nature communications

Article

https://doi.org/10.1038/s41467-025-59523-6

Deep learning-enhanced anti-noise triboelectric acoustic sensor for human-machine collaboration in noisy environments

Received: 23 May 2024

Accepted: 24 April 2025

Published online: 08 May 2025

Check for updates

Chuanjie Yao[1,2], Suhang Liu[1,2], Zhengjie Liu[1,2], Shuang Huang[1,2,3], Tiancheng Sun[1,2], Mengyi He[1,2], Gemin Xiao[4], Han Ouyang[5], Yu Tao[4], Yancong Qiao[3], Mingqiang Li[4], Zhou Li[6], Peng Shi[7], Hui-jiuan Chen[1,2] & Xi Xie[1,2,3]

Human-machine voice interaction based on speech recognition offers an intuitive, efficient, and user-friendly interface, attracting wide attention in applications such as health monitoring, post-disaster rescue, and intelligent control. However, conventional microphone-based systems remain challenging for complex human-machine collaboration in noisy environments. Herein, an anti-noise triboelectric acoustic sensor (Anti-noise TEAS) based on flexible nanopillar structures is developed and integrated with a convolutional neural network-based deep learning model (Anti-noise TEAS-DLM). This highly synergistic system enables robust acoustic signal recognition for human-machine collaboration in complex, noisy scenarios. The Anti-noise TEAS directly captures acoustic fundamental frequency signals from laryngeal mixed-mode vibrations through contact sensing, while effectively suppressing environmental noise by optimizing device-structure buffering. The acoustic signals are subsequently processed and semantically decoded by the DLM, ensuring high-fidelity interpretation. Evaluated in both simulated virtual and real-life noisy environments, the Anti-noise TEAS-DLM demonstrates near-perfect noise immunity and reliably transmits various voice commands to guide robotic systems in executing complex post-disaster rescue tasks with high precision. The combined anti-noise robustness and execution accuracy endow this DLM-enhanced Anti-noise TEAS as a highly promising platform for next-generation human-machine collaborative systems operating in challenging noisy environments.

With the accelerated development of artificial intelligence (AI) and the internet of things technologies, human-machine interaction (HMI) has attracted extensive attention in applications of health monitoring, post-disaster rescue, intelligent control, and so on[1–3]. The HMI systems have also evolved from initial manipulation of external devices (keyboard, remote controls, etc.) to more direct and efficient modalities, including body-movement interaction, electroencephalogram-based control, and voice interaction. Although the HMI with body-movement

A full list of affiliations appears at the end of the paper. e-mail: xiexi27@mail.sysu.edu.cn

offers operational simplicity and ease of implementation, the lack of intuitiveness and limited command repertoire restrict its utility in complex interactive tasks[4–8]. Electroencephalogram-based HMI (e.g., electromyography or electroencephalogram) holds theoretical promise for multifunctional control by decoding diverse electrophysiological signals. However, the incomplete understanding of complex electrophysiological patterns[9–11], coupled with the bulkiness of multi-channel sensor arrays and signal acquisition hardware[12–14], currently hinders both accuracy and practical wearability. As a common and natural communication medium, speech signals encode not only semantic content but also paralinguistic cues such as speaker identity and emotional state[15]. Human-machine voice interaction (HMVI) leverages these advantages, offering an intuitive, efficient, and information-rich interface for complex collaborative tasks[16,17].

The rapid advancement of speech recognition technology and intelligent devices has broadly prompted the applications of conventional HMVI, driving their widespread commercialization[18]. At present, regular commercial microphones mainly use rigid acoustic sensors (e.g., moving coil or electret condenser) for signal acquisition, which inherently limits HMVI system wearability[19]. Flexible wearable acoustic sensors based on diverse sensing mechanisms, including piezoresistive[20,21], piezoelectric[22,23], and triboelectric sening[24–26], potentially provide more practical solutions for the HMVI applications, demonstrating desirable performance in acoustic signal acquisition. Among these, triboelectric acoustic sensors possess advantages in capturing periodic subtle vibrational signals, owing to their high sensitivity, high response speed, and cost-effectiveness[27–35]. The triboelectric sensors are simple in structure, which can be developed into wearable acoustic sensors, to convert weak vibration into acoustic signals by attaching them to the throat area. However, human speech production involves complex coordination of multiple physiological components (throat, tongue, and facial muscles) to generate air vibrations[36–38], meaning that laryngeal vibration alone cannot capture the complete set of speech recognition features. In addition, conventional acoustic sensors are typically designed for daily life conditions, which are often seriously interfered with by noisy environments (e.g., crowded areas, fire emergencies, or heavy rainfall).

Deep learning model (DLM) has emerged as a powerful tool for feature extraction, data analysis, and target classification, enabling efficient processing and recognition of speech signals for semantic interpretation, biometric identification, and sentiment analysis[39–42]. Recently, emerging studies have integrated DLM with flexible wearable acoustic sensors to further enhance HMVI efficiency[43–45]. For instance, Ren et al. reported an intelligent artificial pharyngeal sensor based on laser-induced graphene, which achieved accurate recognition of daily vocabulary for people with vocal impairments with the assistance of the DLM[38]. Lee et al. developed a flexible resonant acoustic sensor based on the piezoelectric principle and combined it with a DLM to achieve accurate biometric authentication[22]. In addition, Chen et al. developed a high-fidelity waterproof acoustic sensor and realized highly accurate biometric authentication with the support of the DLM[26]. Notably, DLMs can compensate for feature-deficient or waveform-distorted acoustic signals by leveraging their advanced data processing and pattern recognition capabilities to reconstruct and interpret speech information with high accuracy. Further exploration of the systematic integration of DLMs and triboelectric acoustic sensors will provide a practical approach for HMVI collaboration in noisy environments.

In this work, we developed an anti-noise triboelectric acoustic sensor (Anti-noise TEAS) based on flexible nanopillar structures, which was further integrated with a convolutional neural network (CNN)-based DLM to enable robust acoustic signal recognition (ASR) for complex HMVI in noisy environments. The Anti-noise TEAS could directly acquire the acoustic fundamental frequency signals from laryngeal mixed-mode vibrations through contact sensing, enabling to capture of multiple types of acoustic signals with high sensitivity, high stability, and a broad response frequency range. By optimizing the device layer structure, the sensor could effectively buffer the interference of the noisy environments, while the accurate meaning of the distorted acoustic signals was recognized with the assistance of the CNN-based DLM. Therefore, the Anti-noise TEAS enhanced by DLM (Anti-noise TEAS-DLM) achieved remarkable speech recognition accuracy exceeding 99% in high-noise conditions, demonstrating near-perfect immunity to ambient interference. In simulated virtual and real-life scenarios, the Anti-noise TEAS-DLM was applied to transmit voice commands for guiding robotic systems to perform complex post-disaster casualty rescue tasks. While conventional microphone-based systems failed to complete tasks owing to noise interference, the Anti-noise TEAS-DLM maintained reliable noise resistance, enabling precise robotic task execution. The Anti-noise TEAS-DLM developed in this work provided a practical solution for HMVI collaborative tasks in noisy environments, including post-disaster rescue, collaborative operations, and wilderness exploration. This AI-driven system was beneficial for promoting the development of a diversified system for HMVI to meet the complex scenarios in real-world noisy conditions.

Results

Figure 1a illustrates the working mechanism of the deep learning-enhanced Anti-noise TEAS to perform complex human-machine collaboration in noisy environments. The Anti-noise TEAS possessed desirable anti-noise interference capability, which could capture the fundamental acoustic signals of laryngeal vibrations through contact sensing in noisy scenarios, such as rainstorms, gusty winds, earthquakes, and other noisy environments. Regular acoustic sensors usually detect the speech acoustic signals after propagation through the air medium, which are easily interfered with and even drowned out by environmental noise. However, with special anti-noise structure design and optimized materials, the Anti-noise TEAS could effectively buffer the interference of environmental noise. By attaching to the cartilage of the human throat, the Anti-noise TEAS could directly capture the mixed-mode signals of acoustic (the weak vibration of vocal folds) signals and mechanical motion (the tiny movements of muscle) signals through contact sensing (Fig. 1b). These mixed-mode acoustic signals contained partial speech recognizable features, which could reflect the semantic content to some extent. While the special structure and contact sensing of the Anti-noise TEAS were beneficial to buffer the environmental noise, the acoustic signals from laryngeal vibrations could also be compromised or distorted, losing the resonance features of the tuning organs, such as the mouth, tongue, and pharynx. This resulted in difficulties in directly recognizing these signals by human hearing and traditional speech processing techniques. As a powerful tool with feature extraction and data analysis capabilities, a CNN-based DLM could potentially enhance the Anti-noise TEAS in the aspect of recognition of the distorted signals. After data processing and feature extraction (Fig. 1c), the CNN-based DLM was utilized to parse and classify the acoustic signals, and obtained both the semantic and identity information (Fig. 1d). Finally, the command signals were wirelessly transmitted to the swarms of robots, which could perform the complex human-machine collaborative tasks in harsh scenarios (Fig. 1e).

