Deep Learning-based Noise-Robust Flexible Piezoelectric Acoustic Sensors for Speech Processing

Young Hoon Jung
Korea Advanced Institute of Science and Technology
Trung Pham
Korea Advanced Institute of Science and Technology (KAIST)
Dias Issa
Korea Advanced Institute of Science and Technology (KAIST)
Hee Seung Wang
Korea Advanced Institute of Science and Technology
Jae Hee Lee
Korea Advanced Institute of Science and Technology (KAIST)
Mingi Chung
Korea Advanced Institute of Science and Technology (KAIST)
Bo-Yeon Lee
https://orcid.org/0000-0002-2150-1803
Gwangsu Kim
Korea Advanced Institute of Science and Technology (KAIST)
Chang D. Yoo
Korea Advanced Institute of Science and Technology
Keon Jae Lee (keonlee@kaist.ac.kr)
Korea Advanced Institute of Science and Technology (KAIST)

Article

Keywords: Flexible piezoelectric acoustic sensors, VUI, artificial intelligence, NPAS

Posted Date: August 13th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-799114/v1

DOI:
Article type:

Deep Learning-based Noise Robust Flexible Piezoelectric Acoustic Sensors for Speech Processing

Young Hoon Jung[1], Trung Xuan Pham[2], Dias Issa [2], Hee Seung Wang[1], Jae Hee Lee[1], Mingi Chung[1], Bo-Yeon Lee[3], Gwangsu Kim[2], Chang D. Yoo[2] and Keon Jae Lee[1*]*

[1]Y. H. Jung, [+] H. S. Wang, J. H. Lee, M. Chung, Prof. K. J. Lee
Department of Materials Science and Engineering, Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon
34141, Republic of Korea
E-mail: keonlee@kaist.ac.kr

[2]T. X. Pham, [+] D. Issa, Prof. G. Kim, Prof. C. D. Yoo
Department of Electrical Engineering, Korea Advanced Institute of Science and Technology (KAIST)
291 Daehak-ro, Yuseong-gu, Daejeon
34141, Republic of Korea
E-mail: cd_yoo@kaist.ac.kr

[3]Dr. B.-Y Lee
Department of Nature-Inspired System and Application, Korea Institute of Machinery and Materials (KIMM)
156 Gajeongbuk-Ro, Yuseong-gu, Daejeon
34103, Republic of Korea

[+] These authors contributed equally to this work.

Keywords: Flexible piezoelectric, Acoustic sensor, Deep learning algorithm, Noise-robust, Speaker Recognition, Speech Enhancement

Abstract

Flexible piezoelectric acoustic sensors (f-PAS) have attracted significant attention as a promising component for voice user interfaces (VUI) in the era of artificial intelligence of things (AIoT). The signal distortion issue of highly sensitive biomimetic f-PAS is one of the most challenging obstacle for real-life applications, due to the fundamental difference compared with the conventional microphones. Here, a noise-robust flexible piezoelectric acoustic sensor (NPAS) is demonstrated by designing the multi-resonant bands outside the noise dominant frequency range. Broad voice coverage up to 8 kHz is achieved by adopting an advanced piezoelectric membrane with the optimized polymer ratio. Deep learning-based speech processing of multi-channel NPAS is demonstrated to show the outstanding improvement in speaker recognition and speech enhancement compared to a commercial microphone. Finally, the NPAS independently identified the multi-user voices in a crowd condition, showing simultaneous speaker separation.

Introduction

Voice user interface (VUI), the most intuitive human-machine interaction (HMI), is a promising technology for personalized services, such as smart home appliances, intelligent virtual assistant, and biometric authentication in the Artificial Intelligence of Things (AIoT) era .[1–7] Commercialized microelectromechanical system (MEMS) microphones exhibit a flat response with low sensitivity in the range from 20 Hz to 8 kHz [8–10]. To enhance the signal-to-noise ratio (SNR) for far-distance detection, these capacitive microphones should be integrated with amplifying circuits, which results in the corresponding increase in noise and power consumption [11]. Furthermore, the single channel of MEMS microphones generates insufficient voice information, causing low accuracy in voice recognition. In contrast, the human ear solves the above issue by adopting the resonant vibration of basilar membrane and the multi-channel voice detection with 10,000 hair cells [12–14]. Recently, flexible piezoelectric acoustic sensors (f-PAS), mimicking the human cochlea, have attracted significant attention for sensing the voice spectrum by controlling resonant frequency bands via ultrathin trapezoidal membrane [15–19]. Biomimetic f-PAS with high sensitivity and multi-channel signals exhibited an exceptional speaker recognition rate in miniaturized dimensions [19].

The extremely sensitive response of lead-zirconate-titanate (PZT) based f-PAS can induce the significant interference between voice signals and ambient sounds [19,20]. A precise detection of voice features (0.1 – 8 kHz), regardless of surrounding noise and environmental conditions, is crucial to identify the speech signals from multi-users [21–23]. The previous f-PAS demonstrated the speaker recognition in anechoic conditions with a limited frequency coverage of up to 4 kHz [17,19]. For commercial application, the f-PAS should prove consistent and broad frequency response with wide voice coverage bands in noisy environments. The conventional MEMS microphones have overcome the distorted voice signals via noise reduction circuits, statistical/adaptive filters, and machine learning (ML) based noise filtering [21,24–26]. Recently, a new approach of noise-robust ML algorithms further improved the performance in speaker recognition and speech enhancement by treating voice input data like an image or calculating a weighted value for each signal [27,28]. However, these noise

filtering and ML technologies were designed to process signals from capacitive-type MEMS microphones with a single-channel input [29–31]. Therefore, the multi-channel f-PAS should be investigated based on totally different resonant mechanisms with new algorithms.

Herein, we report a noise-robust flexible piezoelectric acoustic sensor (NPAS) for highly accurate speech processing in real-life environments. The noise-robustness was realized via three different approaches: i) the frequency response of NPAS, ii) the image-like sound processing by convolutional neural network (CNN), iii) the newly designed deep U-net model for speech enhancement. The NPAS achieved noise-robust response by positioning the multi-resonant bands outside the noise dominant frequency range. To cover the entire human voice spectrum, the the frequency coverage of biomimetic NPAS was expanded up to 8 kHz by using Nb-doped PZT (PNZT) membrane. The highly sensitive NPAS proved the clear sound detection with less noise-interference, regardless of distance and sound pressure level (SPL), showing 35 times higher sensitivity and SNR than the conventional microphones. The speaker-separable characteristics of NPAS in noisy conditions were confirmed by visualizing the mel-frequency cepstral coefficient (MFCC) of multi-voice signals in a t-Distributed Stochastic Neighbor Embedding (t-SNE) plot. Deep learning algorithms were introduced to further improve the speech processing of NPAS in noisy conditions. Using the CNN algorithm, the multi-channel NPAS achieved a 96% speaker recognition rate with a 62% reduction in error rate compared to the commercialized microphone. Speech enhancement of the NPAS was also demonstrated by the selective processing of multi-channel signals via deep U-net model. Finally, the AI-based NPAS successfully separated multi-user voices from a crowd, indicating independent speaker's speeches can be identified and digitalized simultaneously.

Results

Biomimetic NPAS and deep learning-based speech processing

Fig. 1a schematically illustrates the overall concept of deep learning-based speech processing via highly sensitive biomimetic NPAS. (i) The flexible piezoelectric membrane with a noise-robust resonant response was fabricated by mimicking the mechanism of human cochlear. The basilar membrane of trapezoidal shape detects multi-frequency components depending on the width, which can allow the human audible range from 20 Hz to 20 kHz [12–14]. With a voice/noise frequency analysis, this biomimetic structure enables multi-tunable resonant bands for a noise-insensitive piezoelectric response by using the inversely proportional relationship between the resonance frequency and the width of NPAS [32],

$$f_r \propto \frac{t}{l^2}\sqrt{\frac{E}{\rho}} \quad (1)$$

where f_r is the resonance frequency, and t, l, E, and ρ indicate the thickness, width, elastic modulus, and density of NPAS membrane, respectively. Most of the voice information for speaker/speech recognition is distributed in frequency range of 0.1 – 8 kHz while noise sources of industry, office, home, and transportation environments are dominant below 0.1 kHz [33–37]. These distinctive frequency characteristics of voice and noise signals were utilized to achieve a less-distorted NPAS by locating resonance frequencies outside the noise range (only in the phonetic spectrum). To cover the entire voice spectrum up to 8 kHz with high sensitivity, the doping technique was used to improve the piezoelectric coefficient of PZT membrane based on the following equation [38–40],

$$d_{33} \sim \varepsilon_0 \chi_r P \quad (2)$$

where d_{33} is the piezoelectric coefficient, ε_0 is the vacuum permittivity, χ_r is relative permittivity, and P is the polarization. The substitutional donor dopant (Nb^{5+}) increased the dipole/domain mobility and the polarization of perovskite piezoelectric thin film annealed on a sapphire substrate. A flexible PNZT membrane with exceptional piezoelectric properties was fabricated by the detachment from a rigid substrate via inorganic-based laser lift-off (ILLO) method (see "Methods" and Supplementary Fig. 1 for fabrication details) [20,40–48]. The enhanced sensitivity enabled the broad

frequency coverage of resonant acoustic sensors over human voice spectrum, as shown in Supplementary Fig. 2a. (ii) The piezoelectric signals of multi-channel NPAS were generated by human voice and noisy sounds, providing the sufficient data for speech processing. Deep learning algorithms performed the training process of noise-mixed utterance dataset for the optimization of multi-channel NPAS signals. Finally, noise-robust speaker recognition and speech enhancement were demonstrated using the CNN and deep U-net algorithms, respectively.

