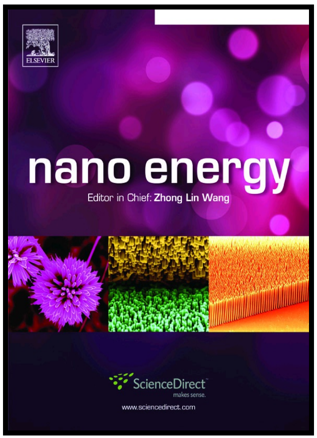

www.elsevier.com/locate/nanoenergy

Machine Learning-based Self-powered Acoustic Sensor for Speaker Recognition

Jae Hyun Han, Kang Min Bae, Seong Kwang Hong, Hyunsin Park, Jun-Hyuk Kwak, Hee Seung Wang, Daniel Juhyung Joe, Jung Hwan Park, Young Hoon Jung, Shin Hur, Chang D. Yoo, Keon Jae Lee

PII: S2211-2855(18)30674-8
DOI: https://doi.org/10.1016/j.nanoen.2018.09.030
Reference: NANOEN3035

To appear in: *Nano Energy*

Received date: 8 August 2018
Revised date: 23 August 2018
Accepted date: 14 September 2018

Cite this article as: Jae Hyun Han, Kang Min Bae, Seong Kwang Hong, Hyunsin Park, Jun-Hyuk Kwak, Hee Seung Wang, Daniel Juhyung Joe, Jung Hwan Park, Young Hoon Jung, Shin Hur, Chang D. Yoo and Keon Jae Lee, Machine Learning-based Self-powered Acoustic Sensor for Speaker Recognition, *Nano Energy,* https://doi.org/10.1016/j.nanoen.2018.09.030

Machine Learning-based Self-powered Acoustic Sensor for Speaker Recognition

Jae Hyun Han [1,a], *Kang Min Bae* [1,b], *Seong Kwang Hong* [1,a], *Hyunsin Park* [b], *Jun-Hyuk Kwak* [c], *Hee Seung Wang* [a], *Daniel Juhyung Joe* [a], *Jung Hwan Park* [a], *Young Hoon Jung* [a], *Shin Hur* [c], *Chang D. Yoo* [b,*], *Keon Jae Lee* [a,*]

[a]Department of Materials Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea.

[b]Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST), 291, Daehak-ro, Yuseong-gu, Daejeon 34141, Republic of Korea.

[c]Department of Nature-Inspired Nanoconvergence System, Korea Institute of Machinery and Materials (KIMM), 156 Gajeongbuk-ro, Yuseong-gu, Daejeon, 34103, Republic of Korea.

[*] Corresponding authors.

Email addresses: cd_yoo@kaist.ac.kr (C.D. Yoo), keonlee@kaist.ac.kr (K.J. Lee).

[1] These authors contributed equally to this work.

ABSTRACT

Herein, we report a new platform of machine learning-based speaker recognition via the flexible piezoelectric acoustic sensor (f-PAS) with a highly sensitive multi-resonant frequency band. The resonant self-powered f-PAS was fabricated by mimicking the operating mechanism of the basilar membrane in the human cochlear. The f-PAS acquired abundant voice information from the multi-channel sound inputs. The standard TIDIGITS dataset were recorded by the f-PAS and converted to frequency components by using a Fast Fourier Transform (FFT) and a Short-Time Fourier Transform (STFT). The machine learning based Gaussian Mixture Model (GMM) was designed by utilizing the most highest and second highest sensitivity data among multi-channel outputs, exhibiting outstanding speaker recognition rate of 97.5 % with error rate reduction of 75 % compared to that of the reference MEMS microphone.

KEYWORDS: Flexible piezoelectric; Self-powered; Acoustic sensor; Machine learning algorithm; Speaker recognition;

1. Introduction

Voice recognition is the most intuitive user-interface for bilateral communication between humans and smart devices [1–4]. Speaker recognition has received spotlight as a next big thing of voice user interface (VUI) such as personalized voice-controlled assistant, smart home appliance, biometric authentication based on artificial intelligence (AI) and internet of things (IoT) infrastructure [5–7].

The conventional speaker recognition was realized by a condenser type microphone, which detects sound by measuring the capacitance value between two conducting layers while supplying continuous power. The condenser type microphone, however, has critical demerits such as low sensitivity, high power consumption, and an unstable circuit due to the large gain of amplification [8–10]. Speaker recognition also suffers from a low recognition rate, caused by limited voice information and optimal algorithms for a simple and accurate process. In addition, the conditions for speaker recognition involve complicated processes such as deep learning of big data, massive training, speaker distance, and background noise [11]. Although complex algorithms are utilized to solve these issues, fundamental research based on intrinsic sound information is required to enhance the voice recognition rate as the facile human mechanism of distinguishing speakers [12,13].

Recently, our group has developed a multi-channel flexible piezoelectric acoustic sensor that can detect three discrete resonant frequencies from sound waveforms inspired by the basilar membrane of the human cochlear [14–17]. Although this acoustic sensor provides sensitive resonant voice information, practical applications have not been demonstrated due to a lack of device capability and a voice processing algorithm.

Herein, we reported a machine learning-based speaker recognition system enabled by a self-powered flexible piezoelectric acoustic sensor (f-PAS) with a highly sensitive multi-resonant frequency band. A flexible piezoelectric membrane was employed using inorganic-based laser lift-off (ILLO) to fabricate the basilar membrane (BM)-inspired f-PAS. The speech waveforms of standard TIDIGITS dataset were recorded by the multi-channel f-PAS and converted into frequency domain signals by using Fast Fourier Transform (FFT) and a Short-Time Fourier Transform (STFT) to obtain the characteristics of the frequency components. A multi-channel machine learning algorithm based on a Gaussian Mixture Model (GMM) was utilized for speaker recognition, resulted in a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot of STFT feature between training dataset and testing utterance. Finally, the f-PAS achieved a 97.5 % speaker recognition rate with the 75 % reduction of error rate compared to that of the reference MEMS microphone.

2. Experimental section

2.1. Fabrication of the f-PAS device

Spin-coating of a PZT chemical solution (MEMS Solution, Inc., soultion concentration of 0.4 M) was conducted at 3000 rpm for 20 s, for making a PZT thin film thickness of 1 μm. A PZT solution was coated on the rigid sapphire (Hi-Solar Co.), and subsequently heat treated at 600 °C for 80 min by RTA for crystallization of the PZT thin film. After the crystallization, the O_2 plasma was treated on the surface of PZT thin film by using inductively coupled plasma-reactive ion etching (ICP-RIE, SNTEK Co.). The ultraviolet (UV) curable adhesive was spin-coated on PZT membrane to attach the polymer membrane with the thickness of 50 μm. After that, the inorganic-based laser lift-off (ILLO) process (XeCl excimer laser, wavelength of 308 nm, scanning frequency of 9 Hz and 650 × 650 μm^2) was conducted to separate the rigid sapphire substrate and transfer to the flexible polymer membrane. Electrodes (Cr and Au electordes, 15 nm and 120 nm of thickness, respectively) were deposited on PZT membrane by sputtering process. The conventional lithography and etching process were applied to pattern the seven channels of IDEs (2 mm line spacing, 100 μm intergap and 60 μm width). A UV sensitive PU passivation layer was coated on the patterned multi channel f-PAS to protect the polymer membane and PZT thin film. Finally, passivated f-PAS device was then interconnected on the multi channel PCB by appling the silver paste.

