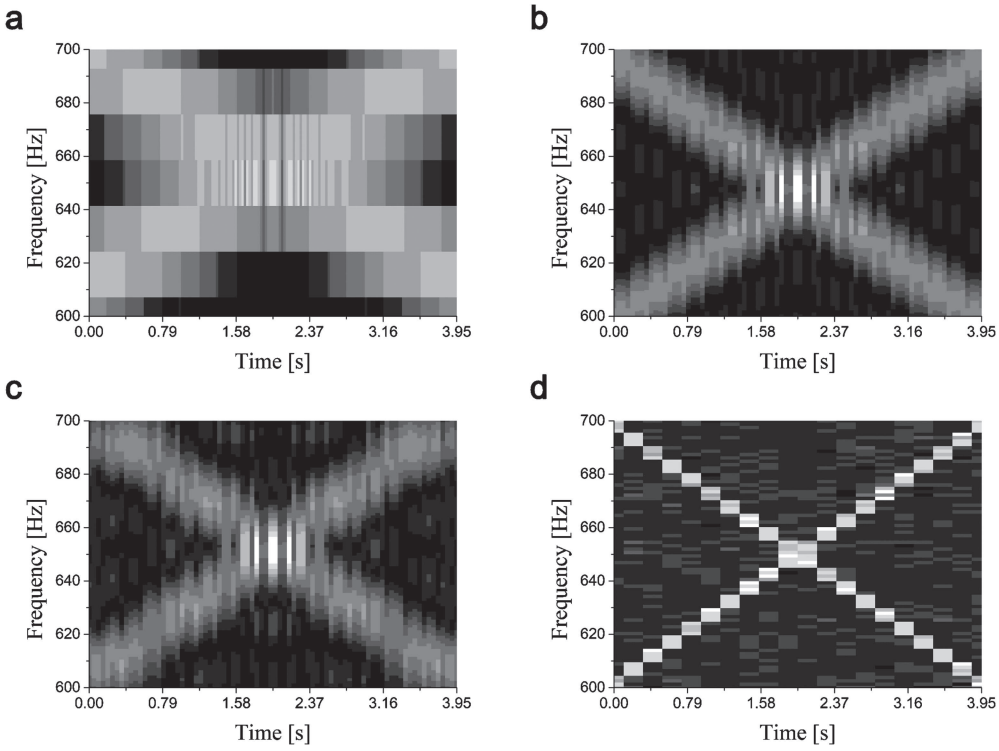

Figure 4. Analysis of signals with two frequency components of 50 ms duration, in 5 Hz intervals, (a) using DFT, (b) using ZFFT, where the input data is down-sampled with a sampling rate of $\frac{1}{20\times\Delta t}$, and (c) using CZT. (d) Results using the ABM combined with the neural network analysis.

not cover the F2 range for all the vowels, as the F2 range for [e] and [i] is 1500 to 3000 Hz. Nevertheless, we find that our system successfully identifies all vowels. One may think that the high-frequency information of F2 is already encoded in the time-series signals of the low-frequency channels, but this is not possible, as high-frequency components above 1400 Hz do not appear in the ABM response (see figure S7 in the SI). Instead, the ability to distinguish the vowels comes from the membrane pattern recognition of the multichannel response, of which some frequency components are lower than F2. This implies that perhaps humans can also distinguish vowels through the low-frequency BM response, as can occur when the high-frequency regions of the BM are damaged. Reference [28] shows the results of vowel recognition of hearing-imapred listeners. In the paper, the hearing-impaired listeners significantly distinguished five English vowels, despite the difficulty of distinguishing F2, and the error rate is similar to the results of the ABM (Error rate is approximately 15%). Therefore, it can be said that while the normal-hearing listener uses both F1 and F2 information to classify the vowel, the hearing-impaired listener uses F1 cues similar to our ABM to distinguish vowels.

We compare the vowel recognition rate of a convolution neural network using different inputs to train the CNN. These include the ABM pattern as well as a number of other common methods used to identify spoken sounds. These include using the raw data (with no pre-processing) as the CNN input, the outputs of the CZT and ZFFT transforms, and a

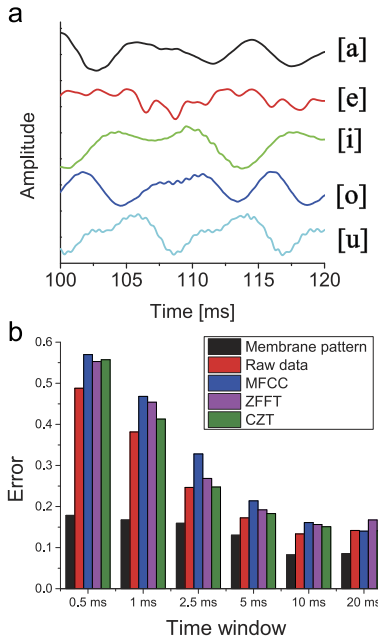

Figure 5. (a) Time series raw acoustic signals collected by a microphone, for the spoken vowels [a], [e], [i], [o], and [u]. (b) Test error rates of the trained CNN, using different time windows for the five vowels, using each of the five data pre-processing methods. The pattern recognition errors from the multi-position detection (black) are compared with the errors from the raw time series data (red), conventional MFCC analysis (blue), ZFFT analysis (magenta), and CZT analysis (green). The horizontal axis shows the size of the time window.