Fig. 1b shows the flexible multi-channel NPAS membrane having the high flexibility and durability under bending deformation, which is crucial to maximize the piezoelectric conversion of minute sounds by resonant vibrations. As shown in the inset optical image, the trapezoidal NPAS film was interconnected to a printed circuit board (PCB) with a curvilinear sound hole for the bottom port microphone. The curved trapezoidal structure of NPAS membrane enabled the linear distribution of the multi-resonant frequencies from the apex region of channel 1 to the base position of channel 7. The multi-resonant vibrations of NPAS film are important to enhance the piezoelectric response over the entire voice spectrum. Fig. 1c presents the three-dimensional nanometer-scale displacements of multi-channel NPAS membrane under white noise at 94 dB SPL (the reference pressure of 1 Pa). By using a laser Doppler vibrometer (LDV), the laser light was irradiated on the entire area of vibrating NPAS film during a frequency sweep from 1 Hz to 8 kHz. The oscillation displacements of piezoelectric membrane were measured by calculating the difference in the frequency of incident and reflected light. The noise-robust piezoelectric response of NPAS was verified by negligible movements in the noise-dominant range. The displacements of a few nanometers were generated below 0.1 kHz, whereas the sound waves in the phonetic spectrum induced the intensive vibrations of ~ 180 nm. These results indicate that the multi-channel NPAS can generate the exact sound signals with less-interfered piezoelectric response and entire voice spectrum coverage.

Characterization of flexible PNZT thin film.

The crystal quality of piezoelectric film is important to improve the sensitivity and detection distance of flexible piezoelectric acoustic sensors [20,42,49]. However, perovskite materials should be annealed at high temperature for the crystallization, which is not compatible flexible plastic substrates [41]. Fig. 2a shows the X-ray diffraction (XRD) analysis data of the PNZT membrane on PET substrates measured by out-of-plane (θ-2θ) scan mode, indicating the polycrystalline perovskite structures after the ILLO process. The θ-2θ XRD peaks show that high-quality inorganic thin film with an identical orientation was achieved on a flexible substrate after the detachment from a rigid sapphire substrate of Supplementary Fig. 3. The XRD results as a function of annealing temperature were also analyzed to exhibit the {100}-oriented high crystallinity above 650℃ by comparing the full width at half maximum (FWHM) of the rocking curve, as depicted in Supplementary Fig. 4 [40,41]. The heat-treated 1 μm thick PNZT membrane was transferred onto a flexible substrate without the mechanical damage of cracks and fracture, as shown in the inset cross-sectional scanning electron microscopy (SEM) images. SEM images of the PNZT surfaces annealed at four different temperatures represent the large average grain size above 650℃, as displayed in Supplementary Fig. 5. This observation suggests that the superior PNZT membrane can be obtained based on the relationship between grain size and piezoelectric properties [41,50]. Supplementary Fig. 6 is the Raman spectra of the PNZT thin film for the characterization of perovskite phases, which reveals that the ILLO process enabled the transfer of piezoelectric membrane annealed at high temperature onto plastic substrates [20,40,41]. The lattice configuration of PNZT crystalline was also maintained as depicted in the high-resolution transmission electron microscopy (HRTEM) images of Supplementary Fig. 7. These structural characterizations confirm that the highly sensitive flexible PNZT membrane was developed because the piezoelectric properties are dependent on crystal orientation, and polymorphic phase [40,42]. Fig. 2b presents the surface analysis results of PNZT membrane after the ILLO process by using X-ray photoelectron spectroscopy (XPS). The XPS spectra obtained from the PNZT surface indicates that laser irradiation did not change the elemental binding energy of the piezoelectric film, compared with Supplementary

Fig. 8. A compositional analysis was also characterized by energy dispersive spectroscopy (EDS) elemental mapping, as shown in the inset of Fig. 2b (see Supplementary Fig. 9, 10 and Supplementary Table 1 for details). An insignificant chemical composition change (Nb: 0.08 at%) was observed on the laser-irradiated PNZT surface. These compositional characterizations indicate that the piezoelectric properties of high-temperature annealed inorganic membrane can be retained on a flexible PET substrate.

Ferroelectric properties are crucial to maximize the resonant sensor performance up to 8 kHz by enhancing the piezoelectric coefficient [38–40]. Fig. 2c depicts the relative permittivity and loss tangent measured to compare the ferroelectric characterizations between PZT and PNZT thin film over the human voice spectrum. The dielectric properties were characterized with the interdigitated electrodes (IDEs) deposited on the piezoelectric membrane, at an oscillation voltage of 5 mV, as shown in the inset optical image of Supplementary Fig. 12b. The flexible PNZT film exhibited higher relative permittivity and similar dielectric loss tangent δ (1000 and 0.04 at 1 kHz) compared to the PZT film (600 and 0.03 at 1 kHz). The improved dipole/domain mobility was attributed to the donor dopant reducing the oxygen vacancy with the defect dipoles of $Nb_{Zr}^{\bullet} - V_{Pb}^{''}$ and $Nb_{Zr}^{\bullet} - V_{Pb}^{''}$ [51–53]. As displayed in Supplementary Fig. 11, 12, the ferroelectric properties were also analyzed as functions of annealing temperature and Nb concentration. The maximum values of permittivity and polarization were shown in 4% Nb-doped PNZT membrane annealed at 650℃. These results suggest that the highly sensitive NPAS can be fabricated by using the optimized PNZT thin film based on Eq. (2). To investigate the PNZT membrane effect on acoustic sensors, the theoretical piezoelectric response at the first resonance was calculated via a finite element method (FEM) simulation. Fig. 2d shows the calculated relative sensitivity of NPAS using the FEM simulation with the following equation,

$$Sens.\ (dBV) = 20\ log\frac{V}{V_0} \tag{3}$$

where *Sens.* is the sensitivity, *dBV* is the units of decibels with respect to 1 volt, V is the root mean square voltage, and V_0 is the reference of 1 volt. The PNZT membrane presented 4 dBV higher sensitivity compared to a PZT film, which proves that the Nb dopant can enhance sensor performance

with superior piezoelectric properties. A broad resonant bandwidth (Δf ~ 400 Hz) of PNZT film was obtained with the 25 μm thick flexible substrates, indicating the low quality factor (Q factor, ~ 1.7) at the first resonance frequency of NPAS. The Q factor is inversely proportional to the bandwidth as in the following equation,

$$\mathrm{Q} = \frac{f_0}{\Delta f} \tag{4}$$

where f_0 is the resonance frequency, and Δf is the frequency bandwidth below 3 dB of the resonant peak value. The effect of polymer ratio on the bandwidth and sensitivity was also theoretically calculated to verify that the 25 μm thick PET could be used for broadening the detectable frequency range of NPAS, as depicted in Supplementary Fig. 13. These material analysis and simulation results indicate that the NPAS can cover the entire voice spectrum up to 8 kHz by using highly sensitive PNZT thin film on the optimized plastic substrates. Note that the NPAS exhibited higher figures of merit for sensitivity and frequency coverage compared to previous resonant piezoelectric acoustic sensors (see Supplementary Fig. 2b and Supplementary Table 2) [15,16,18,20,54–57].

Multi-resonant characterization of NPAS.

The frequency components of sound sources should be analyzed to design the multi-resonant bands. Fig.3a shows the Fast Fourier Transform (FFT) signals of original voice (i) and noise (ii), comparing the frequency domain characteristics. The voice FFT response indicates the major components of the human voice (male/female utterance of 1088153/9014) are above 0.1 kHz while the dominant range of noise signals (indoor/outdoor environmental sound) are below 0.1 kHz [33–37]. Note that the most discriminative information for speaker/speech recognition is located in the high frequency region (3.5 – 8 kHz) of voice spectrum [58,59]. Fig.3b displays the relative sensitivity of NPAS with multi-resonance bands over the human voice range. The relative sensitivity refers to the frequency response of NPAS compared to a reference capacitive microphone (G.R.A.S. 46BE). The NPAS covered the entire voice spectrum up to 8 kHz by combining the improved piezoelectric properties of inorganic membrane and resonance bands of multi-channel, showing higher sensitivity

than capacitive microphone in the high frequency region. The frequency response was measured in the free field condition of anechoic chamber with the white noise (a mixture of frequencies with equal intensity) at 94 dB SPL, which could obtain the electrical output signals without external noise and wave reflection [60]. The relative sensitivity of NPAS was plotted by acquiring the highest electrical signal among seven channels. The detailed frequency distribution of seven NPAS channels from 0.1 to 8 kHz are depicted in Supplementary Fig. 15. As displayed in Supplementary Fig. 16, the NPAS was optimized by comparing the frequency response depending on Nb concentration, that verified the correlation between sensitivity and piezoelectric properties (Supplementary Fig. 12). The highly sensitive response of acoustic sensors is required for clear sound recognition from a far-distance since the sound pressure decreases as a function of the distance [61]. Based on the calculation with Eq. (3), the maximum relative sensitivity of NPAS was 398 times (52 dBV) higher in units of voltage than the reference condenser microphone at the first resonance of 650 Hz. Supplementary Fig. 14 presents the frequency response of NPAS compared to the commercial G.R.A.S microphone over the noise-dominant range (1 ~ 100 Hz). As depicted in the inset of Fig. 3b, the noise-robust piezoelectric response was confirmed by comparing the maximum output voltage of NPAS membrane over the noise-dominant range and the phonetic spectrum. The negligible piezoelectric voltage was measured below 0.1 kHz, while the sound over voice frequency range generated the significant electrical output. The multi-resonant responses of NPAS were theoretically investigated to prove that the enhanced piezoelectric properties improved the voice spectrum coverage. Fig. 3c illustrates the FEM results for the NPAS membrane to analyze the mechanical displacements under the monochromatic sound waves of resonance modes. At the 4th mode of 1810 Hz, the maximum displacement of 180 nm was observed near the region of channel 4. As the resonance frequency was increased up to the 12nd mode, the maximum displacement region shifted towards the narrow position of channel 7 (see Fig. 3c and Supplementary Fig. 17). Note that the sensor performance decreases as the IDE channel width becomes narrow because of the linear relationship between sensitivity and active piezoelectric area. However, the middle region of highly sensitive NPAS membrane was resonated intensively with the

13rd harmonic mode sound, which enabled the broad resonance band in the range of 6.5 – 8 kHz, as displayed in Fig. 3b and Supplementary Fig. 18.