2.2. Electrical signal measurements

The electrical signals of the f-PAS from the sound waveforms were measured by a National Instruments (NI) Sound Module and a Keithley sourcemeter (Keithley 2612A). A speaker and a function generator were used to apply the monochromatic waveform, frequency sweep (white noise), and voice sounds of TIDIGITS dataset to the f-PAS.

2.3. Measurement of voice signals from f-PAS and commercial MEMS microphone

To record the signals from the f-PAS and the reference MEMS microphone in the commercialized cellular phone (Samsung, Galaxy S8), they were located above the same distance from the speaker. The signals from f-PAS and commercial MEMS microphone were measured by applying the TIDIGITS sound signals. All voice signals were emitted through the speaker, and measured through the National Instruments (NI) Sound Module. Sampling frequency of measured electrical signals from both f-PAS and the reference MEMS microphone was fixed as 8 kHz, with the direct connection of f-PAS and MEMS microphone to the NI Sound Module.

2.4. GMM algorithm for speaker recognition

GMM method with a weighted sum of multiple probability density functions (PDF) was used including mean vector and covariance matrix. The covariance matrix was constrained to be diagonal to avoid overfitting of GMM process. The parameters of GMM algorithm were learned using Expectation Maximization (EM) algorithm, which is an iterative method to find maximum likelihood or maximum posteriori of parameters.

2.5. Comparison of speaker recognition rate between f-PAS and commercial MEMS microphone

From the speaker recognition process via GMM algorithm, speaker recognition rates of f-PAS and commercial MEMS microphone was compared depending on the increasing of mixture numbers. For the acquisition of training dataset, each electrical signal from the f-PAS and MEMS microphone was measured by applying the TIDIGITS sound signals. Consequently, the f-PAS training dataset was trained by using the f-PAS electrical signals, and the commercial MEMS mic training dataset was trained by using the MEMS microphone signals, respectively. Error rates of f-PAS and commercial MEMS microphone were calcuated from the simple equation of 100 (%) – recognition rate (%).

3. Results

3.1. Machine learning-based speaker recognition system from multi-channel f-PAS

Figure 1a schematically illustrates the overall concept of a machine learning-based speaker recognition enabled by the multi-channel f-PAS. A flexible $Pb(Zr_{0.52}Ti_{0.48})O_3$ (PZT) membrane providing piezoelectric responses was fabricated by utilizing a similar protocol of ILLO as our previous paper [18–21]. A PZT sol-gel film was deposited on a sapphire substrate by spin coating and then crystallized using rapid thermal annealing (RTA) at 700 ºC for 75 min [22]. The inorganic piezoelectric thin film was detached from the mother sapphire substrate by irradiating an excimer laser of XeCl (wavelength of 308 nm, scanning frequency of 9 Hz, and photon energy of 4.03 eV) from the backside of the mother substrate, and subsequently transferred onto a polymer membrane of 50 μm thickness [17,23,24]. A concaved trapezoidal design was applied to the f-PAS to produce multiple piezoelectric signals. Compared to our previous piezoelectric acoustic sensor with large size and sharp three resonant frequencies, the multi-resonant band f-PAS was integrated in a small single chip that covered entire voice frequency range with high sensitivity [17]. Seven channels of interdigitated electrodes were patterned on the PZT membrane using photolithography and an etching process [25]. The self-powered flexible piezoelectric membrane was attached to a printed circuit board (PCB) to generate an electrical signal by the displacement vibration. Resonant oscillations in the flexible piezoelectric membrane were produced by human voice stimulation of the TIDIGITS dataset, which is composed of acoustic signals of digit sequences to evaluate voice recognition [26]. The multiple piezoelectric signals were generated by a resonant vibration of multi-channel self-powered acoustic sensor, acquiring abundant voice information [27]. Machine learning-based training of TIDIGITS voice data

(40 people, 2800 speeches) was performed by using a GMM algorithm, which is modified for multi signal processing [28–30]. Finally, the speeches of a random person were compared with the trained dataset for speaker recognition. Figure 1b shows a photograph and schematic illustration of the multi-channel f-PAS and corresponding electrical output generated by sound input. The f-PAS output voltage as a function of time was generated from standard male speech of TIDIGITS (Male, voice of 84149) [26], indicating that the piezoelectric acoustic sensor successfully recorded the human voice. Note that our self-powered sensing of sound waveforms is operated without an external power supply [31], which is a significant advantages for low power mobile or always-on IoT applications [32-34]. In contrast, commercialized condenser type acoustic sensors require continuous external power source to measure the capacitance between two conducting layers even when they are not sensing the voice.

3.2. Experimental setup and electrical output signals of f-PAS from incoming sound wave

Figure 2a shows a photograph of the sound recording experimental setup for speaker recognition. An anechoic chamber consisting of acoustic absorbent and a stationary table was employed to remove the external noise and sound waveforms reflection [35,36]. A speaker was placed under the f-PAS (and commercialized reference microphone) to input a frequency sweep from 100 to 4000 Hz at 94 dB sound pressure level (SPL) via a function generator. Relative responses (log scale) of the f-PAS were measured by a PXI sound module, and then normalized to 0 dB at 1 kHz from the average of seven channels. Figure 2b presents the relative responses of the multi-channel f-PAS over the voice frequency range (from 100 Hz to 4 kHz) under the condition of white noise, a random signal of equal intensity. Among the seven channels, the relative responses from four channels (2, 4, 5 and 6) were plotted to present the different frequency response of the corresponding channels. Multiple resonant piezoelectric outputs were measured at various resonance frequencies of 630, 1060, 1290 and 1710 Hz. The highest relative responses of 26.7 dB and 23.5 dB could be measured through channels 2 and 5 at 630 Hz and 1060 Hz, respectively. The maximum relative sensitivity of channel 2 presented 30 mV at 1kHz as shown in Figure S1, which is higher than that of the MEMS microphone [37]. This high sensitivity of f-PAS enables the applications such as speaker recognition, which was limited to the previous acoustic sensors [17]. The resonance-based multi-channel f-PAS could recognize minute sound information compared to the highly amplified MEMS microphone, providing ample voice input for accurate speaker recognition as expressed in Figure S2. Relative responses from other channels (1, 3 and 7) are shown in Figure S3. Note that the optimal combination from discrete portions of multi-channel frequency signals can be intentionally selected for increasing a speaker recognition rate. Figure 2c shows sound signals and frequency domain data of original TIDIGITS speech and the f-PAS. Figure 2c-i shows original sound waveform of TIDIGITS speech (Woman, voice of 063). This sound signal of the time domain was converted into the frequency domain by using a FFT and a STFT. Figure 2c-ii presents the sound waveform, FFT and STFT of the same TIDIGITS speech recorded by channel 2 of the f-PAS. The FFT algorithm represented the sound information as a function of the frequency domain, showing the major frequency components contained in the acoustic signal. The STFT algorithm contains multiple frames of the sound FFT signal over a short shifting period to analyze the time varying characteristics of frequency domain signals. The definition of STFT is given by equation (1) [38],