conventional Mel-frequency cepstral coefficient (MFCC) analysis. The MFCC method provides a representation of the power spectrum of a sound, using a cosine transform of the log power spectrum on a nonlinear Mel frequency scale [29]. For the MFCC and raw data recognition, we used the neural network structure in reference [25]; using this, we achieved a 91% accuracy for a 50 ms time window. For the CZT and ZFFT analyses, we set the frequency band to 200 Hz to 3000 Hz in order to cover most of the vowel frequency components. When using the ABM to transform sound, we measured the ABM membrane displacement patterns using ten lasers, where position 1 is located 4 cm from the apex, and positions 2–10 are spaced evenly by 0.8 cm toward the base. The CNN is trained for each time window with a unique data set, then tested with different data.

As shown in figure 5(b), we find that all methods give roughly equivalent performance when using relatively long time windows (5–20 ms), while the performance of the ABM displacement window is clearly superior for shorter time windows, especially below 2.5 ms.

We find that as the vowel sampling time window decreases, the accuracy of each of the standard methods decreases drastically, to about 50%, while the ABM pattern accuracy is only slight reduced, from 90% to 80%. When the Fourier components of a speech signal are analyzed in a short time window, the frequency resolution is poor due to the resultingly poor frequency resolution. The ABM pattern recognition sidesteps this limitation, because the ABM pattern recognition involves both spatial and temporal information, where the spatial response is convolved with some frequency information, even for very brief excitations.

To understand qualitatively why the mechanical response pattern improves the performance, we consider simplified model of the membrane as a harmonic oscillator with mass m, friction constant γ, and angular resonance frequency ω_0. The displacement of the oscillator $x(t)$ in response to a driving force $F(t)$ follows the Newtonian equation of motion, with displacement

$$x(t) = \int_{-\infty}^{t} \mathrm{d}t' G(t-t')F(t')/m, \tag{3}$$

where the Green function response $G(t-t')$ is given by

$$G(t-t') = \frac{1}{\pi}\exp(-\frac{\gamma}{2}(t-t'))\frac{\cos\left[\sqrt{4\omega_0^2-\gamma^2}(t-t')\right]}{\sqrt{4\omega_0^2-\gamma^2}}. \tag{4}$$

Note that $G \propto \exp(-\gamma(t-t')/2)$, so the displacement $x(t)$ at time t integrates the sound over a preceding time window $\sim 2/\gamma$. In the ABM, measurements of the mechanical Q factor indicate that $2/\gamma \sim 1$ ms, so the displacement $x(t)$ of the oscillator encodes roughly that length of the sound signal.

The integrated response is however not sufficient to explain the dramatic reduction of the error rate of the ABM for short time windows, shown in figure 5. The ABM membrane's motion is however sensed in multiple locations, each of which can be roughly modeled as an independent harmonic oscillator, with a different resonance frequency. The sound identification involves the pattern from many (in our experiment, $n = 10$) such oscillators. In auditory science, this position-dependent frequency analysis is called 'place coding'. The ABM-enabled place coding used here, combined with the known strong pattern recognition of our neural network, is what enables the good sound recognition capability demonstrated in this work.

6. Conclusion

We have demonstrated an ABM with a frequency-dependent spatial response, which when combined with a trained pattern-recognizing neural network shows outstanding performance in frequency resolution as well as speech recognition of short-time phonemes. This artificial system mimics the human auditory system's ability to resolve and distinguish very short segments of speech, and shows strong

potential for the analysis of more complex spoken sounds. We believe the superior performance of our system in distinguishing vowels, compared to other frequency-analyzing methods, stems from the system's ability to recognize minute differences in the spectra of [o] and [u] (see figure S7 in the SI). To discern small differences using Fourier transform methods requires long integration times; however, as we demonstrated in the two-tone experiments, a pre-trained pattern recognition system can overcome this limitation, a capability that successfully transfers to vowel recognition as well. As the purpose of this study was to investigate the auditory signal processing of humans using a biomimetic system, future applications in speech recognition through membrane-based speech signal processing may become possible, even though the practical application of this approach is uncertain. Compared to other spectral zoom methods (ZFFT and CZT), the pattern recognition of the ABM can improve speech recognition in short time intervals. This may provide an important clue as to why real-time speech recognition is possible for humans but still challenging using artificial systems.

Acknowledgments

We thank Jong Hun Park for the information on CZT. This work was supported by a National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP), NRF2017R1A2B3010002.