In flexible resonant acoustic sensors, highly efficient piezoelectric conversion is crucial to detect minute sound without the amplification circuit that causes increase in power consumption and noise fluctuation, as presented in Supplementary Fig. 19 [11,61]. Fig. 3d presents the electrical voltage outputs of multi-channel NPAS under monochromatic sound waves of 94 dB SPL. As shown in Supplementary Fig. 17, piezoelectric signals of the 1st, 2nd, and 3rd resonances were shown at channels 2, 3, and 4, respectively. The high frequency response of NPAS was compared with the commercialized G.R.A.S microphone by measuring the sensitivity at sound waves of 2, 3, 4, 5, 6, and 7 kHz, as depicted in the inset of Fig. 3d. The magnified sinusoidal voltages of NPAS and the reference microphone are displayed in Supplementary Fig. 20. The outstanding peak-to-peak voltage of NPAS (~ 130 mV at 1st resonance) was up to 35 times higher than the commercial condenser-type microphone (~ 3.7 mV). To detect far-distant voice clearly with less noise-interference, acoustic sensors require high sensitivity and SNR without amplification [11,61]. Fig. 3e shows the sensitivity expressed logarithmically in units of dBV under monochromatic sound, converted from the NPAS voltage signal of Fig. 3d by using Eq. (3). Self-powered NPAS exhibited the exceptional sensitivity of – 26 dBV at the first resonance mode, providing the solution for the amplifier-induced limitations of MEMS microphones. Fig. 3f displays the SNRs of NPAS under the monochromatic sound waves of the 1st – 3rd resonances (i), and 2 – 7 kHz (ii). The SNRs of self-powered NPAS were calculated by subtracting the sensitivity peaks and electrical noise base line. The highly sensitive NPAS exhibited the exceptional SNRs of 94, 83, and 82 dBV at each harmonic frequency (Fig. 3f, i), showing the less-interference property compared to 63 dB of the commercial capacitive microphone. As presented in Fig. 3f, ii, the noise-robust characteristics were also proved at non-resonant high frequencies due to the broad resonant bandwidth.

Fig. 3g shows the sensitivity and SNR of highly sensitive NPAS as a function of distance. The relationship between the sound pressure level and the distance is defined by the following equation,

$$\Delta L_p = 20 \log \frac{r_f}{r_i} \quad (5)$$

where ΔL_p is the difference in the sound pressure level, r_f is the final distance between the speaker and NPAS, and r_i is the initially measured distance from the sound source. As described in Supplementary Fig. 21a, the reference pressure of 94 dB SPL was obtained from the initial position (r_i), where the sensitivity and SNR of NPAS were – 26 dBV and 94 dBV, respectively. The sensitivity and SNR of NPAS were inversely proportional to the distance because the SPL decreased depending on the measurement distance [49]. In addition, Supplementary Fig. 21b displays the sensitivity of NPAS as functions of distance and incident angle for the directional characterization. The exceptional sensitivity and SNR of NPAS were maintained at different distance and angle, indicating the capability of far-distant voice recognition without the directional distortion. The inset of Fig. 3g shows the linear relationship between the piezoelectric voltage and pressure at the first resonance. The output voltage of highly sensitive NPAS increased from 0.1 mV to 190 mV by increasing the pressure up to 6.3 Pa. The linearity of NPAS suggests the feasibility of voice recognition in wide sound pressure range for practical applications.

Deep learning-based noise robust speaker recognition.

The voice features generated by acoustic sensors should be similar to the original sound sources for accurate speaker recognition [21–23]. Fig. 4a presents the waveform of original sound signal and NPAS signal by using TIDIGIT speech (man, voice of 1088153) and real-world noise (knock sound of Supplementary Fig. 22, 23). As displayed in Fig. 4b, the time-domain voltage signals were converted into the frequency-domain MFCC using a Discrete Cosine Transform (DCT) for the feature extraction [21]. The multi-channel NPAS represented the time-varying frequency characteristics similar to original speech data, generating the abundant voice information with the entire phonetic spectrum coverage for the speaker recognition of neural network model. Fig. 4c shows the seven-channel NPAS flowchart of CNN-based deep learning process for the speaker recognition with multi-resonant signals. The forty speakers were randomly selected from the standard TIDIGITS dataset for recording the

voice information (20 men and 20 women speakers, 77 speeches per each speaker, a total of 3080 voice data). The 2800 speech data were used to train the CNN classifier while 280 voice data were utilized to evaluate the performance of speaker recognition. The MFCC features were extracted after the Short-Time Fourier Transform (STFT) process was performed in each frame of the noisy speech signal [21]. The CNN algorithm was trained with the MFCC features by minimizing the objective function of cross-entropy loss,

$$l = -\sum_{i=1}^{C} y_i \log\left(f_i(x_i)\right) \quad (6)$$

where l is the cross-entropy loss function, C is the number of speakers, y_i is the label, f_i is the probability predicted by the CNN model, and x_i is the MFCC input. As illustrated in Supplementary Fig. 24, the speaker recognition was conducted using an attention method that automatically applies the weighted values in the crucial channels of NPAS under noise levels from – 10 dB to 20 dB.

To verify the noise-robust voice detection, the NPAS speech signals were compared with MEMS microphone. Fig. 4d displays the MFCC features visualized in the scatter t-SNE plots by using the test datasets of the commercial microphone (i), and NPAS (ii). The t-SNE plots embedded the high-dimensional voice characteristics of 40 speakers into a two-dimensional (2D) space, which showed the probability distributions of similar speech clusters even under the extremely noisy conditions of – 10 dB SNR. The noise level, SNR, is defined by the following equation,

$$\mathrm{SNR} = 20\log\frac{V_{signal}}{V_{noise}}$$

where V_{signal} and V_{noise} are the voltage of signal and noise, respectively. Compared to the complexly mixed distribution of MEMS features, the NPAS features were clearly separated into different speaker clusters, indicating the higher recognition performance with noise-robustness. Fig. 4e shows the superior noise-robust characteristics of NPAS by comparing the decrease in speaker recognition rate as a function of noise levels. The NPAS exhibited only 8% reduction in recognition rate under the high noisy level of – 10 dB with 10 noises, whereas the accuracy of commercialized microphone decreased from 91% to 68%. It is noteworthy that the difference in accuracy rate was higher as the noise level increased. In addition, the recognition rate of NPAS outperformed the MEMS microphone

for both clean speech signal and single noisy signal (Supplementary Fig. 25, 26). Fig. 4f presents the speaker recognition error rate of the MEMS microphone and NPAS depending on the number of noises. The NPAS achieved a 61% reduction of error rate under the harsh conditions of – 10 dB with 40 noises, compared to commercial MEMS microphone. The outstanding speaker recognition was successfully demonstrated using noise-robust and highly sensitive frequency response of multi-channel NPAS with the channel attention-based CNN architecture.

Speech enhancement via deep learning model.

Fig. 5a shows a flow chart of Deep U-net based Speech Enhancement (DEEP-SEA) model to extract the de-noised speech signal from a time-domain noisy waveform. The DEEP-SEA model was newly designed to improve the performance by using the attention block and Gated Recurrent Unit (GRU) between the encoder and decoder. The TIDIGITS dataset was utilized to investigate speech enhancement under 3 different SNR levels from – 5 to 5 dB. The encoder layers of DEEP-SEA model transformed the TIDIGITS signals, providing the voice features to the decoder through the skip connections and the attention module (see "Methods" for detailed process). To enhance the speech quality of output data from the decoder, the DEEP-SEA model was trained by minimizing the following loss equation,

$$L_{SE}(w, \hat{w}) = \|w - \hat{w}\|_1 + \frac{1}{N}\sum_{n=1}^{N} L_{STFT}^{n}(w, \hat{w}) \quad (7)$$

where L_{SE} is the speech enhancement loss function, w is the ground clean waveform, $\hat{w}$ is the de-noised waveform, $\|\cdot\|_1$ is the Manhattan distance, N is the number of STFT losses, and L_{STFT}^{n} is the n-th resolution of multi-resolution STFT loss. End-to-end speech enhancement was performed with a single waveform by averaging the multi-channel signals or selecting one channel data among the seven NPAS signals. Supplementary Fig. 27 presents the waveforms of MEMS and NPAS signals (man, voice of 11292OO, and 5 noises mixture) comparing the time-domain data before and after speech enhancement. The clean voice waveform was obtained with the DEEP-SEA model by removing other sound signals, as shown in Supplementary Movie 1. As displayed in Fig. 5b, the STFT

algorithm was utilized to represent the speech information as a function of frequency. The speech-enhanced STFT spectrograms of MEMS microphone and NPAS showed the similar frequency components contained in voice signals. The NPAS exhibited more distinct voice features than the commercial microphone in the STFT spectrogram even without the noise filtering process, indicating the noise-robust characteristics of voice-resonant NPAS.