$$\mathrm{X(k,n)} = \sum_{\mathrm{m}=-\infty}^{\infty} w(m)x(T*n-m)e^{-j\frac{2\pi k}{N}m} \tag{1}$$

where x is the original signal, N is the window size, T is the window shift period and w is the window function. Both commercial MEMS microphone and f-PAS go through the same noise reduction of GMM algorithm called hamming window. Hamming window was used to prevent the spectrum distortion during the STFT process. The voice signal converted to STFT was utilized to perform speaker recognition using the GMM machine learning algorithm. Similar data obtained from channel 5 are shown in Figure S4. Note that the slight difference in the acoustic signals between the original sound and the f-PAS is ascribed to the non-flat response of the resonant f-PAS frequency band.

3.3. GMM-based machine learning algorithm for the speaker recognition of f-PAS

Figure 3a shows a flow chart of GMM based machine learning to process speaker recognition from the multi-channel voice data of the f-PAS. The GMM based algorithm consists of training and testing using the standard TIDIGITS dataset (20 men and 20 women speakers, 77 speeches per each speaker, a total of 3080 voice data). The forty speakers randomly selected from the TIDIGITS dataset were recorded by the multi-channel f-PAS and then converted to STFT features, which is calculated by weighted summation of STFT absolute value. Among the 3080 TIDIGITS datasets of speaker recognition algorithm, the 2800 data were used for training data and the remaining 280 data were utilized for testing data. A statistical GMM method depending on the number of mixture was used for determining speaker recognition, as a simple illustration of Figure S5. A mixture was defined as the number of Gaussian distributions representing STFT feature. Figure 3b shows the trained STFT features visualized in t-SNE plot. The t-SNE plot embeds high-dimensional data of similar objects into a low-dimensional space, which is related with probability distribution of similar clusters among 2800 training data. Figure 3c presents the majority voting method used to test the speaker recognition process. When an input test signal is given, the STFT method generates multiple frames of FFT signals over time period. The likelihood of the test speaker was calculated according to each frame by comparing the training and testing results, and finally majority voted speaker was determined as the frame increases. This speaker recognition process can be expressed by below equation (2) [39],

$$\hat{C} = \arg\max_{C} \sum_{n=1}^{M} 1[y_n = C] \tag{2}$$

where y denotes the predicted class for each frame and M denotes the number of frames in a STFT. The inset shows a schematic illustration of the speaker recognition process. Every speaker recognition was evaluated by averaging 10 iterations of the speaker recognition rate to increase the reliability, as described in our previous reports of machine learning algorithms [40–43].

3.4. Frequency responses of f-PAS and comparison of speaker recognition rate between f-PAS and commercial MEMS microphone

Figure 4a presents the relative responses of the multi-channel f-PAS under a frequency sweep from 100 to 4000 Hz at the 94 dB SPL condition. As shown in two-averaging (red line) and seven-averaging (blue line), the selection of the channels for speaker recognition was determined using the frequency responses of the multi-channel f-PAS. Two-averaging was acquired by averaging the most highest and the second highest responses selected from seven channel outputs as demonstrated in Figure S6, whereas seven-averaging was obtained by averaging all piezoelectric signals of the f-PAS. Note that two-averaging and seven-averaging have the same amount of data, but seven-averaging process includes low sensitive signals compared to two-averaging process. The two-averaging exhibited higher sensitivity compared to the seven-averaging (e.g., 26.7 dB higher at 650Hz) over the entire voice frequency band. Figure 4b presents the GMM algorithm based speaker recognition rate of two-averaging, seven-averaging and a commercialized reference MEMS microphone according to the number of mixture. The speaker recognition rates of the two-averaging, seven-averaging and MEMS were increased and saturated as the number of mixture was increased to 30, exhibiting 97.5, 92, and 90 % recognition rate, respectively. When a small number of mixtures was applied, the speaker recognition rate of the MEMS microphone was higher than that of two-averaging and seven-averaging, respectively. However, as the number of mixture was increased, the recognition rate of the f-PAS surpassed that of the reference microphone due to the fact that the multi-channel f-PAS had more voice information for Gaussian profiles compared to MEMS sensor. Figure 4c shows the speaker recognition error rate of two-averaging, seven-averaging, and the MEMS microphone, calculated from the Figure 4b of 30 mixtures. The error rate of the two-averaging was 2.5 %, exhibiting 75 % lower than that of the reference MEMS microphone. This excellent error rate of speaker recognition was attributed to multi-channel sound signals and the highly sensitive response of the two-averaging f-PAS data.

4. Conclusions

In summary, we successfully demonstrated a machine learning-based speaker recognition system using a flexible piezoelectric acoustic sensor by mimicking the operation mechanism of the basilar membrane. The f-PAS can obtain the abundant and intrinsic voice information from the highly sensitive multi-channel membrane, which is beneficial for identifying speakers. Incoming voice information of TIDIGITS dataset were converted into FFT and STFT multi-data to acquire the frequency characteristics of the human voice. The GMM based speaker recognition algorithm was implemented for the training of 2800 data size, which is embedded in the t-SNE plot of probability distribution. The 280 test process was performed by comparing the STFT features and predicting the speaker over time frames. Finally, the f-PAS achieved a 97.5 % speaker recognition rate, which is 75 % reduction of the error rate compared to that of the commercialized MEMS microphone. The significant improvement of the speaker recognition rate indicates that the f-PAS application can be further extended to voice-based biometric authentication and highly accurate speech recognition, which we are currently investigating.