The Anti-noise TEAS was composed of a flexible carbon nanotubes/polydimethylsiloxane (CNTs/PDMS) nanopillars substrate as a positive friction electrode and a fluorinated ethylene propylene (FEP) as a negative friction electrode (Fig. 1f). As a soft lithography technique, replica-molding was able to replicate the micro/nano-structures on the surface of the sample cost-effectively and rapidly (Supplementary Fig. S1). Herein, the nanopillars structure was fabricated on a flexible CNTs/PDMS substrate (Supplementary Fig. S2) by replica-molding from the surface structure of cicada wings. This nano-scale

https://doi.org/10.1038/s41467-025-59523-6

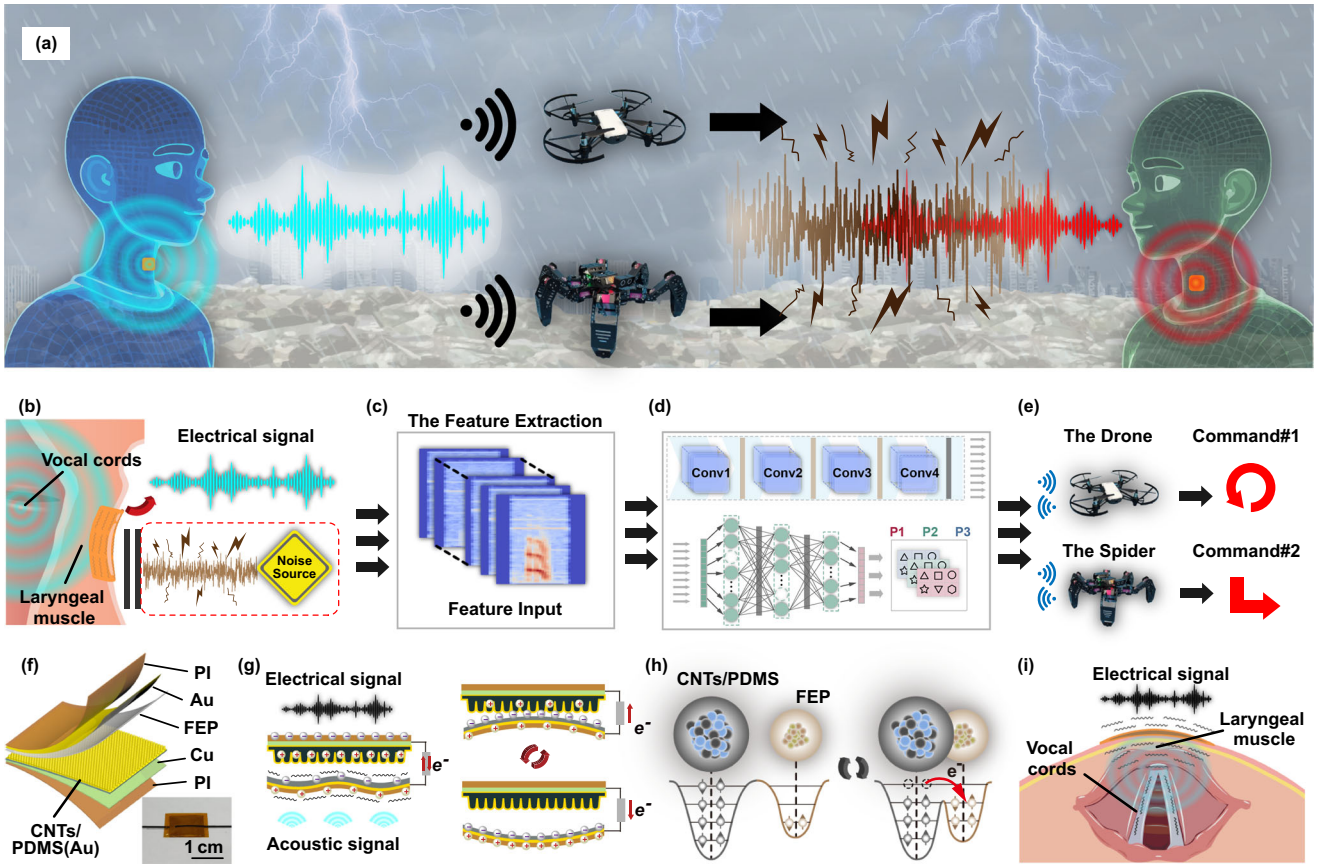


Fig. 1 | Schematic of the Anti-noise TEAS-DLM for human-machine collaboration in noisy environments. a Diagram of the working mechanism of the deep learning-enhanced Anti-noise TEAS to perform complex human-machine collaborative tasks in noisy scenarios. **b** The Anti-noise TEAS detected mixed-mode signals of acoustic (weak vibrations of the vocal cords) signals and mechanical motion (tiny movements of muscle) signals via contact sensing, which could block the interference of environmental noise. **c** Pre-processing and feature extraction of the acoustic signals to obtain the input feature spectrogram of the convolutional neural network (CNN)-based DLM. **d** Schematic diagram of the CNN-based DLM for ASR. **e** Controlling robotic systems, including the robot, the drone, etc., through the Anti-noise TEAS-DLM. **f** Structural schematic and photographs of the Anti-noise TEAS, with scale bar of 1 cm. **g** Working principle of the Anti-noise TEAS to realize contact-separation acoustic sensing. **h** Schematic of the atomic-scale model of the Anti-noise TEAS, where electrons were transferred from the positive materials to the negative materials when the two friction electrodes contacted. **i** Diagram of the Anti-noise TEAS attached to the laryngeal cartilage position to detect mixed-mode acoustic signals via the coupling of contact electrification and electrostatic induction.

pillar structure could effectively expand the actual contact area of the friction layer interface, inducing more charge transfer and higher electrical output. In addition, a layer of ~300 nm Au on the surface of the flexible CNTs/PDMS nanopillars substrate was sputtered, to improve the substrate conductivity. After attaching a layer of conductive copper tape on the backside of the substrate, the positive friction electrode of the Anti-noise TEAS was obtained. A thin FEP film, which was coated with an Au layer on backside, served as a negative friction electrode. The friction layer interface of the FEP film was corona-treated to increase the surface charge density at the interface[46,47]. The as-prepared positive electrode was assembled with the FEP negative electrode to form the triboelectric sensors in the contact-separation mode, where a 300 μm-thick polyethylene glycol terephthalate (PET) film was used as the spacer between the two electrodes. The Anti-noise TEAS was encapsulated with Kapton tape, to increase the robustness of sensors and the interface contact between the sensors and skins[48]. The working principle of the Anti-noise TEAS was based on contact-separation sensing through the coupling of contact electrification and electrostatic induction (Fig. 1g). As a diaphragm, the FEP negative electrode would generate vibrations in response to acoustic signals. Due to the different electronic affinity capability of two friction materials, electrons were transferred from the positive frictional interface to the negative frictional interface, while contact occurred between the two friction electrodes (Fig. 1h). Thus, when the FEP diaphragm was close to the positive electrode, electrons would flow from the positive electrode to negative electrode through an external circuit. In contrast, when the FEP diaphragm was away from the positive electrode, electrons would flow back from the negative electrode to the positive electrode through the external circuit. By attaching the Anti-noise TEAS to the human larynx (Supplementary Fig. S3 and Supplementary Note S1), the sensor would generate corresponding electrical signals (current and voltage signals) in response to the laryngeal vibrations, including weak vibration of the vocal folds and tiny muscle movements, which could reflect the amplitude and frequency of the acoustic fundamental frequency signals in the human voice (Fig. 1i).