The objective measures were evaluated to compare the de-noised speech quality of MEMS and NPAS signals as a function of noise level. A high score in the objective evaluation means that the enhanced speech signals are high-quality sound without noise [62]. As presented in Fig. 5c, the standard metrics based on human auditory perception are the perceptual evaluation of speech quality (PESQ), and short-time objective intelligibility (STOI). A high PESQ indicates that the de-noised output signals are similar to the original clean data while the STOI measures the comprehensibility of speech data by comparing the time-frequency components. The NPAS achieved a higher score in both PESQ and STOI than the commercialized microphone (32% and 8% improvement, respectively), that proved the clear and unmodulated speech signal after the enhancement processing. Fig. 5d displays the evaluated composite measure for signal distortion (CSIG) of the commercial microphone and NPAS. The CSIG score of MEMS microphone was up to 3.9 at the low noise level of 5 dB, whereas the NPAS exhibited the less-distorted sound signal with the score of 4.1 in the extremely noisy condition of – 5 dB. Fig. 5e shows the comparison of the speech signals between the commercial microphone and the NPAS by calculating a composite measure for background noise intrusiveness (CBAK). An outstanding improvement of 109% was achieved with the two-averaging NPAS signals, showing superior voice enhancement of multi-resonant NPAS compared to the MEMS microphone. The exceptional speech quality was attributed to the selective data processing of multi-channel NPAS signals, as displayed in Supplementary Fig. 28, 29. The NPAS data for single channel 3 and two-averaging showed the high score in CSIG and CBAK, respectively. Note that the single channel 3 has the notable high frequency characteristics over the range of 4 ~ 5 and 7 ~ 8 kHz. Fig. 5f presents the enhancement in composite measure for overall speech quality (COVL) of NPAS via the intentional

channel selection. The COVL score of MEMS microphone, seven-averaging, two-averaging, and single channel 3 were rated as 2.6, 3.0, 3.5, and 3.6 at the high noisy level of – 5 dB SNR, respectively. The NPAS channel 3 exhibited a 40 % increase in COVL score compared to the commercialized microphone. The exceptional quality of speech-enhanced NPAS signals was enabled by the noise-robust STFT features and multi-channel voice data.

Fig. 5g schematically illustrates the speaker separation using the multi-channel NPAS. In the experiment, the speakers and crowd were located at 2 m and 3 m distant from the NPAS, respectively. The voices of different speakers can be regarded as noisy data due to the interference among speech signals. The multi-user signals were detected based on the frequency response and directional characteristics of NPAS (Supplementary Fig. 13, 18). The independent vector analysis (IVA) algorithm was utilized to separate the voice waveform of each speaker by real-time processing (Supplementary Movie 2). Note that the IVA algorithm requires a microphone array to separate the multi-speaker with high accuracy [63]. Fig. 5h shows the comparison of signal-to-distortion ratio (SDR), and signal-to-interference ratio (SIR) for each speaker depending on the number of NPAS channels. The high SDR and SIR are important metrics of speaker separation, representing clear speech data without the voice of each other user. As the number of channels was increased from 1 to 7, the SDR and SIR of speaker A were improved up to 6 and 9.3, respectively. The separated speech signals of speaker B exhibited more distorted but less interfered voice information compared to speaker A, that was analyzed by 1.8 dB lower SDR, and 0.5 dB higher SIR, respectively. These results suggests that the multi-channel NPAS could be used as an acoustic sensor array for separating multi-speakers in a crowd. Fig. 5i displays the separation performance of seven-channel NPAS by measuring the SDR and SIR of speaker A as a function of iterations. The SDR and SIR values of separated speech signals were enhanced by ~ 1.6 times (3.6 dB and 4.3 dB, respectively) when the number of iterations was 10. This efficient data processing was attributed to the multi-channel speech signals of the single NPAS chip, indicating the potential of NPAS in real-time IoT applications of multi-speaker separation and recording.

Discussion

In summary, we developed a noise-robust and broad spectrum-covering NPAS for deep learning-based speech processing by mimicking the multi-resonant mechanism of human cochlear. The multi-channel NPAS with voice-resonant bands achieved the insensitive response to noise components over the entire voice spectrum up to 8 kHz. The broad frequency coverage of NPAS was enabled by using the optimized PNZT membrane of superior piezoelectric properties. The biomimetic NPAS membrane detected the far-distant and minute voice signals without the distortion and interference while showing outstanding sensitivity and SNR (- 26 and 94 dBV) at the first resonance mode of 650 Hz. The noise-robust and speaker-separable characteristics of NPAS were visualized in a t-SNE plot using the MFCC features of multi-voice signals. The NPAS with multi-channel attention CNN achieved a 61% reduction in speaker recognition error rate compared to the commercialized microphone in the condition of 40 noises mixture and – 5 dB noise level. The DEEP-SEA model was developed to enhance the noise filtering performance by adding the attention block for the optimization of NPAS signals. The selective channel processing of DEEP-SEA model improved the objective quality metric of NPAS speech signals up to 109% compared to the conventional MEMS microphone. Finally, the multi-speaker separation was successfully processed by using NPAS and IVA algorithm.

Methods

Fabrication of the NPAS

A PNZT chemical solution (QUINTESS Co. Ltd., 0.4 M) was spin-coated on rigid sapphire substrates (Hi-Solar Co.), followed by a rapid thermal annealing (RTA) procedure for crystallization. The deposition process was repetitively conducted to form a 1 μm thick PNZT film. Subsequently, the surface of PNZT membrane was treated with O_2 plasma using inductively coupled plasma-reactive ion etching (ICP-RIE, SNTEK Co.). The ultraviolet (UV) sensitive polyurethane (PU, Norland Optical Adhesive) was spin-coated to attach a 25 μm thick PET to the crystallized PNZT thin film. The PNZT membrane was transferred onto flexible substrates by irradiating XeCl laser (wavelength of 308 nm) on the transparent mother substrates. After the ILLO process, seven IDEs channels (Cr/Au, thickness of 10 and 100 nm) were patterned on the surface of detached PNZT thin film using conventional microfabrication. The multi-channel NPAS was covered with a PU passivation layer to prevent mechanical and electrical damage. Finally, a poling process was conducted to align the piezoelectric dipoles after interconnection of NPAS and PCB.

Material Characterizations

The crystallographic properties of the PNZT thin film were characterized by a multipurpose thin-film X-ray diffractometer (D/MAX-2500, RIGAKU), a high resolution Raman/PL system (LabRAM HR Evolution Visible/NIR, HORIBA), and a field emission transmission electron microscope (Talos F200X, FEI). The compositional analyses of PNZT on both sapphire and PET plastic substrates were conducted using a multi-purpose X-ray photoelectron spectroscope (Sigma Probe, Thermo VG Scientific) and an energy dispersive X-ray spectroscope (SU5000, Hitachi). The morphological images were investigated with a focused ion beam scanning electron microscope (Helios G4, FEI), a field emission scanning electron microscope (SU5000, Hitachi), and an optical microscope (VHX-1000E, Keyence). The polarization-electric field hysteresis was analyzed using a ferroelectric measurement system (Precision Premier II, Radiant Technologies).

Mechanical and Electrical Signal Measurement

The mechanical displacements of the NPAS were characterized using an LDV (He-Ne laser, wavelength of 633 nm) with the frequency sweep of a mouth simulator (type 4227-A, Bruel & Kjaer). Electrical signals were measured via a National Instruments (NI) Sound Module under white noise and monochromatic sinusoidal sound waves generated with a function generator and mouth simulator. The NPAS characteristics were compared with a commercialized reference microphone (G.R.A.S. 46BE, condenser type) under the same conditions of 94 dB SPL. The TIDIGITS dataset was recorded by NPAS and a commercial phone (Samsung, Galaxy S8) under the same conditions to compare the speaker recognition and speech enhancement results.

Resonance Simulation

FEM simulation (COMSOL Multiphysics 5.2 software) of the NPAS was conducted to theoretically calculate the sensitivity, spectrum bandwidth, resonant frequencies, and vibrational displacements. The curvilinear membrane shape was constructed as an actual NPAS structure (5 mm of w1, 20 mm of w2, and 30 mm of l). The resonant frequency (Eq. (1)) was simulated to investigate the resonance distribution, and oscillation displacement in the NPAS membrane. The sensitivity of PZT and PNZT thin film for the frequency response of NPAS was compared by using the Eq. (3). The resonant bandwidth was also calculated as a function of the polymer to piezoelectric film ratio.

Speaker Recognition

A deep learning-based network (CNN) was utilized to classify 40 speakers for both data collected using a commercial cellular phone (Samsung, Galaxy S8) and the multi-channel NPAS. The noisy voice dataset was prepared by mixing 10 ~ 40 types of noises (indoor and outdoor sound sources) with clean TIDIGTIS speeches. After the waveforms were sliced into multiple frames through a pre-emphasis filter, a window function was applied to each frame. The filter banks were computed by using the STFT-converted frames, which enabled extraction of the MFCC features with a DCT

method. The CNN classifier was trained for 3000 epochs until the convergence with MFCC features extracted from Phone S8 and NPAS. Error rates were compared by calculating the simple equation of (100 – recognition rate (%)).