Acknowledgements

This research was supported by the Wearable Platform Materials Technology Center (WMC) (NRF-2016R1A5A1009926) and Nano·Material Technology Development Program (NRF-2016M3A7B4905621). This was also supported by the Center for Integrated Smart Sensors funded by the Ministry of Science and ICT as Global Frontier Project (CISS-2016M3A6A6929958). This work was also supported by the ICT R&D program of MSIP/IITP [2016-0-00563, Research on Adaptive Machine Learning Technology Development for Intelligent Autonomous Digital Companion] and Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (2017-0-01780, The technology development for event recognition/relational reasoning and learning knowledge based system for video understanding).

References

[1] E. Formisano, F. De Martino, M. Bonte, R. Goebel, “Who” is saying “what”? Brain-based decoding of human voice and speech, Science 322 (2008) 970–973.

[2] T.K. Perrachione, S.N. Del Tufo, J.D.E. Gabrieli, Human voice recognition depends on language ability, Science 333 (2011) 595.

[3] H. Guo, X. Pu, J. Chen, Y. Meng, M.-H. Yeh, G. Liu, Q. Tang, B. Chen, D. Liu, S. Qi, C. Wu, C. Hu, J. Wang, Z. L. Wang, A highly sensitive, self-powered triboelectric auditory sensor for social robotics and hearing aids. Sci. Robot. 3 (2018) eaat2516.

[4] X. Pu, H. Guo, J. Chen, X. Wang, Y. Xi, C. Hu, Z. L. Wang, Eye motion triggered self-powered mechnosensational communication system using triboelectric nanogenerator. Sci. Adv. 3 (2017) e1700694.

[5] R. Blossey, Self-cleaning surfaces - Virtual realities, Nat. Mater. 2 (2003) 301–306.

[6] D.J. Ward, D.J.C. MacKay, Artificial intelligence: Fast hands-free writing by gaze direction, Nature. 418 (2002) 838.

[7] J.P. Campbell, Speaker recognition: A tutorial, Proc. IEEE. 85 (1997) 1437–1462.

[8] W. Sui, W. Zhang, K. Song, C.H. Cheng, Y.K. Lee, Breaking the size barrier of capacitive MEMS microphones from critical length scale, TRANSDUCERS 2017 - 19th Int. Conf. Solid-State Sensors, Actuators Microsystems (2017) 946–949.

[9] N. Mohamad, Modelling and Optimisation of a Spring-Supported Diaphragm Capacitive MEMS Microphone, Engineering 02 (2010) 762–770.

[10] J.W. Weigold, T.J. Brosnihan, J. Bergeron, X. Zhang, A MEMS Condenser Microphone for Consumer Applications, 19th IEEE Int. Conf. Micro Electro Mech. Syst. (2006) 86–89.

[11] J.H.L. Hansen, T. Hasan, Speaker recognition by machines and humans: A tutorial review, IEEE Signal Process. Mag. 32 (2015) 74-99.

[12] M. Mills, E. Melhuish, Recognition of mother’s voice in early infancy, Nature 252 (1974) 123.

[13] J. Yang, J. Chen, Y. Su, Q. Jing, Z. Li, F. Yi, X. Wen, Z. Wang, Z.L. Wang, Eardrum-inspired active sensors for self-powered cardiovascular system characterization and throat-attached anti-interference voice recognition, Adv. Mater. 27 (2015) 1316-1326.

[14] T. Inaoka, H. Shintaku, T. Nakagawa, S. Kawano, H. Ogita, T. Sakamoto, S. Hamanishi, H. Wada, J. Ito, Piezoelectric materials mimic the function of the cochlear sensory epithelium, Proc. Natl. Acad. Sci. 108 (2011) 18390–18395.

[15] P. Belin, R.J. Zatorre, P. Lafaille, P. Ahad, B. Pike, Voice-selective areas in human auditory cortex, Nature 403 (2000) 309–312.

[16] G. Von Békésy, Travelling Waves as Frequency Analysers in the Cochlea, Nature 225 (1970) 1207–1209.

[17] H.S. Lee, J. Chung, G.T. Hwang, C.K. Jeong, Y. Jung, J.H. Kwak, H. Kang, M. Byun, W.D. Kim, S. Hur, S.H. Oh, K.J. Lee, Flexible inorganic piezoelectric acoustic nanosensors for biomimetic artificial hair cells, Adv. Funct. Mater. 24 (2014) 6914–6921.

[18] H. Palneedi, J.H. Park, D. Maurya, M. Peddigari, G.T. Hwang, V. Annapureddy, J.W. Kim, J.J. Choi, B.D. Hahn, S. Priya, K.J. Lee, J. Ryu, Laser Irradiation of Metal Oxide Films and Nanostructures: Applications and Advances, Adv. Mater. 30 (2018) 1705148.

[19] D.J. Joe, S. Kim, J.H. Park, D.Y. Park, H.E. Lee, T.H. Im, I. Choi, R.S. Ruoff, K.J. Lee, Laser–Material Interactions for Flexible Applications, Adv. Mater. 29 (2017) 1606586.

[20] S. Kim, J.H. Son, S.H. Lee, B.K. You, K. Il Park, H.K. Lee, M. Byun, K.J. Lee, Flexible crossbar-structured resistive memory arrays on plastic substrates via inorganic-based laser lift-off, Adv. Mater. 26 (2014) 7480–7487.

[21] H.E. Lee, S. Kim, J. Ko, H.I. Yeom, C.W. Byun, S.H. Lee, D.J. Joe, T.H. Im, S.H.K. Park, K.J. Lee, Skin-Like Oxide Thin-Film Transistors for Transparent Displays, Adv. Funct. Mater. 26 (2016) 6170–6178.

[22] C.K. Jeong, S.B. Cho, J.H. Han, D.Y. Park, S. Yang, K. Il Park, J. Ryu, H. Sohn, Y.C. Chung, K.J. Lee, Flexible highly-effective energy harvester via crystallographic and computational control of nanointerfacial morphotropic piezoelectric thin film, Nano Res. 10 (2017) 437–455.

[23] K. Il Park, J.H. Son, G.T. Hwang, C.K. Jeong, J. Ryu, M. Koo, I. Choi, S.H. Lee, M. Byun, Z.L. Wang, K.J. Lee, Highly-efficient, flexible piezoelectric PZT thin film nanogenerator on plastic substrates, Adv. Mater. 26 (2014) 2514–2520.

[24] I. Choi, H.Y. Jeong, H. Shin, G. Kang, M. Byun, H. Kim, A.M. Chitu, J.S. Im, R.S. Ruoff, S.Y. Choi, K.J. Lee, Laser-induced phase separation of silicon carbide, Nat. Commun. 7 (2016) 13562.

[25] B.H. Mun, B.K. You, S.R. Yang, H.G. Yoo, J.M. Kim, W.I. Park, Y. Yin, M. Byun, Y.S. Jung, K.J. Lee, Flexible one diode-one phase change memory array enabled by block copolymer self-assembly, ACS Nano 9 (2015) 4120–4128.