The acoustic response characteristics of acoustic sensors were important for the application of the HMVI. We employed a standard player, a sound pressure level (SPL) detector, and a precision source/measurement unit (SMU) to evaluated the acoustic response characteristics of the Anti-noise TEAS (Supplementary Fig. S4). The standard player was used as a sound source generator, which could apply acoustic signals of different frequencies and SPL to the Anti-noise TEAS. The SPL detector was used to detect acoustic signals and environmental noise, which could provide standardized SPL reference values. The SMU could detect and record the acoustic-current response signals of the Anti-noise TEAS in real time. The effect of designed parameters of the Anti-noise TEAS, such as spacer thickness

https://doi.org/10.1038/s41467-025-59523-6

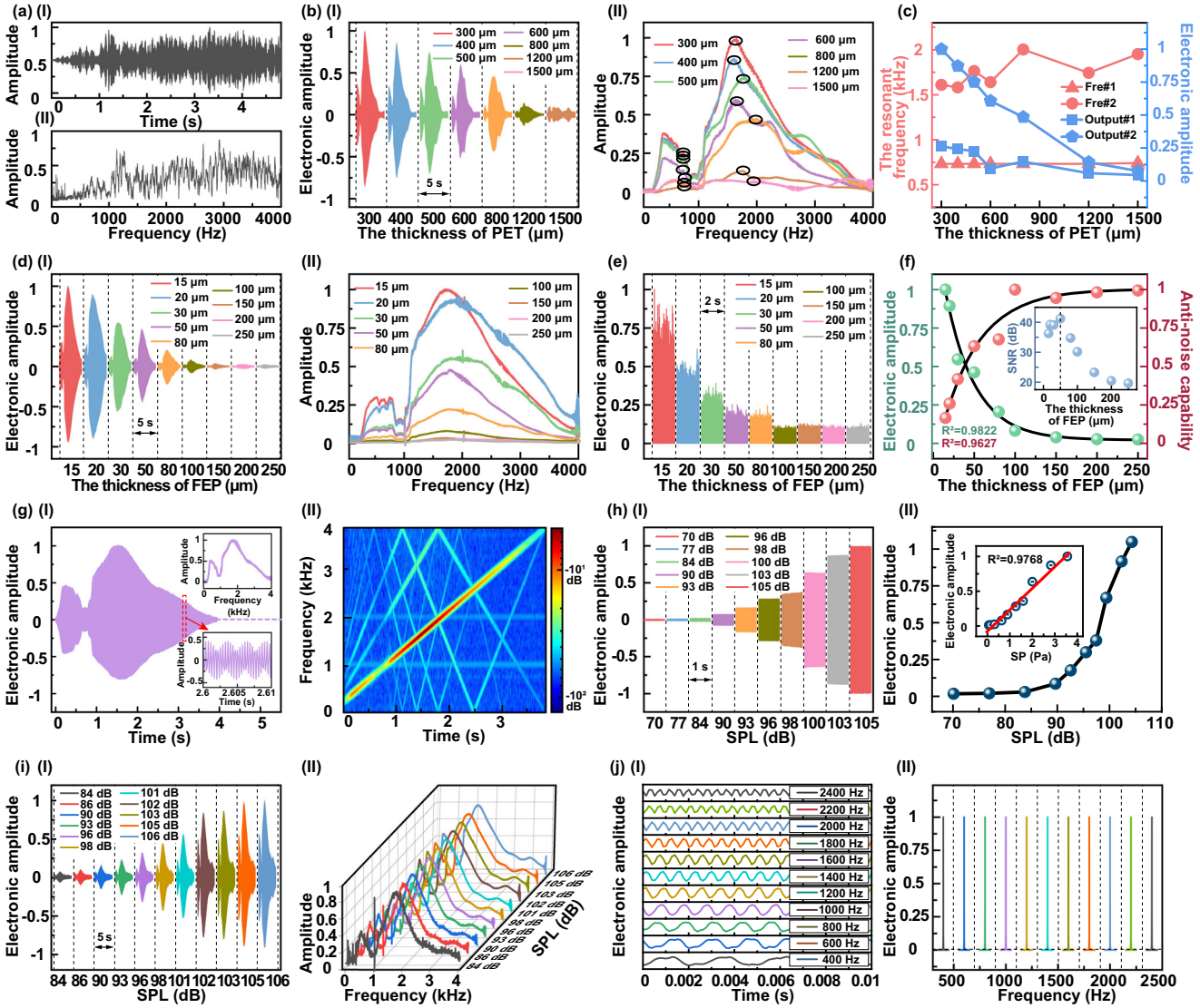


Fig. 2 | Acoustic response characteristics of the Anti-noise TEAS. a (I) Waveform and (II) spectrogram of a linear sweep audio source recorded by a regular standard microphone. **b** (I) Sweep signal curves and (II) frequency characteristic curves recorded by the Anti-noise TEAS with different spacing sizes. **c** Frequency and peak value of the resonant frequencies vs the different spacing distances for the Anti-noise TEAS. **d** (I) Sweep signal curves and (II) frequency characteristic curves recorded by the Anti-noise TEAS with different FEP diaphragm thicknesses. **e** White noise signals recorded by the Anti-noise TEAS with different FEP diaphragm thicknesses through non-contact sensing. **f** Comparison of output performance and anti-noise capability of the Anti-noise TEAS vs the FEP diaphragm thicknesses. The inset presented the SNR of the Anti-noise TEAS vs the FEP diaphragm thicknesses. **g** (I) Typical linear sweep audio source curve obtained with the Anti-noise TEAS. The inset showed the frequency characteristic curve and detailed waveform, respectively. (II) Spectrogram of the sweep signal recorded by the Anti-noise TEAS. The red color in the spectrogram represented higher intensity, and the blue color meant lower intensity. **h** (I) Waveforms of acoustic signals (2 kHz) detected by the Anti-noise TEAS with different SPL from 70 dB to 105 dB. (II) Curves of different SPL vs the output amplitude of the acoustic-current response signals of the Anti-noivs TEAS. The inset exhibited curves of different SP vs the amplitude of the current. **i** (I) Waveforms of sweep signals and (II) frequency characteristic curves of different SPL detected by the Anti-noise TEAS. **j** (I) Waveforms and (II) frequency characteristics of different single-frequency acoustic signals detected by the Anti-noise TEAS.

and FEP thickness, on the acoustic response characteristics was systematically investigated. Firstly, the Anti-noise TEAS was prepared with various PET spacer thicknesses, ranging from 300 μm to 1500 μm. The acoustic response characteristics were evaluated by a sweep source signal with a frequency of 50–4000 Hz, a SPL of 102 dB, and a duration time of 5 s (Fig. 2a), which basically covered the main fundamental frequency range during human speech communications. The acoustic-current response signals recorded by the Anti-noise TEAS with different spacing thicknesses are shown in Fig. 2b-I. When the spacing distance was <800 μm, the envelope contours of the acoustic-current response signals exhibited high consistency, while the overall amplitude of the signals decreased with increasing spacing distance. This was likely because the electrostatic induction in the sensing process decayed as the spacer distance increased, resulting in a diminishing amplitude of the acoustic-current response signals. When the spacer distance increased to >800 μm, the stability of the Anti-noise TEAS during sensing was affected, resulting in a significant deformation of the envelope contours and a further decrease in the amplitude. Fast Fourier transforms (FFT) were performed on the acoustic-current response signals at different spacing distances to obtain linear standard sweep frequency characteristic curves. The frequency characteristic curves of the Anti-noise TEAS presented a

similar situation to the acoustic-current response signals, i.e., the overall amplitude of the frequency characteristic curves diminished with the increase of the spacing distance (Fig. 2b-II). The 1st and the 2nd resonance frequencies were marked, and the relationship of frequencies (red curve) and peak values (bule curve) with spacing distance were plotted (Fig. 2c). The amplitude of the resonance frequency peaks declined with the increase of the spacing distance, and resonance frequency rose slightly with the increase of the spacing thickness, i.e., a red shift of the resonance frequency occurred. Since the flexible CNTs/PDMS substrates naturally featured various height undulations of ~200 μm, the PET layer was optimized with 300 μm thickness as the spacer, to ensure sufficient separation of friction electrodes.