Speech Enhancement

De-noised voice signals were extracted via end-to-end speech enhancement using a single noisy waveform with a sampling rate of 16 kHz. To produce an input waveform, the single voice signal was generated by averaging or selecting the NPAS signal from seven IDEs. The model consists of two main components: i) encoder and decoder networks composed of 1D convolutional layers with standard U-net skip-connections; ii) an attention layer followed by Gated Recurrent Unit (GRU) blocks between the encoder and decoder. Both the encoder and decoder networks are composed of L layers, a kernel size of K, and stride equal to S (L=5, K=8, and S=4 in this model). The i-th encoder layer includes two 1D convolutional layers with the number of channels as in the equations below,

$$C_{in,i} = 48 \times 2^{i-2}, C_{out,i} = 2C_{in,i} = 48 \times 2^{i-1} \quad (7)$$

where C_{in} is the number of input channels, and C_{out} is the number of output channels ($C_{\mathrm{in},1} = 1$, and $1 \leq i \leq \mathrm{L}$). The Exponential Linear Unit (ELU) activated the outputs of the first convolutional layer, which was then passed to the second convolutional layer and Gated Linear Unit (GLU) activation. The outputs of each encoder layer were passed to the subsequent layer and the corresponding decoder layer via skip connection, providing the output of the encoder (denoted as X) to the attention layer. After the output of the attention layer was passed over the two GRU layers with a hidden size of $48 \times 2^{L-1}$, the latent representation was produced as $\mathrm{Z} = \mathrm{X} + \mathrm{GRU}(\mathrm{Attention}(\mathrm{X}))$. This representation was further passed to the decoder network having transpose convolutional layers constructed in the same manner as the encoder layers. In contrast with the encoder, the layers in the decoder network were numbered in a reverse direction from L to 1. The skip connections connect the i-th decoder input with the output of the i-th encoder. The DEEP-SEA model was trained utilizing a multi-resolution STFT loss,

$$L_{STFT}^{n}(w,\hat{w}) = L_{SC}(w,\hat{w}) + L_{Mag}(w,\hat{w})$$

$$= \frac{\|STFT(w) - STFT(\hat{w})\|_F}{\|STFT(w)\|_F} + \|log|STFT(w)| - log|STFT(\hat{w})|\|_1 \quad (8)$$

where L_{SC} is the spectral convergence loss, L_{Mag} is the magnitude loss, $\|\cdot\|_F$ is the Frobenius norm, and $|STFT(\cdot)|$ is the STFT magnitudes of waveform. The ranges of STOI, PESQ, and CSIG/CBAK/COVL were 0 ~ 100, - 0.5 ~ 4.5, and 1 ~ 5, respectively (a high score indicates the high quality of de-noised sound wave).

Speaker Separation

An independent vector analysis (IVA) was used to investigate the separation of the multi-users voices with the NPAS. The test was conducted by setting parameters such as the input sound sources (~ 14), the number of iterations (~ 30), the distance between NPAS and sound sources (2 ~ 3 m), and the number and sensitivity of the NPAS channels. The frequency bins of each voice signal were regarded as a single vector, representing the mixture of multi-speaker voices as the multiplication of mixing matrix and each voice matrix. The algorithm was processed to find the de-mixing matrix for obtaining the original voice of each speaker. The speaker was separated with a small amount of computations by minimizing the mutually common information among NPAS channels. The performance was evaluated by calculating the SDR and SIR with the original and separated voice sources.

References

1. Formisano, E., De Martino, F., Bonte, M. & Goebel, R. 'Who' is saying 'what'? Brain-based decoding of human voice and speech. *Science* **322**, 970–973 (2008).
2. Perrachione, T. K., Del Tufo, S. N. & Gabrieli, J. D. E. Human voice recognition depends on language ability. *Science* **333**, 595 (2011).
3. Li, W. *et al.* Nanogenerator-based dual-functional and self-powered thin patch loudspeaker or microphone for flexible electronics. *Nat. Commun.* **8**, 15310 (2017).
4. Ward, D. J. & MacKay, D. J. C. Fast hands-free writing by gaze direction. *Nature* **418**, 838–838 (2002).
5. Zhu, M. *et al.* Haptic-feedback smart glove as a creative human-machine interface (HMI) for virtual/augmented reality applications. *Sci. Adv.* **6**, 1–15 (2020).
6. Liu, Y. *et al.* Epidermal mechano-acoustic sensing electronics for cardiovascular diagnostics and human-machine interfaces. *Sci. Adv.* **2**, e1601185 (2016).
7. Jin, T. *et al.* Triboelectric nanogenerator sensors for soft robotics aiming at digital twin applications. *Nat. Commun.* **11**, 5381 (2020).
8. Walser, S. *et al.* MEMS microphones with narrow sensitivity distribution. *Sensors Actuators, A Phys.* **247**, 663–670 (2016).
9. Je, C. H., Lee, J., Yang, W. S., Kim, J. & Cho, Y. H. A surface-micromachined capacitive microphone with improved sensitivity. *J. Micromechanics Microengineering* **23**, 055018 (2013).
10. Ali, W. R. & Prasad, M. Piezoelectric MEMS based acoustic sensors: A review. *Sensors Actuators, A Phys.* **301**, 111756 (2020).
11. Kwon, H. sang & Lee, K. C. Double-chip condenser microphone for rigid backplate using DRIE and wafer bonding technology. *Sensors Actuators, A Phys.* **138**, 81–86 (2007).
12. Eatock, R. Adaptation in Hair Cells. *Annu. Rev. Neurosci.* **23**, 285–314 (2000).

13. Gillespie, P. G. & Müller, U. Mechanotransduction by Hair Cells: Models, Molecules, and Mechanisms. *Cell* **139**, 33–44 (2009).

14. Caprara, G. A., Mecca, A. A. & Peng, A. W. Decades-old model of slow adaptation in sensory hair cells is not supported in mammals. *Sci. Adv.* **6**, 1–13 (2020).

15. Lee, H. S. *et al.* Flexible inorganic piezoelectric acoustic nanosensors for biomimetic artificial hair cells. *Adv. Funct. Mater.* **24**, 6914–6921 (2014).

16. Han, J. H. *et al.* Basilar membrane-inspired self-powered acoustic sensor enabled by highly sensitive multi tunable frequency band. *Nano Energy* **53**, 198–205 (2018).

17. Han, J. H. *et al.* Machine learning-based self-powered acoustic sensor for speaker recognition. *Nano Energy* **53**, 658–665 (2018).

18. Jung, Y. H. *et al.* Flexible Piezoelectric Acoustic Sensors and Machine Learning for Speech Processing. *Adv. Mater.* **32**, 1–18 (2020).

19. Wang, H. S. *et al.* Biomimetic and flexible piezoelectric mobile acoustic sensors with multiresonant ultrathin structures for machine learning biometrics. *Sci. Adv.* **7**, eabe5683 (2021).

20. Park, K. Il *et al.* Highly-efficient, flexible piezoelectric PZT thin film nanogenerator on plastic substrates. *Adv. Mater.* **26**, 2514–2520 (2014).

21. Furui. S. Digital Speech Processing Synthesis, and Recognition. (Marcel Dekker, New York, 2018)

22. Benesty. J. *et al.* Noise Reduction in Speech Processing. (Springer, New York, 2009)

23. Gong, Y. Speech recognition in noisy environments: A survey. *Speech Commun.* **16**, 261–291 (1995).

24. Williams, M. D., Griffin, B. A., Reagan, T. N., Underbrink, J. R. & Sheplak, M. An AlN MEMS piezoelectric microphone for aeroacoustic applications. *J. Microelectromechanical Syst.* **21**, 270–283 (2012).

25. Vihari, S., Murthy, A. S., Soni, P. & Naik, D. C. Comparison of Speech Enhancement Algorithms. *Procedia Comput. Sci.* **89**, 666–676 (2016).

26. Pascual, S., Bonafonte, A. & Serra, J. SEGAN: Speech enhancement generative adversarial network. *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH* **2017-August**, 3642–3646 (2017).

27. Xu, Z. & Sun, J. Model-driven deep-learning. *Natl. Sci. Rev.* **5**, 22–24 (2018).

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

29. Gerkmann, T., Krawczyk-Becker, M. & Le Roux, J. Phase processing for single-channel speech enhancement: History and recent advances. *IEEE Signal Process. Mag.* **32**, 55–66 (2015).

30. Theodoridis. S. Machine Learning: A Bayesian and Optimization Perspective (Academic Press, London, 2015).

31. Schmidt, M. N. & Olsson, R. K. Single-channel speech separation using sparse non-negative matrix factorization. *INTERSPEECH 2006 9th Int. Conf. Spok. Lang. Process. INTERSPEECH 2006 - ICSLP* **5**, 2614–2617 (2006).

32. Sillero, E. *et al.* Static and dynamic determination of the mechanical properties of nanocrystalline diamond micromachined structures. *J. Micromechanics Microengineering* **19**, 115016 (2009).

33. Farina, A. Soundscape Ecology. (Springer, London, 2014).

34. Broner, N. The effects of low frequency noise on people-A review. *J. Sound Vib.* **58**, 483–500 (1978).

35. Broner, N. Low frequency and infrasonic noise in transportation. *Appl. Acoust.* **11**, 129–146 (1978).

36. Berglund, B., Hassmén, P. & Job, R. F. S. Sources and effects of low-frequency noise. *J. Acoust. Soc. Am.* **99**, 2985–3002 (1996).

37. Bengtsson, J., Persson Waye, K. & Kjellberg, A. Evaluations of effects due to low-frequency noise in a low demanding work situation. *J. Sound Vib.* **278**, 83–99 (2004).

38. Li, F. *et al.* Ultrahigh piezoelectricity in ferroelectric ceramics by design. *Nat. Mater.* **17**, 349–354 (2018).

39. Li, F. *et al.* Giant piezoelectricity of Sm-doped Pb(Mg1/3Nb2/3)O3-PbTiO3 single crystals. *Science.* **364**, 264-268 (2019).

40. Jeong, C. K. *et al.* Flexible highly-effective energy harvester via crystallographic and computational control of nanointerfacial morphotropic piezoelectric thin film. *Nano Res.* **10**, 437–455 (2017).