[26] H. Hirsch, K. Hellwig, S. Dobler, Speech recognition at multiple sampling rates, Proc. Eur. Conf. Speech Commun. Technol. 2001 (2001) 1837–1840.

[27] S. Egusa, Z. Wang, N. Chocat, Z.M. Ruff, A.M. Stolyarov, D. Shemuly, F. Sorin, P.T. Rakich, J.D. Joannopoulos, Y. Fink, Multimaterial piezoelectric fibres, Nat. Mater. 9 (2010) 643–648.

[28] Y. Lecun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.

[29] V. Mnih, K. Kavukcuoglu, D. Silver, A.A. Rusu, J. Veness, M.G. Bellemare, A. Graves, M. Riedmiller, A.K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, D. Hassabis, Human-level control through deep reinforcement learning, Nature 518 (2015) 529–533.

[30] D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Mastering the game of Go with deep neural networks and tree search, Nature 529 (2016) 484–489.

[31] G.-T. Hwang, Y. Kim, J.-H. Lee, S. Oh, C.K. Jeong, D.Y. Park, J. Ryu, H. Kwon, S.-G. Lee, B. Joung, D. Kim, K.J. Lee, Self-powered deep brain stimulation via a flexible PIMNT energy harvester, Energy Environ. Sci. 8 (2015) 2677–2684.

[32] G.-T. Hwang, V. Annapureddy, J.H. Han, D.J. Joe, C. Baek, D.Y. Park, D.H. Kim, J.H. Park, C.K. Jeong, K. Il Park, J.J. Choi, D.K. Kim, J. Ryu, K.J. Lee, Self-Powered Wireless Sensor Node Enabled by an Aerosol-Deposited PZT Flexible Energy Harvester, Adv. Energy Mater. 6 (2016) 1600237.

[33] D.H. Kim, H.J. Shin, H. Lee, C.K. Jeong, H. Park, G.T. Hwang, H.Y. Lee, D.J. Joe, J.H. Han, S.H. Lee, J. Kim, B. Joung, K.J. Lee, In Vivo Self-Powered Wireless Transmission Using Biocompatible Flexible Energy Harvesters, Adv. Funct. Mater. 27 (2017) 1700341.

[34] G.-T. Hwang, M. Byun, C.K. Jeong, K.J. Lee, Flexible Piezoelectric Thin-Film Energy Harvesters and Nanosensors for Biomedical Applications, Adv. Healthcare Mater. 4 (2015) 646-658.

[35] K. Song, K. Kim, S. Hur, J.H. Kwak, J. Park, J.R. Yoon, J. Kim, Sound pressure level gain in an acoustic metamaterial cavity, Sci. Rep. 4 (2014) 7421.

[36] H.F. Ma, T.J. Cui, Three-dimensional broadband ground-plane cloak made of metamaterials, Nat. Commun. 1 (2010) 21.

[37] C.H. Je, J. Lee, W.S. Yang, J. Kim, Y.H. Cho, A surface-micromachined capacitive microphone with improved sensitivity, J. Micromechanics Microengineering. 23 (2013) 055018.

[38] J. Allen, Short term spectral analysis, synthesis, and modification by discrete Fourier transform, IEEE Trans. Acoust. 25 (1977) 235–238.

[39] R. Boyer, J. Moore, MJRTY—A Fast Majority Vote Algorithm, Autom. Reason. (1991) 105–117.

[40] J. Cho, C.D. Yoo, Underdetermined convolutive BSS: Bayes risk minimization based on a mixture of super-Gaussian posterior approximation, IEEE/ACM Trans. Audio Speech Lang. Process 23 (2015) 828–839.

[41] S. Kim, C.D. Yoo, S. Nowozin, P. Kohli, Image segmentation usinghigher-order correlation clustering, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2014) 1761–1774.

[42] J.H. Choi, C.D. Yoo, Underdetermined high-resolution DOA estimation: A 2 ρth-order source-signal/noise subspace constrained optimization, IEEE Trans. Signal Process 63 (2015) 1858–1873.

[43] M. Jin, F.K. Soong, C.D. Yoo, A syllable lattice approach to speaker verification, IEEE Trans. Audio, Speech Lang. Process 15 (2007) 2476–2484.