The effect of the FEP diaphragm with various thicknesses (ranging from 15 μm to 250 μm) on the acoustic response characteristics was also evaluated with the sweep source signal. The acoustic-current response signals recorded by the Anti-noise TEAS with different FEP diaphragm thicknesses were exhibited in Fig. 2d-I. The envelope contours of the response signals of different FEP diaphragm thicknesses were pretty similar, and the output amplitude of the signals diminished as the FEP diaphragm thickness increased. The negative electrode stiffened as the thickness of the FEP diaphragm increased, resulting in less vibration of the friction layer during acoustic sensing, which in turn led to a reduction in the output signals. The FFT was performed on the response signals to obtain the frequency characteristic curves at various FEP diaphragms (Fig. 2d-II). The amplitude of the frequency characteristic curves decreased as the FEP thickness increased. The maximum peak of the frequency characteristic curves was identified as the output amplitude of the Anti-noise TEAS (Fig. 2f). The output performance of the sensor exhibited an exponential negative correlation with the FEP diaphragm thickness ($R^2 = 0.9822$). When the FEP diaphragm increased to 100 μm, the response signals of the Anti-noise TEAS decreased to ~0.11-fold compared to that of the 15 μm thickness. While a thinner FEP diaphragm was desirable for enhancing output amplitude of the Anti-noise TEAS, the thicknesses effects on the anti-noise capability of sensor needed to be further investigated. Since the Anti-noise TEAS was a contact sensor oriented towards the detection of fundamental frequency acoustic signals of human larynx, the interference signals sensed spatially (non-contact) were one of the main noise sources during its operation. A white noise signal (SPL = 102 dB) was applied onto the Anti-noise TEAS, and recorded the interference signals. The amplitude of the noise signals detected by the Anti-noise TEAS declined with the increase of the FEP diaphragm thickness (Fig. 2e). The inverse of the average noise amplitude was used as an indicator to evaluated the anti-noise capability, where an exponential positive correlation with the thickness of the FEP diaphragm ($R^2 = 0.9627$) was revealed (Fig. 2f). Comparing the effective output signals (contact sensing) and noise interference signals (spaced sensing) detected with different FEP diaphragm thicknesses, a relationship between the signal-to-noise ratio (SNR) and the FEP diaphragm thickness (inset of Fig. 2f) was obtained. The curves of the effective output signal and the noise interference signal intersected at ~50 μm of the FEP diaphragm. The SNR of the Anti-noise TEAS reached its maximum value at 50 μm thickness of the FEP diaphragm. To balance good output performance and anti-noise capability, the FEP film with 50 μm thickness was used as the negative diaphragm of the Anti-noise TEAS.

Therefore, the optimal structure design of the Anti-noise TEAS was the 300 μm-thick PET film as the spacer and the 50 μm-thick FEP film with as the negative diaphragm. With standardized preparation processes and optimal structure design, the Anti-noise TEAS was able to simultaneously possess desirable output performance and noise immunity. The optimal Anti-noise TEAS possessed a wide frequency response range (Fig. 2g-I), which covered the main range of fundamental frequency of acoustic signal in communications. In addition, the Anti-noise TEAS was capable to record the acoustic signal in different frequency ranges (Supplementary Fig. S5). The frequency characteristic curves were converted to spectrogram by short-time Fourier transform (STFT) with a window of 800 ms and a moving step of 200 ms. The color depth in the spectrogram reflected the magnitude of the acoustic signal, with darker colors (red) representing larger magnitudes and lighter colors (blue) indicated lower magnitudes. As presented in Fig. 2g-II, the intensity of the acoustic spectrogram was mainly concentrated on the diagonal line, indicating that the Anti-noise TEAS possessed a uniform linear response to sweep source. In addition, the octave signal of the sweep source was clearly visible in the spectrogram, which implied that the Anti-noise TEAS possessed the potential to detect acoustic signals at higher frequencies.

As an important metric, acoustic sensitivity of the Anti-noise TEAS was also evaluated by acoustic signals (at 2 kHz) with various SPL from 70 dB to 105 dB. The curves of the SPL vs the acoustic-current response signals of the sensor were exhibited in Fig. 2h-I, demonstrated that the amplitude of the Anti-noise TEAS raised continuously with SPL increasing. The formula for converting SPL to sound pressure (SP) was shown as followed:

$$SPL = 20 \lg \frac{SP}{SP_{ref}} \tag{1}$$

$$S = \frac{A}{SP} = \frac{A}{SP_{ref} \times 10^{\frac{SPL}{20}}} \tag{2}$$

Where SPL was the sound pressure level, SP was the sound pressure, SP_{ref} was the standard reference sound pressure with a value of 2×10^{-5} Pa, and A was the amplitude of output current. The SPL was converted to SP according to Eq. (1) and linearly fitted to the amplitude of the response signals (Fig. 2h-II). The response signals exhibited a great linear positive relationship with the SP ($R^2 = 0.9768$), with a sensitivity of 0.05 nA Pa^{-1}, which was represented by an amplitude with the SPL of 94 dB (SP = 1). In addition, the Anti-noise TEAS was also employed to detect sweep sources with different SPL ranging from 84 dB to 106 dB. The response signals of the Anti-noise TEAS exhibited a positive correlation with SPL increasing (Fig. 2i-I), while the envelopes of signals maintained a great consistency. Subsequently, the acoustic-current signals were converted into frequency characteristic curves with FFT, which exhibited a high degree of uniformity across various SPLs. However, due to differences in SNR at various SPL, the frequency characteristic curves exhibited fluctuations and distortion at lower SPL, while the overall signal envelopes remained relatively stable (Supplementary Fig. S6). To further validate the effectiveness of the Anti-noise TEAS in discerning acoustic signals of different single frequencies, the sensor was employed to detect various single-frequency sinusoidal acoustic signals ranging from 400 Hz to 2400 Hz. As shown in Fig. 2j-I, the Anti-noise TEAS was able to record the waveforms at different frequencies effectively (Supplementary Fig. S7). The frequency characteristic curves at different frequencies were obtained by FFT, demonstrating that the Anti-noise TEAS possessed good singularity and consistency in acquiring acoustic signals (Fig. 2j-II and Supplementary Note S2).

As an essential indicator of the Anti-noise TEAS, the durability of the Anti-noise TEAS was evaluated with a periodic acoustic signal at 200 Hz for a total period of ~7×10^5. During the long-term acoustic signal detecting process, the amplitude of the acoustic-current response signal did not undergo significant change and decline (Fig. 3a). The initial and final signal waveforms exhibited good consistency, with the overall fluctuation range of the output current being less than 1.5%. This indicated that the Anti-noise TEAS possessed excellent durability and stability after long-term simulation. Furthermore, to assess the feasibility of the Anti-noise TEAS in HMVI,