ORCID iDs

Woo Seok Lee https://orcid.org/0000-0002-9381-1253
Kang-Hun Ahn https://orcid.org/0000-0003-2358-593X

References

[1] Dallos P 1996 Overview: cochlear neurobiology *The Cochlea* (Berlin: Springer) pp 1–43
[2] Robles L and Ruggero M A 2001 Mechanics of the mammalian cochlea *Physiol. Rev.* **81** 1305–52
[3] Ren T 2002 Longitudinal pattern of basilar membrane vibration in the sensitive cochlea *Proc. Natl Acad. Sci.* **99** 17101–6
[4] Mohamed A-r, Dahl G E and Hinton G 2012 Acoustic modeling using deep belief networks *IEEE Trans. Audio Speech Lang. Process.* **20** 14–22
[5] Hoshen Y, Weiss R J and Wilson K W 2015 Speech acoustic modeling from raw multichannel waveforms *2015 IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)* (Piscataway, NJ) pp 4624–8
[6] Seide F, Li G and Yu D 2011 Conversational speech transcription using context-dependent deep neural networks *12th Annual Conf. of the Int. Speech Communication Association*
[7] Arisoy E, Sainath T N, Kingsbury B and Ramabhadran B 2012 Deep neural network language models *Proc. of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-Gram Model? On the Future of Language Modeling for HLT* (Association for Computational Linguistics) pp 20–8
[8] Lyons R G 2004 *Understanding Digital Signal Processing* 3rd edn (Englewood Cliffs, NJ: Prentice-Hall)
[9] Porat B 1996 *A Course in Digital Signal Processing* (New York: Wiley)
[10] Proakis J G 2001 *Digital Signal Processing* (*Principles Algorithms and Applications*) (London: Pearson Education)
[11] Von Békésy G and Wever E G 1960 *Experiments in Hearing* vol 8 (New York: McGraw-Hill)
[12] Von Békésy G 1970 Travelling waves as frequency analysers in the cochlea *Nature* **225** 1207
[13] White R D and Grosh K 2005 Microengineered hydromechanical cochlear model *Proc. Natl Acad. Sci.* **102** 1296–301
[14] Chen F, Cohen H I, Bifano T G, Castle J, Fortin J, Kapusta C, Mountain D C, Zosuls A and Hubbard A E 2006 A hydromechanical biomimetic cochlea: experiments and models *J. Acoust. Soc. Am.* **119** 394–405
[15] Shintaku H, Nakagawa T, Kitagawa D, Tanujaya H, Kawano S and Ito J 2010 Development of piezoelectric acoustic sensor with frequency selectivity for artificial cochlea *Sens. Actuators,* A **158** 183–92
[16] Wittbrodt M J, Steele C R and Puria S 2006 Developing a physical model of the human cochlea using microfabrication methods *Audiol. Neurotol.* **11** 104–12
[17] Jung Y, Kwak J-H, Lee Y, Kim W and Hur S 2013 Development of a multi-channel piezoelectric acoustic sensor based on an artificial basilar membrane *Sensors* **14** 117–28
[18] Jang J, Lee J, Jang J H and Choi H 2016 A triboelectric-based artificial basilar membrane to mimic cochlear tonotopy *Adv. Healthcare Mater.* **5** 2481–7
[19] Jeon H, Jang J, Kim S and Choi H 2018 Characterization of a piezoelectric ALN beam array in air and fluid for an artificial basilar membrane *Electron. Mater. Lett.* **14** 101–11
[20] Gong S *et al* 2020 A soft resistive acoustic sensor based on suspended standing nanowire membranes with point crack design *Adv. Funct. Mater.* **30** 1910717
[21] LeCun Y, Haffner P, Bottou L and Bengio Y 1999 Object recognition with gradient-based learning *Shape, Contour and Grouping in Computer Vision* (Berlin: Springer) pp 319–45
[22] Krizhevsky A, Sutskever I and Hinton G E 2012 Imagenet classification with deep convolutional neural networks *Advances in Neural Information Processing Systems* pp 1097–105
[23] Nair V and Hinton G E 2010 Rectified linear units improve restricted Boltzmann machines *Proc. of the 27th Int. Conf. on Machine Learning (ICML-10)* pp 807–14
[24] Rumelhart D E, Hinton G E and Williams R J 1985 Learning internal representations by error propagation *Technical Report* California University San Diego La Jolla Institute for Cognitive Science
[25] Kim H, Lee W S, Yoo J, Park M and Ahn K H 2019 Origin of the higher difficulty in the recognition of vowels compared to handwritten digits in deep neural networks *J. Kor. Phys. Soc.* **74** 12–8
[26] Thomas I B 1968 The influence of first and second formants on the intelligibility of clipped speech *J. Audio Eng. Soc.* **16** 182–5
[27] Holmes J N, Holmes W J and Garner P N 1997 Using formant frequencies in speech recognition *Fifth European Conf. on Speech Communication and Technology*
[28] Richie C, Kewley-Port D and Coughlin M 2003 Discrimination and identification of vowels by young, hearing-impaired adults *J. Acoust. Soc. Am.* **114** 2923
[29] Sahidullah M and Saha G 2012 Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition *Speech Commun.* **54** 543–65