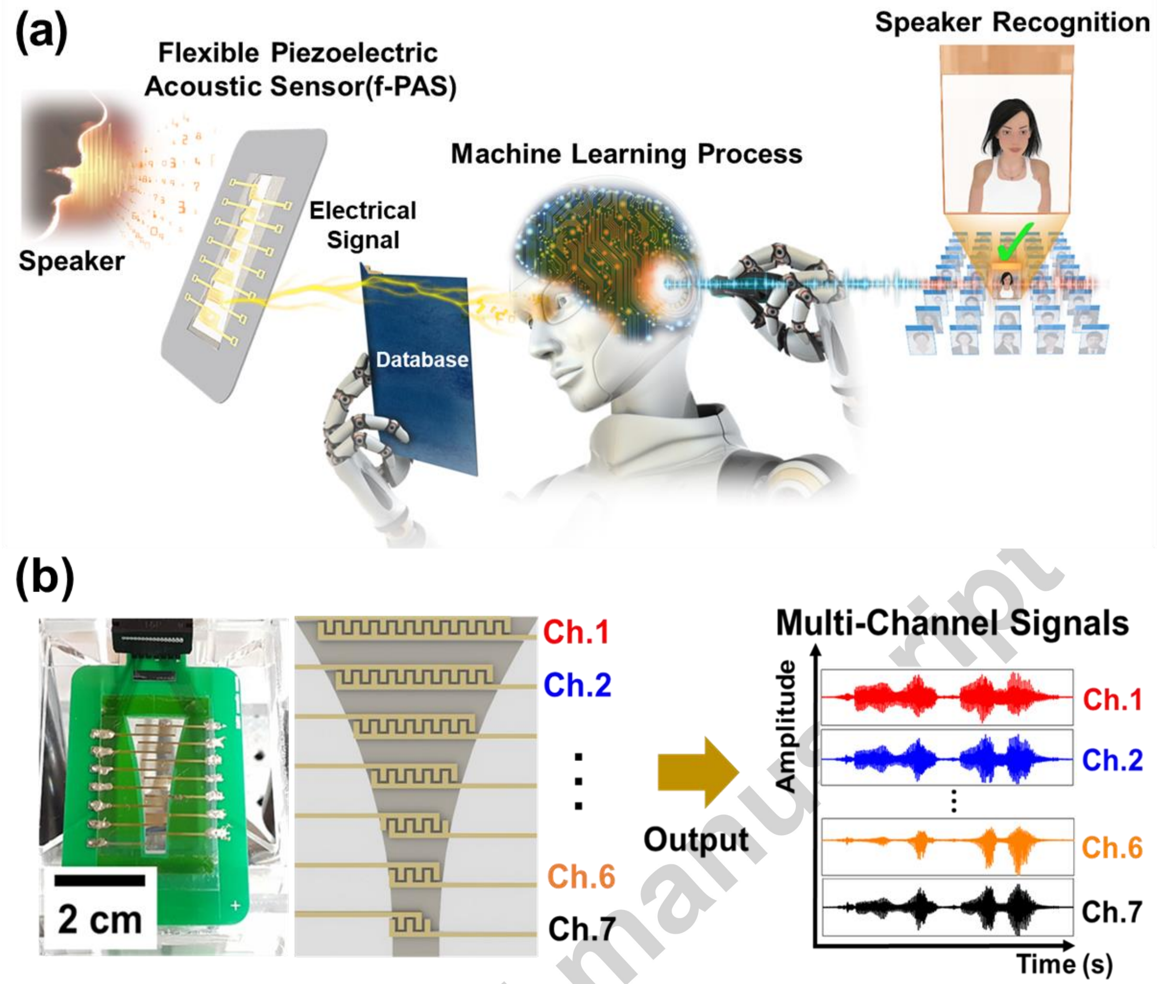

Figure 1. Overall schematic of the machine learning-based speaker recognition system and its multiple signals from f-PAS multi-channel. (a) Overall schematic of the machine learning-based speaker recognition system. The f-PAS with multi-channel provides multi signals while the piezoelectric membrane vibrates in response to the speaker's voice. The training and test procedure demonstrate the machine learning process using GMM algorithm. Speaker recognition is implemented by this process. (b) A photograph of f-PAS and a schematic of multi-channel flexible piezoelectric acoustic sensor with its output voice signal, which consists of multiple signals.

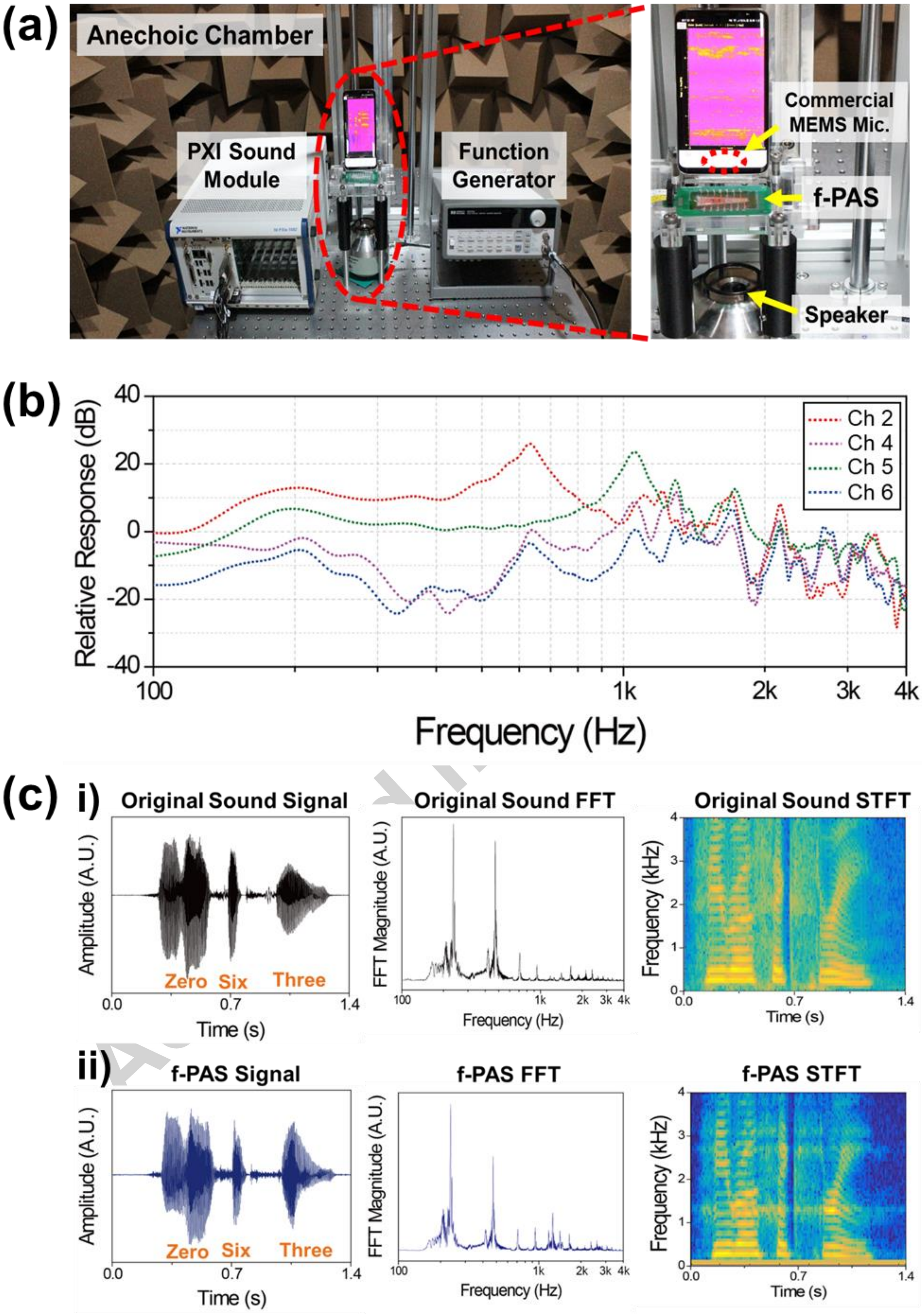
(a)
Anechoic Chamber
PXI Sound Module
Function Generator
Commercial MEMS Mic.
f-PAS
Speaker
(b)
Relative Response (dB)
40
20
0
-20
-40
100
1k
2k
3k
4k
Frequency (Hz)
Ch 2
Ch 4
Ch 5
Ch 6
(c)
i)
Original Sound Signal
Amplitude (A.U.)
Zero
Six
Three
0.0
0.7
1.4
Time (s)
Original Sound FFT
FFT Magnitude (A.U.)
100
1k
2k
3k
4k
Frequency (Hz)
Original Sound STFT
Frequency (kHz)
4
3
2
1
0
0.0
0.7
1.4
Time (s)
ii)
f-PAS Signal
Amplitude (A.U.)
Zero
Six
Three
0.0
0.7
1.4
Time (s)
f-PAS FFT
FFT Magnitude (A.U.)
100
1k
2k
3k
4k
Frequency (Hz)
f-PAS STFT
Frequency (kHz)
4
3
2
1
0
0.0
0.7
1.4
Time (s)

Figure 2. Experimental setup for measuring the signals from the f-PAS and its electrical output signals from the incoming sound wave. (a) Photograph of overall experimental setup for measuring the displacement and electrical output of the f-PAS in anechoic chamber. The inset shows the commercialized cellular phone located on the speaker and the f-PAS. (b) Relative response of f-PAS over the frequency sweep of 100 to 4000Hz. Resonance frequency of channels 2 and 5 appeared at 650Hz and 1080Hz, respectively. (c) Comparison of original and the f-PAS sound signals (Woman, voice of 063). i) original sound signal, FFT response and STFT of original sound signals. ii) the f-PAS signal, FFT response and STFT of the f-PAS signals. The STFT is multiple frames of the time varying FFT signal over short pieces of period to analyze the time-dependent characteristics of frequency domain signals.