41. Hwang, G.-T. *et al.* Self-Powered Wireless Sensor Node Enabled by an Aerosol-Deposited PZT Flexible Energy Harvester. *Adv. Energy Mater.* **6**, 1600237 (2016).

42. Park, D. Y. *et al.* Self-Powered Real-Time Arterial Pulse Monitoring Using Ultrathin Epidermal Piezoelectric Sensors. *Adv. Mater.* **29**, 1702308 (2017).

43. Laser–Material Interactions for Flexible Applications. *Adv. Mater.* **29**, 1606586 (2017).

44. Jeon, T. *et al.* Laser Crystallization of Organic-Inorganic Hybrid Perovskite Solar Cells. *ACS Nano* **10**, 7907–7914 (2016).

45. Jin, H. M. *et al.* Flash Light Millisecond Self-Assembly of High χ Block Copolymers for Wafer-Scale Sub-10 nm Nanopatterning. *Adv. Mater.* **29**, 1700595 (2017).

46. Lee, H. E. *et al.* Monolithic Flexible Vertical GaN Light-Emitting Diodes for a Transparent Wireless Brain Optical Stimulator. *Adv. Mater.* **30**, 1800649 (2018).

47. Lee, H. E. *et al.* Micro Light-Emitting Diodes for Display and Flexible Biomedical Applications. *Adv. Funct. Mater.* **29**, 1808075 (2019).

48. Peng, Y. *et al.* Achieving high-resolution pressure mapping via flexible GaN/ ZnO nanowire LEDs array by piezo-phototronic effect. *Nano Energy* **58**, 633–640 (2019).

49. Rathe, E. J. Note on two common problems of sound propagation. *J. Sound Vib.* **10**, 472–479 (1969).

50. Sun, S. *et al.* Structural origin of size effect on piezoelectric performance of Pb(Zr,Ti)O3. *Ceram. Int.* **47**, 5256–5264 (2021).

51. Zhu, W., Fujii, I., Ren, W. & Trolier-Mckinstry, S. Domain wall motion in A and B site donor-doped Pb (Zr 0.52 Ti 0.48) O 3 films. *J. Am. Ceram. Soc.* **95**, 2906–2913 (2012).

52. Nguyen, M. D. *et al.* Effect of dopants on ferroelectric and piezoelectric properties of lead zirconate titanate thin films on Si substrates. *Ceram. Int.* **40**, 1013–1018 (2014).

53. Horchidan, N. *et al.* A comparative study of hard/soft PZT-based ceramic composites. *Ceram. Int.* **42**, 9125–9132 (2016).

54. Shintaku, H. *et al.* Development of piezoelectric acoustic sensor with frequency selectivity for artificial cochlea. *Sensors Actuators, A Phys.* **158**, 183–192 (2010).

55. Lang, C., Fang, J., Shao, H., Ding, X. & Lin, T. High-sensitivity acoustic sensors from nanofibre webs. *Nat. Commun.* **7**, 11108 (2016).

56. Lang, C. *et al.* High-output acoustoelectric power generators from poly(vinylidenefluoride-co-trifluoroethylene) electrospun nano-nonwovens. *Nano Energy* **35**, 146–153 (2017).

57. Shao, H. *et al.* Efficient conversion of sound noise into electric energy using electrospun polyacrylonitrile membranes. *Nano Energy* **75**, 104956 (2020).

58. Lu, X. & Dang, J. An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification. *Speech Commun.* **50**, 312–322 (2008).

59. Chettri, B., Kinnunen, T. & Benetos, E. Subband Modeling for Spoofing Detection in Automatic Speaker Verification. Preprint at https://arxiv.org/abs/2004.01922 (2020).

60. Song, K. *et al.* Sound pressure level gain in an acoustic metamaterial cavity. *Sci. Rep.* **4**, 4–9 (2014).

61. Dehé, A. Silicon microphone development and application. *Sensors Actuators, A Phys.* **133**, 283–287 (2007).

62. Défossez, A., Synnaeve, G. & Adi, Y. Real time speech enhancement in the waveform domain. *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH* **2020-October**, 3291–3295 (2020).

63. Wang, Z. Q., Le Roux, J. & Hershey, J. R. Multi-Channel Deep Clustering: Discriminative Spectral and Spatial Embeddings for Speaker-Independent Speech Separation. *ICASSP, IEEE Int. Conf. Acoust. Speech Signal Process. - Proc.* **2018-April**, 1–5 (2018).

Acknowledgements

This work was supported by Wearable Platform Materials Technology Center (WMC) (NRF-2016R1A5A1009926), and Convergent Technology R&D Program for Human Augmentation (NRF-2020M3C1B8081519) through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT. This work was also supported by the Technology Innovation Program(Fashionable Smart Protection Suit for Monitoring Sterilizing and Blocking of Harmful Factors, 20012500) funded By the Ministry of Trade, Industry & Energy(MOTIE, Korea)

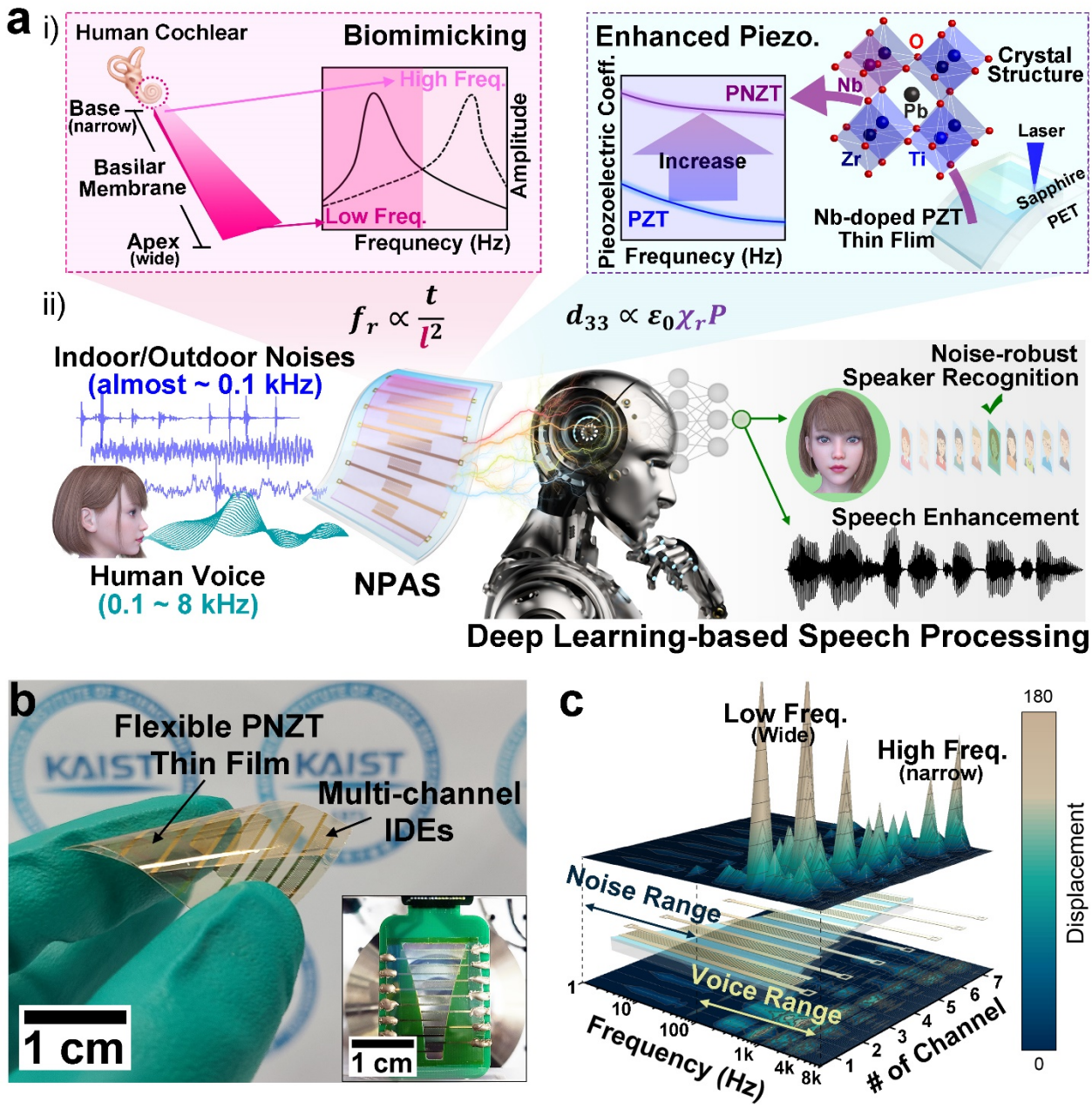
Author Contributions

Y.H.J and T. X. Pham contributed equally to this work. Y.H.J. and K.J.L. designed and carried out the experiments and data analysis. J.H.L and B.-Y Lee performed the characterization of PNZT membrane. W.H.S. and M. Chung performed signal acquisition. T. X. Pham, I. Dias, G. Kim, and C.D.Y. designed the deep learning algorithm. Y.H.J. and W.H.S. contributed FEM simulations. Y.H.J., C.D.Y., and K.J.L. contributed to writing the manuscript.