https://doi.org/10.1038/s41467-025-59523-6

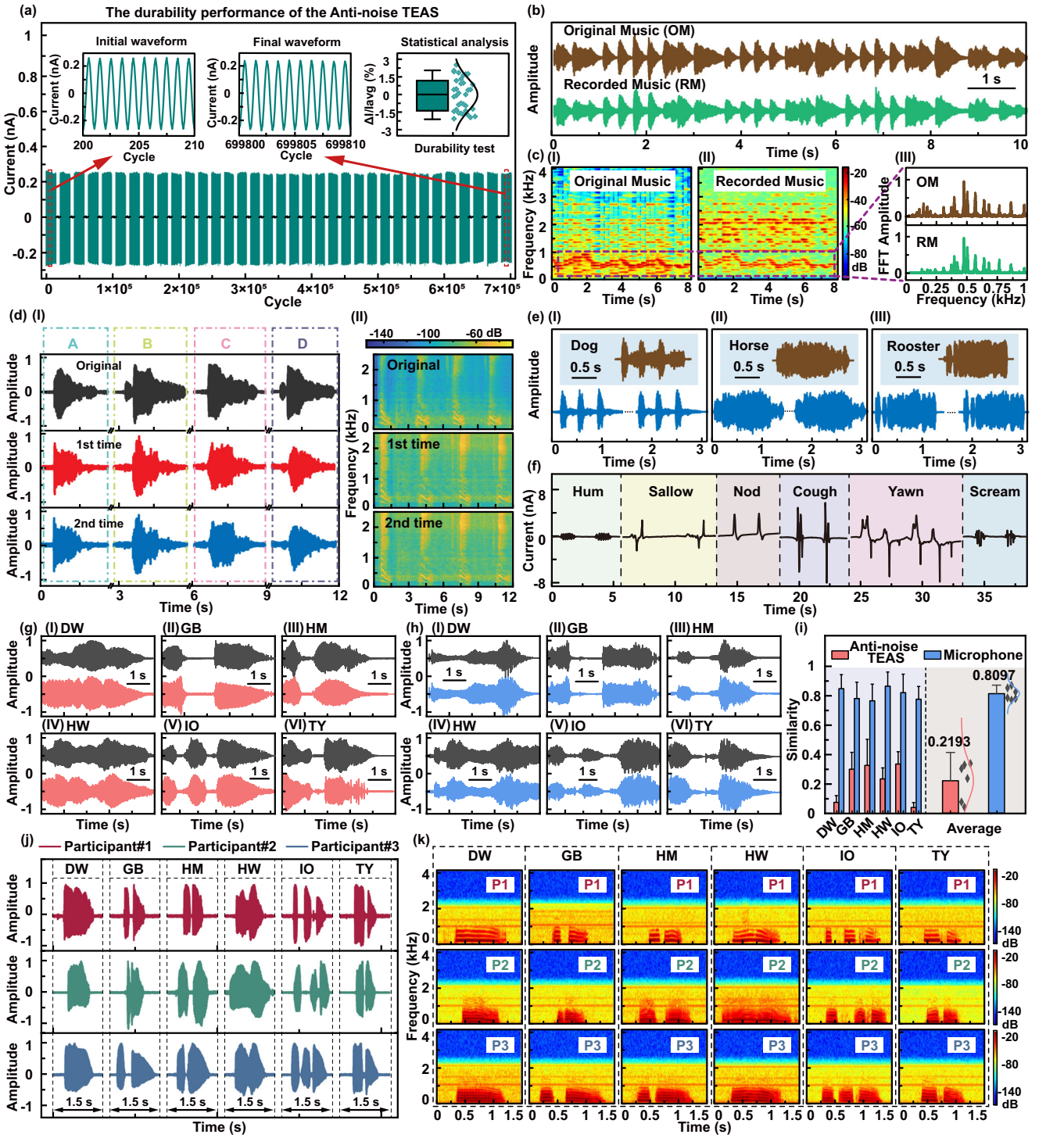


Fig. 3 | Acoustic sensing performance of the Anti-noise TEAS. a Durability test of the Anti-noise TEAS, utilizing an acoustic signal at 200 Hz for ~7×10^5. The inserts presented the output signals during the initial and final periods, along with a statistical analysis of output current at various stages. Graph bars represented standard deviation (SD), and the center line presented the average value. The sample size was 35. **b** Comparison of signal waveforms from the same piece of classical music recorded using the Anti-noise TEAS and a standard microphone. **c** Spectrogram comparison of the music recorded using (I) the Anti-noise TEAS and (II) the microphone, along with the (III) power spectra under the 1 kHz range. **d** (I) Waveform comparisons of A, B, C, and D were detected using the Anti-noise TEAS. (II) Spectrogram comparisons between the detected and original ones. **e** Waveform comparisons of a (I) dog, (II) horse, and (III) rooster detected by the Anti-noise TEAS with their original waveforms. **f** Waveforms of motion artifact interference caused by different physiological activities such as hum, sallow, etc. **g** Comparison of 6 acoustic signals detected by the Anti-noise TEAS with standard speech signals, such as thank you, help me, etc. **h** Comparison of six acoustic signals detected by the microphone with standard speech signals. **i** Correlation statistics of acoustic signals detected by the Anti-noise TEAS and the microphone compared to standard speech signals. Graph bars represented SD. The sample size was 6. **j** Representative waveforms of 6 acoustic signals detected by three different participants using the Anti-noise TEAS. **k** Representative spectrograms of 6 acoustic signals detected by three different participants using the Anti-noise TEAS.

the sensor was employed to detect various types of acoustic signals, including music, speech, and animal calls. A piece of classical music was recorded by the Anti-noise TEAS and compared to the original waveform acquired with a commercial microphone. As exhibited in Fig. 3b, the waveform profile of the acoustic signal detected by the Anti-noise TEAS was consistent with that of the microphone. While conventional microphones typically utilize a sampling rate of 44.1 kHz or higher, the Anti-noise TEAS was designed to focus on detecting the fundamental acoustic signals of human voices, only employing a sampling rate of 8 kHz. This difference would result in slight distortions and variations in the details of the waveforms. With a window of 1024 ms and a moving step of 256 ms, the music waveforms were converted to spectrograms by STFT. As shown in Fig. 3c-I, II, the spectrograms of the music recorded by the Anti-noise TEAS closely matched those of the original music, particularly in the frequency range below 1 kHz. The FFT power spectra below 1 kHz also confirmed the similarity between the original music and the recorded one (Fig. 3c-III). Although the music signal recorded by the Anti-noise TEAS exhibited slight distortions and discrepancies in spectral details, it overall reproduced the original music signal accurately. Additionally, the Anti-noise TEAS was utilized to detect acoustic signals corresponding to English letter combinations A, B, C, and D twice. Figure 3d-I exhibited the waveforms of the original speech signals for A, B, C, and D, alongside the waveforms of the two speech signals recorded by the Anti-noise TEAS. The waveforms of the two speech signals recorded with the Anti-noise TEAS exhibited good repeatability and consistency with the original signal waveforms. After transforming the acoustic signals into spectrograms by STFT (Fig. 3d-II), the distribution of the Anti-noise TEAS spectrograms demonstrated clear similarity to the original spectrograms, indicating good reproducibility between the spectrograms obtained from the twice tests (Supplementary Fig. S8). Thus, the Anti-noise TEAS was proved effective in recording and differentiating acoustic signals with semantic content. Furthermore, the Anti-noise TEAS was also employed to record the calls of 3 distinct animals, including a dog, a horse, and a rooster. The waveform contours of the animal call signals detected with the Anti-noise TEAS closely matched those of the original signals (Fig. 3e), preserving the characteristic peaks and overall contours of the original animal calls effectively. Additionally, the waveforms obtained from the Anti-noise TEAS exhibited high consistency, which demonstrated the sensor's excellent repeatability in acoustic signal acquisition.

During the acquisition of fundamental acoustic signals, the Anti-noise TEAS was affixed onto the human larynx, maintaining close contact with the skin. With a skeletal conduction-like approach, the Anti-noise TEAS obtained fundamental acoustic signals directly from the wearer's larynx during communication, enabling contact-based acoustic signal detection. Nevertheless, the Anti-noise TEAS was susceptible to motion artifact interference by various physiological activities, such as breathing, nodding, and swallowing, during contact sensing. To evaluate motion artifact interference, the Anti-noise TEAS was affixed onto the wearer's larynx and recorded the motion artifacts induced by the movements, including swallowing, nodding, coughing, and so on. The motion artifact caused by different physiological activities exhibited distinctive recognizable characteristics (Fig. 3f), which demonstrated good repeatability and recognizability among interference signals. These interference signals could be discerned and removed from the fundamental acoustic signals by employing suitable filters and de-baselining algorithms. The effects of vigorous human movements on the output signals of the Anti-noise TEAS were further evaluated (Supplementary Note S3).

Human speech communication was a highly intricate process, where the respiratory organs (lungs, trachea, etc.), vocal organs (vocal cords, larynx, etc.), and articulatory organs (tongue, lip, etc.) collaborated to generate air vibrations, thereby producing corresponding acoustic signals. Within this process, the vibration signals presented on the surface of the throat area typically comprised mixed-mode signals originating from the vibration of the vocal folds and subtle muscular movements. These mixed-mode signals contained certain speech recognizable features, which could reflect the semantic content of the acoustic signals to some extent. With the optimal structure and contact sensing, the Anti-noise TEAS could mitigate the environmental noise interference effectively, yet the acquired acoustic signals were plagued by feature deficiency and waveform distortion. The Anti-noise TEAS could directly record the tiny vibration signals of the vocal folds and the subtle muscular movement signals on the human larynx through contact sensing. Nevertheless, these mixed-mode signals lacked the resonance characteristics contributed by the oral cavity, tongue, pharynx, and other articulatory organs. Consequently, it might be difficult to effectively recognize the complete information encoded within these signals using only human hearing or traditional speech processing techniques. The Anti-noise TEAS was affixed to the larynx of a participant to record six mixed-mode fundamental acoustic signals (for 10 times), including don't worry (DW), help me (HM), hello world (HW), it's OK (IO), and thank you (TY). For comparison, the commercial microphone was also employed under identical conditions to record the 6 acoustic signals, serving as a control group. Figure 3g exhibited the 6 acoustic signals recorded by the Anti-noise TEAS compared with the standard waveforms, while Fig. 3h showed a comparison between the 6 acoustic signals recorded by the microphone and the standard waveforms. The acoustic signals recorded by the Anti-noise TEAS generally exhibited similarity to the standard waveforms in terms of contour, albeit displaying some differences in details. However, the acoustic signals recorded by the microphone closely resembled the standard waveforms both in general contour and details. The average correlation between the recorded and the standard signals was analyzed based on the Pearson correlation coefficient. As exhibited in Fig. 3i, the correlation among the acoustic signals recorded by the Anti-noise TEAS was notably low, with an averaged similarity score of 0.2193, whereas the correlation of the signal recorded by the microphone was as high as 0.8097. Thus, although the anti-noise structure and contact sensing mode of the Anti-noise TEAS enabled inhibition of environmental noise, the acquired mixed-mode signals would suffer from serious feature deficiency and waveform distortion. Furthermore, the correlation among the 6 acoustic signals recorded by the Anti-noise TEAS showed considerable variability, with the highest correlation around 0.3362 (for IO) and the lowest correlation recorded at merely 0.0406 (for TY). Those differences might stem from the various proportions of semantic features embedded within the mixed-mode components (vocal fold vibration and muscle movement) and the resonance components (resonance generated by the articulatory organs). Specifically, when the speech signals comprised a higher proportion of mixed-mode components, the Anti-noise TEAS was better equipped to capture effective, acoustically recognizable features, thereby resulting in a higher correlation. However, these mixed-mode acoustic signals lacked sufficient recognizable features. Traditional speech processing techniques or human hearing perception were insufficient for effectively identifying the complete information embedded within these modality-deficient acoustic signals. To address this limitation, more advanced signal processing and analysis tools, such as DLMs, were required to achieve effective recognition of acoustic signals.