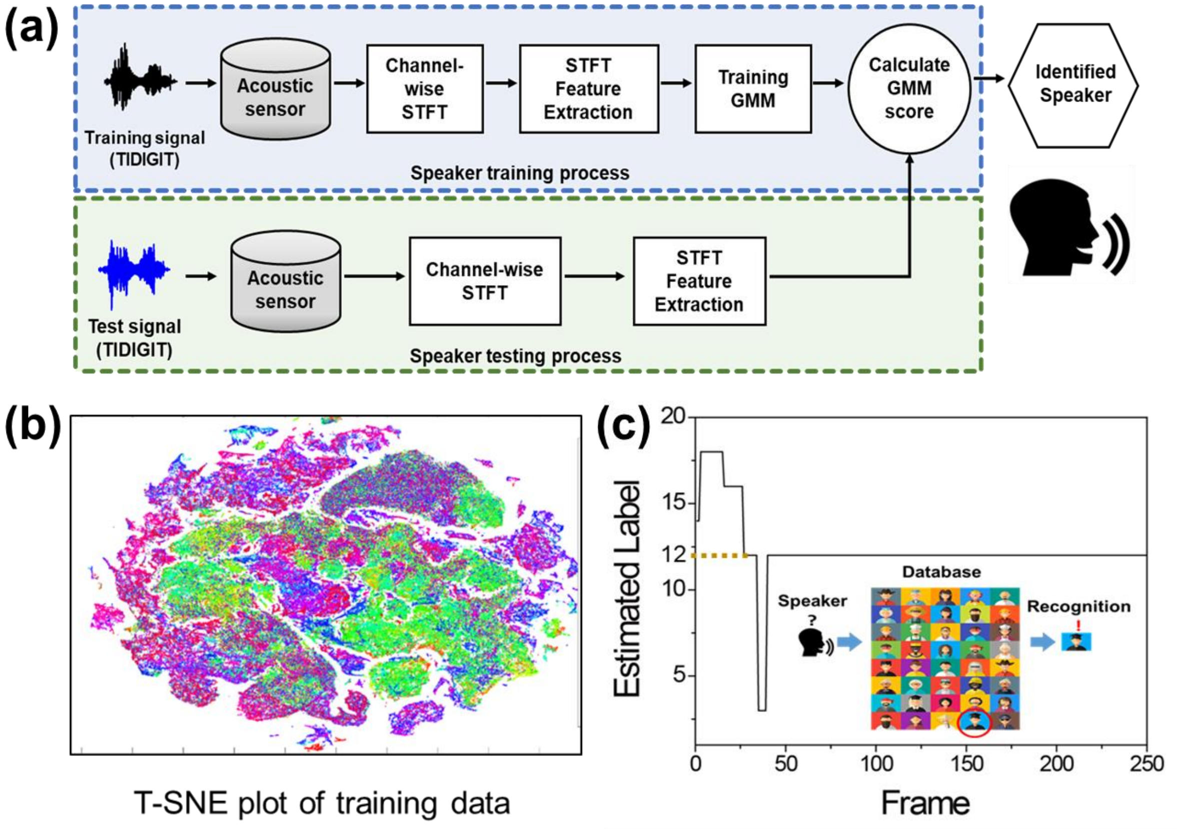

Figure 3. Visual representations of GMM-based machine learning algorithm for the speaker recognition. (a) GMM algorithm flow chart of speaker training process and test process using the standard TIDIGITS dataset (20 men and 20 female speakers, 77 speeches per each speaker, a total of 3080 voice data). The TIDIGITS dataset of 90 % percent used for training data, and 10 % for testing data. (b) The trained STFT features visualized in t-SNE plot by 2800 training data of 40 people. The t-SNE plot embeds high-dimensional data of similar objects into a low-dimensional space, which is related with probability distribution. (c) The majority voting method to test the speaker recognition process over the frames in case of finding 12th speaker.