Competing Interests

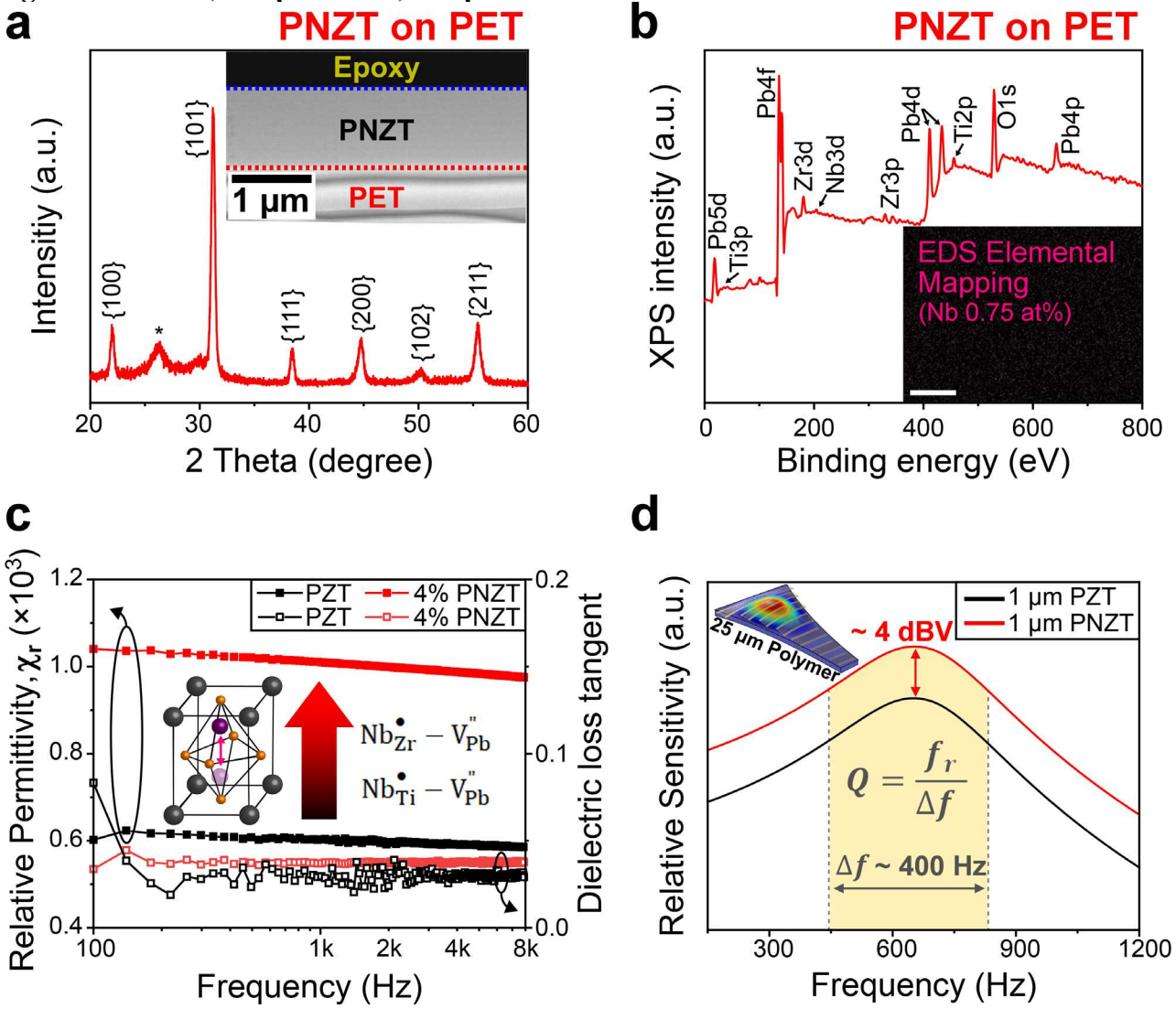
The authors declare no competing interests.

Fig. 1: Overall concept of biomimetic NPAS and deep learning-based speech processing.



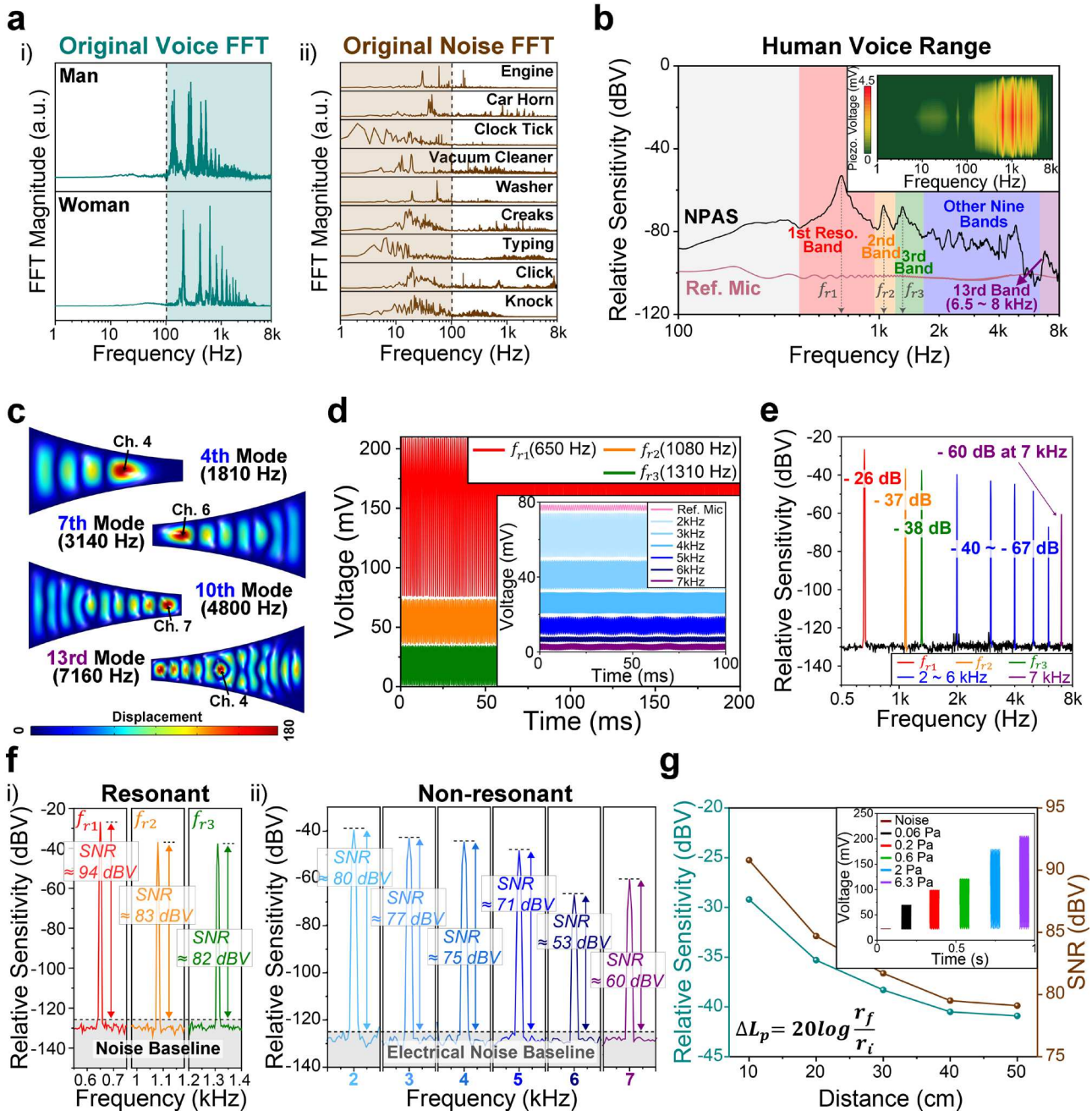
a Schematic illustration of biomimetic self-powered acoustic sensor and deep learning-based speech processing under noise condition: (i) Basilar membrane-inspired NPAS fabricated with highly sensitive PNZT thin film to include multi-resonant frequency bands into the human voice range from 100 to 8000 Hz. (ii) Noise-robust speaker recognition and speech enhancement of the NPAS using deep learning algorithms. **b** Photograph of the flexible multi-channel NPAS membrane bent by human fingers. Inset shows the NPAS attached on a PCB. **c** Multi-resonant displacement of the NPAS thin film measured by LDV under a frequency sweep from noise-dominant range to human voice spectrum.

Fig. 2: Structural, compositional, and piezoelectric characterizations of flexible PNZT thin film.