Additionally, we also detected the mixed-mode acoustic signals involving three participants. Specifically, participant#1 (male), participant#2 (female), and participant#3 (male) were employed to utter identical phrases, including DW, HM, HW, IO, and TY, while their acoustic signals were recorded by the Anti-noise TEAS. Figure 3j exhibited representative waveforms of 6 acoustic signals recorded by the 3 participants using the Anti-noise TEAS. Although the recorded acoustic signals showed similarity in the overall waveforms, significant

variabilities occurred in the waveform details and contour features. These differences contained a wealth of speech information, including individual specificities and unique voiceprints among different participants. Such information held significant potential for applications in semantic recognition, voiceprint identification, and biometric authentication. These acoustic signals were converted to the frequency-domain spectra (Supplementary Fig. S9) by FFT. While the frequency range of 3 participants centered from 50 Hz to 1000 Hz, which was located in the fundamental frequency range of human vocal signals, the contours of the spectra exhibited notable distinctions among them. For instance, the first resonance frequencies of the 6 acoustic signals for P#2 were distributed within the range of 250–300 Hz, with a mean first resonance frequency of ~273 Hz. Meanwhile, the first resonance frequencies of P#1 and P#3 were distributed within the range of 100–200 Hz, with mean first resonance frequencies of ~163 Hz and ~166 Hz, respectively (Supplementary Fig. S10). The acoustic signals of these 3 participants were converted into spectrograms by STFT. While the spectrograms of all participants mainly concentrated below 1 kHz, there existed significant variability in the distribution of formants and features (Fig. 3k), where the frequency domain distributions for P#2 were generally higher than those for P#1 and P#3. Hence, the Anti-noise TEAS could effectively capture the fundamental acoustic signals of various participants, while the waveforms, spectrums, and spectrograms exhibited distinct individual variabilities. This indicated that the Anti-noise TEAS was capable of recording the content of the fundamental acoustic signals, which maintained the rich speech recognizable features, such as individual specificity and voiceprint. Integrating these diverse recognizable features with AI models would facilitate the potential applications and scalability of the Anti-noise TEAS in semantic recognition, voiceprint identification, and biometric authentication.

DLM was a powerful tool for feature extraction, data analysis, and target classification, which could efficiently recognize the speech signals, such as semantic recognition, bio-certification, and emotion analysis. Different from human hearing and traditional speech processing techniques, DLM exhibited a robust analytical capability for the recognizable features within speech signals, even in cases of speech signals with feature deficiency and waveform distortion. This capability was highly compatible with the Anti-noise TEAS, which focused on the detection of laryngeal mixed-mode acoustic signals. Thus, a CNN-based DLM with recognition capabilities was established to recognize the acoustic signals detected by the Anti-noise TEAS. Specifically, a CNN-based DLM for multi-label ASR was constructed to identify the semantics and individuals of various acoustic signals recorded by the Anti-noise TEAS. In addition, the recognition performance of the DLM for multi-label ASR was evaluated by various indicators, such as the confusion matrix, t-SNE cluster, and so on. The gradient-weighted class activation mapping (Grad-CAM) was also employed to analyze the activated regions of different spectrograms across 3 participants.

A total of 1244 samples (704 for the train dataset and 540 for the test dataset) of 6 different types of acoustic signals from 3 participants were collected using the Anti-noise TEAS, and followed by preprocessing on the raw data. Specifically, to ensure consistency among the different acoustic signals in the time dimension, the acoustic signals of various lengths were unified into 3.2 s through truncation or padding. Additionally, all signals were normalized to mitigate undesirable effects stemming from anomalous samples during the model training process. A feature extraction approach similar to Mel-frequency cepstral coefficients (MFCCs) was employed to transform the acoustic signals into spectrogram features, which were more suitable for DLM training. As a time-varying feature of acoustic signals, MFCCs have been widely used in speech recognition, speech synthesis, etc. However, the MFCCs transformation of acoustic signals required a discrete cosine transform (DCT), which projected spectral energy onto a new orthogonal basis that might not fully maintain the relative localization of the original signals[25]. In this work, to address the limitation of the discretization, the DCT would not be performed during the MFCC transformation. After taking the logarithm of the spectrum, the Mel-spectrogram was utilized as the spectrogram features of the acoustic signals directly. Consequently, by employing a window of 800 ms and a step of 200 ms, a 128×128-dimensional Log-Mel-spectrogram was obtained by segmenting the speech waveform signals into frames, applying windowing, performing STFT, Mel-filtering, and logarithmic operations (Supplementary Fig. S11).

The DLM for multi-label ASR consisted of convolution layers, max-pooling layers, drop-out layers, flatten layers, fully-connected layers, and sigmoid layers (Fig. 4a). By feeding the Log-Mel-spectrograms into the DLM for multi-label ASR, it could concurrently recognize semantic and individual information within the acoustic signals. In the DLM, it consisted of four convolution layers with 64, 128, 256, and 512 channels, respectively. The 3×3 convolution kernels, batch normalization, and activation functions of each convolution layer were employed to extract and preserve the feature spectrograms of the Anti-noise TEAS, allowing for the extraction and retention of deep nonlinear acoustic features from shallow to deep layers. The Max-pooling layers, which are situated between the convolution layers, employed a 2×2 filter to execute a down-sampling operation on the feature spectrograms generated from the last convolution layers, thereby diminishing the model parameters. The dropout layer with a coefficient of 0.25 was employed to randomly discard some neuron outputs of the model, which could further prevent overfitting. Then, the multi-dimensional feature spectrograms from the convolution layers were flattened into a one-dimensional feature sequence by the flatten layer and fed into the fully-connected layers. The fully-connected layers consisted of a three-layer structure with node numbers of 4096, 1024, and 128, respectively. This architecture formed a fully-connected neural network capable of extracting and reducing the dimensionality of the feature sequence, while retaining the most significant feature components within the acoustic signals. Similarly, between each fully-connected layer, some neuron nodes were randomly dropped out with a dropout layer (0.25), to enhance the final generalization ability of the DLM. The sigmoid activation function was employed to facilitate the multi-label classification of different signals, enabling the identification of various semantic contents, individual objects, and other diversified label classifications simultaneously (Supplementary Fig. S12). The detailed parameters of the DLM for multi-label ASR were exhibited in Supplementary Table S1.

To mitigate the impact of different sample partitioning schemes, the 10-fold cross-validation method was initially employed to perform a pre-validation and hyperparameter selection of the model structure (Supplementary Fig. S13). Subsequently, the training of the DLM for multi-label ASR was performed using all samples in the training dataset. As the training epochs increased, both the accuracy and loss function values converged rapidly (Fig. 4b). Specifically, the accuracy of the model presented a rapid boost as the epochs increased, stabilizing only after ~10 epochs, and consistently remaining above 99% overall. Conversely, the loss function value decreased rapidly, stabilizing after ~10 epochs and remaining below 0.05 overall. These observations indicated that the DLM for multi-label ASR exhibited a desirable convergence speed during the training process, which might be attributed to its structure design and hyperparameter selection. In addition, this also demonstrated that the acoustic signals of the human larynx detected with the Anti-noise TEAS contained a large number of semantic and individual features. Those features could be effectively extracted and retained throughout the training process of the DLM, ultimately enhancing the efficiency and speed of feature extraction.