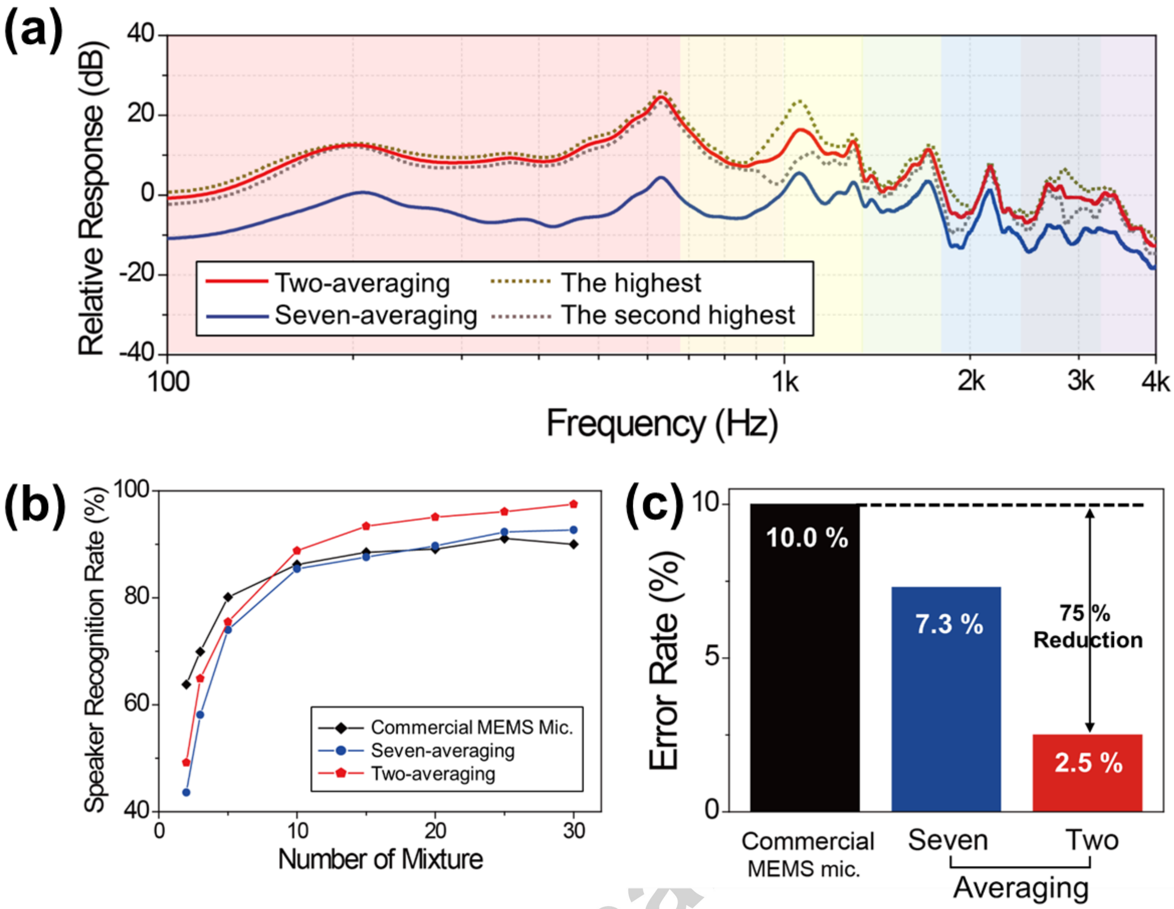

Figure 4. Frequency responses of f-PAS and comparison of speaker recognition rate between f-PAS and commercial MEMS microphone. (a) Relative frequency response of two-averaging (red), seven-averaging (blue), the most highest and the second highest signals over the voice frequency range of 100 Hz to 4 kHz. All signals were normalized to 0 dB at 1 kHz. (b) Comparison of the speaker recognition rate of commercial MEMS microphone, seven-averaging and two-averaging signals of the f-PAS according to the number of mixture. (c) Recognition error rate of commercialized phone and the f-PAS (seven-averaging and two-averaging) in the mixture number of 30.

Jae Hyun Han received his B.S. degree in Materials Science and Engineering (MSE) from Sungkyunkwan University in 2014 and M.S. degree from KAIST in 2016. He is currently working toward his Ph.D. at KAIST under the supervision of Prof. Keon Jae Lee. His doctoral research interests include piezoelectric and triboelectric energy harvesting, flexible acoustic sensors, and laser-material interaction.

Kang Min Bae received his B.S. degree in Electrical Engineering (EE) from Korea University in 2017. He is currently working toward his M.S. at KAIST under the supervision of Prof. Chang D. Yoo. His research interests include deep learning and computer vision.

Seong Kwang Hong received his Ph.D. degree in Electrical Engineering at Hanyang University in 2017. Currently, he works with Prof. K. J. Lee as a postdoctoral research associate in the Department of Materials Science and Engineering (MSE) at Korea Advanced Institute of Science and Technology (KAIST). His research interests include energy harvesting system, flexible electronics, and sensor applications.

Hyunsin Park received his B.S. degree and M.S. degrees in computer science from Kobe University, in 2007 and 2009, respectively. He is currently working toward his Ph.D. at KAIST under the supervision of Prof. Chang D. Yoo. His doctoral research interests include multimedia signal processing and machine learning.

Jun-Hyuk Kwak received his M.S. in Electrical Engineering from Kyungpook national university in 2008. He worked at Korea Institute of Machinery & Materials (KIMM) as senior researcher.

Currently, he is senior researcher in the Department of Research at Center for Advanced Meta-Materials (CAMM). His current research topics are MEMS acoustic sensor, application of acoustic metamaterial, underwater ultrasonic sensor with metamaterial and energy harvesting device by acoustic wave.

Hee Seung Wang received his B.S. and M.S. degree in Materials Science and Engineering (MSE) from Korea University in 2015 and from Korea Advanced Institute of Science and Technology (KAIST) in 2017, respectively. He is currently working toward his Ph.D. and his research topics are triboelectric energy harvesting and flexible piezoelectric acoustic sensor.

Daniel J. Joe received a B.S. degree in Electrical Engineering at the University of Illinois at Urbana Champaign (UIUC) in 2008 and a Ph.D. degree in Electrical and Computer Engineering at Cornell University in 2014. Currently, he is a BK21 Plus postdoctoral research associate in the Department of Materials Sciences and Engineering at KAIST. His current research interests involve various flexible thin-film materials and devices including energy harvesters, acoustic sensors, pressure sensors, wireless power transfer system, etc.

Jung Hwan Park received his Ph.D. in Materials Science and Engineering (MSE) at KAIST. During his Ph.D., he pioneered light-material interaction for flexible and stretchable electronics under the supervision of Prof. Keon Jae Lee. He is currently interested in extending his research field to the novel light-induced fabrication and synthesis for mass production of soft materials and devices.

Shin Hur is a principal researcher in Korea Institute of Machinery and Materials of Nanoconvergence Mechanical System Division. He received his Ph.D. degree from Chungnam National University, South Korea in 2005. His main research interests include biomimetic and piezoelectric acoustic devices based on MEMS fabrication, acoustic metamaterials.

Prof. Chang D. Yoo received the B.S. degree in Engineering and Applied Science from the California Institute of Technology, the M.S. degree in Electrical Engineering from Cornell University and the Ph.D degree in Electrical Engineering from the MIT. He is Director of Korea Institute of Electrical Engineers (KIEE) and Director of the Acoustical Society of Korea (ASK). He is Member of Tau Beta Pi and Sigma Xi. He was on the technical committee member of IEEE machine learning for signal processing society from 2009 to 2011. He had also served as Associated Editor of IEEE Signal Processing Letters, IEEE Transactions on Information Forensics and Security, IEEE Transactions on Audio, Speech and Language Processing.

Prof. Keon Jae Lee received his Ph.D. in Materials Science and Engineering (MSE) at University of Illinois, Urbana-Champaign (UIUC). During his Ph.D. at UIUC, he involved in the first co-invention of “Flexible Single-crystalline Inorganic Electronics”, using top-down semiconductors and soft lithographic transfer. Since 2009, he has been a professor in MSE at KAIST. His current research topics are self-powered flexible electronic systems including energy harvesting/storage devices, IoT sensor, LEDs, large scale integration (LSI), high density memory and laser material interaction for *in-vivo* biomedical and flexible application.

Highlights

A new platform of machine learning-based speaker recognition system was realized by the flexible piezoelectric acoustic sensor (f-PAS) with a highly sensitive multi-resonant frequency band.

The resonant self-powered f-PAS was fabricated by mimicking the operating mechanism of the basilar membrane in the human cochlear with the abundant voice information from the multi-channel sound inputs.

By utilizing the machine learning based Gaussian Mixture Model (GMM), the f-PAS exhibited outstanding speaker recognition rate of 97.5 % with error rate reduction of 75 % compared to that of the MEMS microphone.