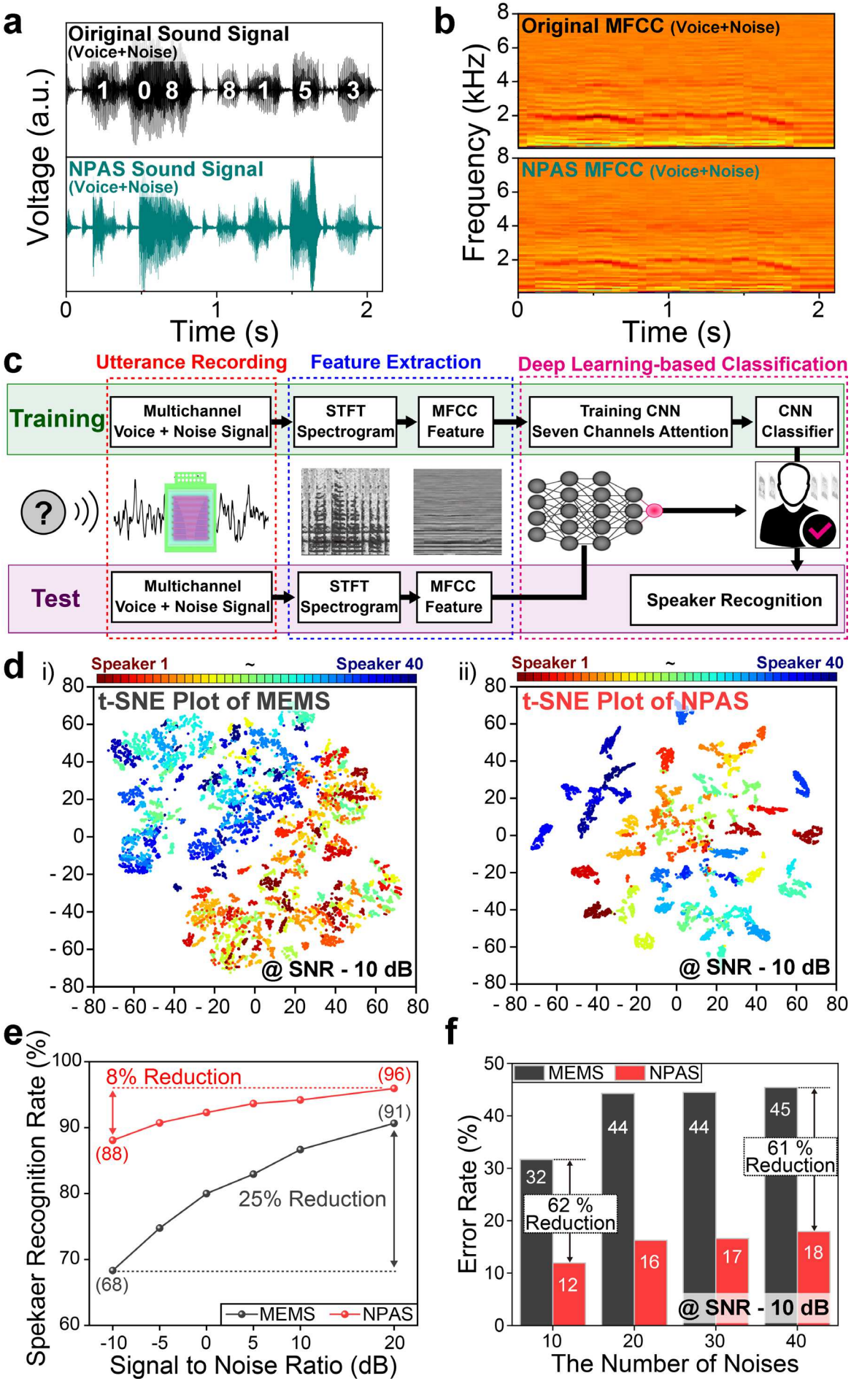
a XRD patterns of PNZT thin films on plastic substrates. The asterisks denote the specific peaks from each substrate. Insets show cross-sectional SEM images of the PNZT membrane on sapphire and PET substrates, respectively. **b** XPS analysis of the PNZT membrane after ILLO transfer. Insets display EDS elemental mapping results (Scale bar: 3 μm). **c** Dielectric and loss properties of the PZT and 4% Nb-doped PZT thin films. **d** Comparison of resonant frequency bands between PZT and PNZT membranes on 20 μm thick polymer.

Fig. 3: Electrical characterizations of NPAS.



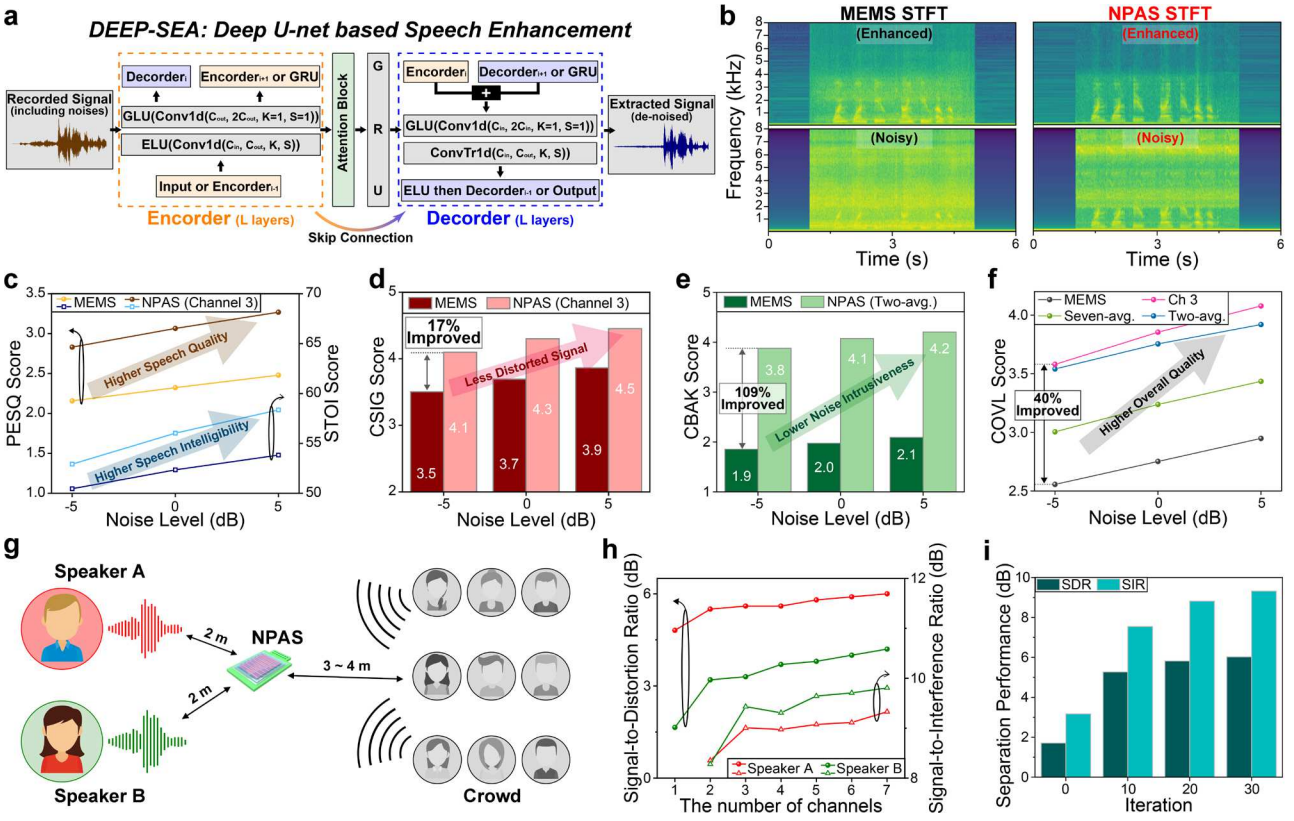
a FFT response of male/female voices (i), and indoor/outdoor noise sounds (ii) from 1 Hz to 8 kHz. **b** Frequency response of NPAS and reference microphone over the human voice spectrum with white noise input. Relative sensitivity of the NPAS plotted by selecting the highest value among multi-channel bands. Inset shows the maximum piezoelectric voltage of the NPAS membrane from 1 Hz to 8 kHz under white noise condition. **c** Multi-resonance of the curved NPAS thin film investigated by FEM simulation. **d** The piezoelectric voltages of the NPAS membrane at the first, second, and third resonance frequencies. The inset shows the output voltage of the other resonances, compared to a reference microphone. **e** Sensitivities of the resonant frequencies expressed in units of dBV under monochromatic sound waves. Red, orange, and green indicate the first, second, and third resonances, while blue and purple denote 6 single frequencies from 2 kHz to 7 kHz. **f** The calculated SNRs of the first, second, third resonances (i) and other monochromatic sinusoidal sound (ii) by the deviation in the sensitivities and electrical noise baseline. **g** The sensitivity and SNR of NPAS calculated as a function of distance at first resonance. Inset displays the output voltage of NPAS at first resonance under different pressure conditions.

Fig. 4: Deep learning-based noise robust speaker recognition of NPAS.



a Comparison of original and NPAS sound waveform including voice and background noises. **b** MFCC features of the original and NPAS signals, presenting the similarity in both the time and frequency domains. **c** Deep learning-based flow chart of training/test procedures for speaker recognition by using 2800/280 voice dataset of TIDIGITS. **d** MFCC features visualized in a t-SNE plot using 280 test datasets (40 people, 7 utterances) for a commercial MEMS microphone (i) and NPAS (ii). The t-SNE plot presents the high-dimensional data of similar speakers in 2D space with probability distribution. **e** Recognition rate of the NPAS exhibiting superior noise-robustness in the condition of 10 noises, compared to a commercial phone. **f** Speaker recognition error rate of the NPAS surpassing the commercialized MEMS microphone according to the number of noises.

Fig. 5: Comparison of speech enhancement between NPAS and commercial MEMS microphone.



a DEEP-SEA model flow chart of overall process for extracting the de-noised voice signal from recording input. **b** Comparison of speech enhancement signals between MEMS microphone and NPAS. Speech enhancement was analyzed using STFT spectrogram. **c-f** The objective measures to investigate the filtering performance of NPAS and commercialized phone. The calculated PESQ and STOI values showing higher speech quality and intelligibility of NPAS compared to the commercial MEMS microphone (**c**). The CSIG evaluation indicating the signal distortion under noise levels from -5 dB to 5 dB (**d**). Comparison of the background noise in the speech signals between NPAS and commercial phone by measuring CBAK score (**e**). The COVL measures to calculate the overall effect on the speech quality of the extracted signal (**f**). **g-i** Experiment of separating each speaker's voice from the crowd by using the multi-channel NPAS. Schematic illustration of the speaker separation condition (**g**). The evaluated SDR and SIR for the separated speech signals according to the number of NPAS channels (**h**). Analysis of SDR and SIR as a function of iterations (**i**).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementrayInformation.docx
- SupplementaryMovie1.mp4
- SupplementaryMovie2.mp4