To validate the feasibility of the DLM for multi-label ASR in acoustic recognition tasks, the samples of the test dataset were fed

https://doi.org/10.1038/s41467-025-59523-6

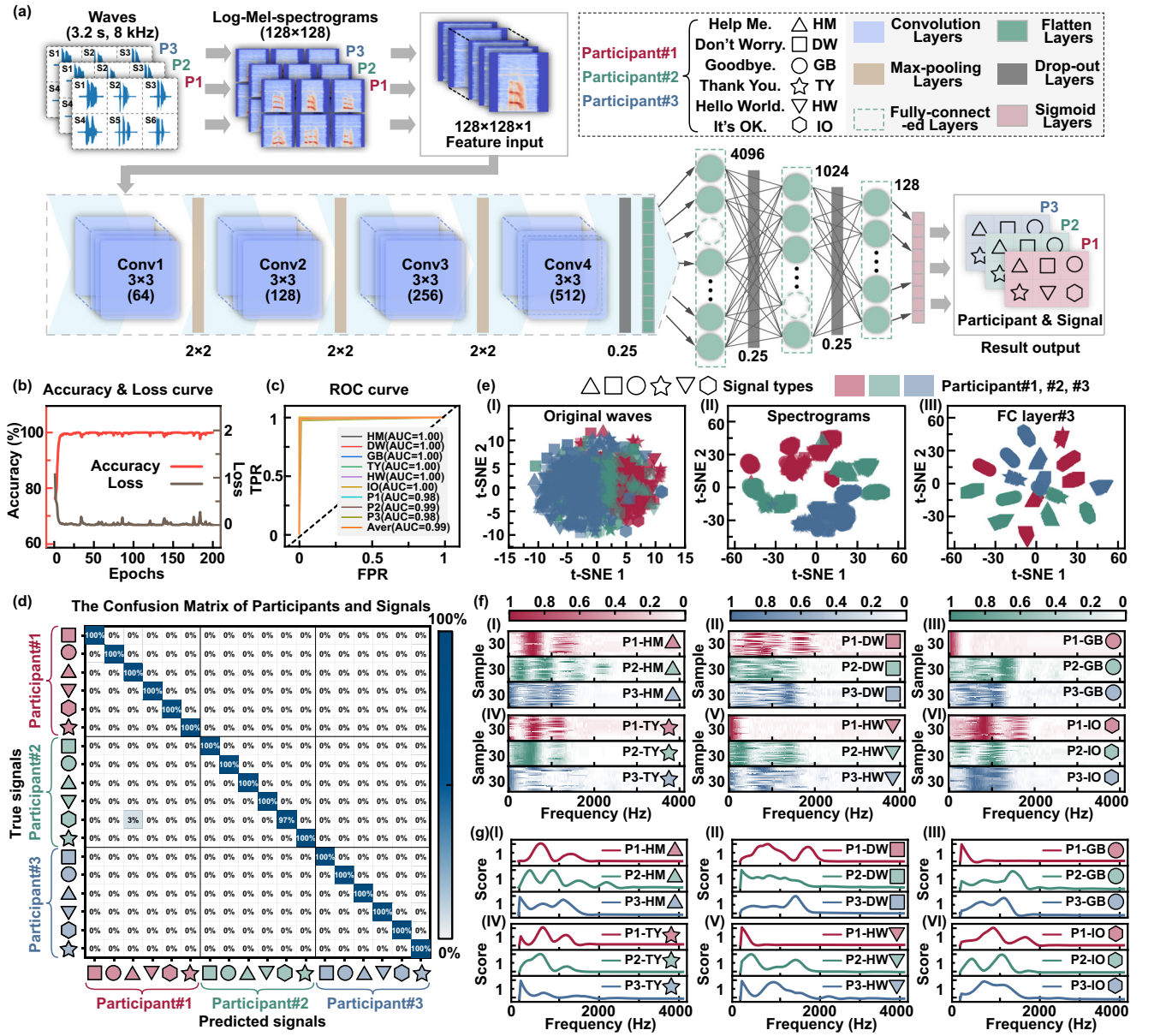


Fig. 4 | The CNN-based DLM for multi-label ASR. a Schematic of the structure of the CNN-based DLM for multi-label ASR. The original acoustic signals of the Anti-noise TEAS were truncated, filled, normalized, and feature extracted into 128 × 128-dimensional Log-Mel-Spectrograms. The DLM for multi-label ASR consisted of convolution layers, max-pooling layers, drop-out layers, flatten layers, fully-connected layers, and sigmoid layers. By inputting the Log-Mel-Spectrograms of 6 different acoustic signals from three different participants into the model, it was possible to recognize the semantic content of the speech signals and individual objects simultaneously. **b** Curves of the relationship between accuracy and loss function value vs the number of epochs of the DLM for multi-label ASR during training. **c** receiver operating characteristic (ROC) curves and area under the ROC curve (AUC) for classification of six different acoustic signals and three participants using the DLM for multi-label ASR. **d** Confusion matrix results using the DLM for multi-label ASR for semantic and individual recognition. Six different symbols represented different acoustic content categories, and three different colors represented different participants. **e** T-SNE of (I) original speech waveforms, (II) Log-Mel-Spectrum feature maps, and (III) output feature values of the third layer of the fully-connected layer. Among them, different symbols represented different semantic content categories, and different colors represented different participants. The sample size was 1244. **f** Heatmaps of gradient-weighted class activation mapping (Grad-CAM) attention score distribution vs frequency for the DLM across (I–IV) Six different semantic content categories, while three different colors represented three different participants. **g** Curves of Grad-CAM attention score distribution vs frequency for the DLM across (I–IV) Six different semantic content categories, while three different colors represented three different participants.

into the model to predict the multi-label classification. Subsequently, receiver operating characteristic (ROC) curves (Fig. 4c) and confusion matrices (Fig. 4d) were performed. The confusion matrix exhibited that the DLM for multi-label ASR possessed high recognition accuracy, with 99.8% accuracy for semantic recognition, 99.8% for individual recognition, and 99.8% overall accuracy. The model predicted the semantic and individual label of the samples separately, rather than directly predicting the semantic + individual label of the samples. This multi-label classification approach was more suitable for the classification of information-rich samples such as acoustic signals, which enabled a more comprehensive characterization of acoustic signals, including semantic information, individual, and emotion. Three separate DLMs for ASR were also established using data from each participant. The accuracy curves, loss curves, and confusion matrices during the training process of the models were presented in Supplementary Figs. S14–S16, while the detailed parameters of the

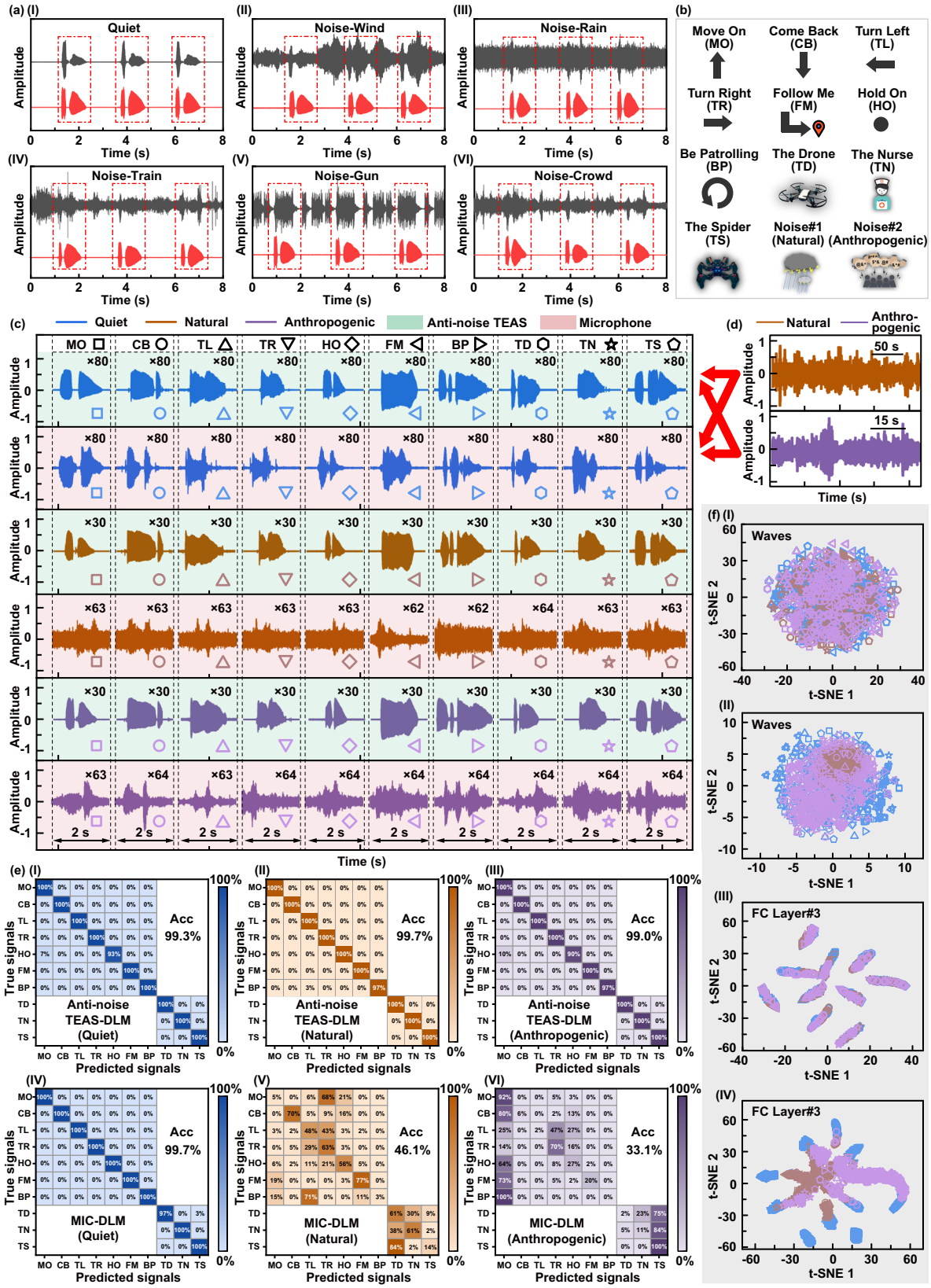
models could be found in Supplementary Tables S2–S4. The curves exhibited fast convergence, with an overall accuracy close to 100%, and the loss curve steadily ~0. The confusion matrices indicated that the acoustic recognition accuracy of the DLMs for each participant reached 98.3%, 100%, and 99.2%, respectively, which demonstrated the effectiveness of acoustic signal classification of individual models. To further evaluate the recognition performance of the DLM for multi-label ASR, the ROC curves for nine label classifications were plotted with true positive rate (TPR) and false positive rate (FPR) from the prediction results. Additionally, the area under the ROC curve (AUC) values was also calculated for each ASR. From the ROC curves, the AUC values of the model were close to or equal to 1 on both the semantic and the individual classification tasks, with an average AUC value of 0.97, which indicated the excellent classification performance of the DLM for multi-label ASR.

To visualize the extraction capability of the DLM for multi-label ASR on critical features of the acoustic signals, the t-SNE cluster algorithm was used to perform clustering on the original acoustic waveforms, feature spectrograms, and output feature queues of the third fully-connected layer, respectively. In the t-SNE cluster maps, various symbols represented six different semantic contents, while distinct colors denoted three different participants. For the original acoustic waveform signals (Fig. 4e-I), the samples from different semantic contents and participants were intermingled, which exhibited minimal demarcation and clustering. Notably, the silhouette coefficient score of original acoustic waveform signals was merely −0.1862, indicating a lack of clear separation and clustering within the samples (Supplementary Fig. S17). For the Log-Mel-spectrograms (Fig. 4e-II), the samples from different participants presented obvious separation and clustering, with an improved silhouette coefficient score of 0.5382. However, there were still obvious regional overlaps between samples of different semantic contents, and the clustering effect was not sufficient to distinguish different types of samples effectively. Furthermore, after feeding the Log-Mel-spectrogram into the model, the output feature values of the third fully-connected layer were intercepted and performed t-SNE clustering operations. The clustering efficacy of the output feature values was notable, and there were clearer clustering patterns among different semantic content and distinct participants (Fig. 4e-III). Minimal overlap was observed between acoustic samples from different categories, resulting in a further improvement of the silhouette coefficient to 0.7841. During the operation of the DLM for multi-label ASR, the convolutional and fully-connected layers were capable of extracting and retaining the most essential features in the acoustic signals of the Anti-noise TEAS efficiently. This capability would facilitate the extraction of the recognizable features within acoustic signals, which was helpful in compensating for the lack of semantic features in the acoustic signals.

DLMs have exhibited desirable performance in ASR owing to its robust computational capabilities, yet the interpretability of deep learning operations remains a challenge. The internal mechanisms and decision-making processes of DLM were often opaque, making it difficult to discern how the model precisely reached a recognition conclusion. To address this issue, the Grad-CAM, as a neural network visualization technique, has undergone continuous development and garnered widespread acceptance in deep learning techniques. The Grad-CAM aided in deciphering the decision inference process of the DLM, which could enhance its interpretability, while preserving the architecture of the model intact without compromising its structure or accuracy. In this work, the acoustic signals of test dataset were fed into the DLM for multi-label ASR, and utilized the Grad-CAM to reverse-engineer the decision-making outcomes of the model, aiming to assess the attention scores of the input feature spectrogram in different regions. The representative Grad-CAM heatmaps of 6 different acoustic signals from 3 different participants were exhibited in Supplementary Figs. S18–S20. In the normalized heatmap of the Grad-CAM, regions with higher scores indicated that the model allocated greater attention to the features in those areas, which signified that the important information embedded within those regions contributed to the recognition of acoustic signals. Conversely, regions with lower scores denoted that the model disregarded redundant and irrelevant information within those regions during the decision-making process. After interpolation, superimposition, summation, and normalization, the average heatmaps of the Grad-CAM for different categories of acoustic signals were obtained (Supplementary Figs. S21–S23). By comparing the average heatmaps of the Grad-CAM with the input Log-Mel-spectrograms, the model's attention scores were concentrated in the low and mid-frequency region (below 1500 Hz). This observation indicated that these frequency ranges contained significant features crucial for semantic and individual recognition, which were consistent with the frequency ranges of the fundamental acoustic signals detected by the Anti-noise TEAS. Furthermore, the heatmaps of the Grad-CAM were performed average operation along the time dimension, and retained only the compressed one-dimensional heatmaps of the average score value across frequencies. Subsequently, these compressed one-dimensional heatmaps from all samples were stacked to generate a frequency heatmap, which depicted the distribution of the Grad-CAM across frequencies (Supplementary Fig. S24). Figure 4f exhibited the frequency-heatmaps of the Grad-CAM for the 6 acoustic signals from 3 participants (different colors). Darker colors within the heatmap signified higher attention score values of the corresponding frequency regions, indicating the model's dependence on these frequencies in the decision-making process. While the focused frequency regions of the model varied across different acoustic signals, the regions with the highest scores predominantly fall within the middle and low frequency ranges of the Log-Mel-spectrograms. In addition, the frequency-heatmaps of the Grad-CAM with the same type of samples (30 samples) were averaged and summed, which obtained the curves that indicate the relationship between the model's attention score values and frequency distribution. The region of highest attention scores for the model varied for different categories of acoustic signals (Fig. 4g). The highest attention scores for most categories of acoustic signals were located below 1000 Hz, while a few other categories were located in the range of 1000–1500 Hz. This demonstrated that, during ASR, the features of highest concern to DLM for multi-label ASR were mainly concentrated in the middle and low frequency ranges below 1500 Hz. Such findings further confirmed the efficacy of the Anti-noise TEAS in capturing fundamental acoustic signals. The Anti-noise TEAS was able to adequately record most of the key information in the speech signals, and had sufficient differentiation of acoustic characteristics, such as different semantic contents and individual objects.

A conventional microphone encountered significant interference in noisy environments, which would hinder efficient and accurate HMVI. Although filtering algorithms and traditional speech processing techniques could solve the problem of noise interference to a certain extent, their efficacy largely depended on the prior recognition of noise categories (i.e., the availability of noise samples beforehand) to optimize signal effectiveness. In the case of unknown noisy environments, traditional speech processing techniques often failed to effectively distinguish and recognize the acoustic signals. Due to the excellent anti-noise capability, the Anti-noise TEAS could collect fundamental acoustic signals through contact sensing in noisy environments, which could be combined with DLM to achieve semantic recognition and intelligent interaction. With the optimal structure and contact sensing, the Anti-noise TEAS naturally mitigated noise interference during the sensing of mixed-mode acoustic signals. To evaluate the anti-noise capability of the Anti-noise TEAS, the sensor was employed to detect the acoustic signal (help me) in various environments, such as quiet, windy, crowded, and compared with a commercial microphone. Across various noise scenarios, the acoustic signal waveforms recorded by the Anti-noise TEAS

https://doi.org/10.1038/s41467-025-59523-6



exhibited remarkable consistency with those recorded in quiet environment (Fig. 5a). Conversely, the acoustic signals detected by the microphone were substantially obscured by ambient noise, which made the contour characteristics of the acoustic signals indistinguishable. The acoustic signals were transformed into spectrograms via STFT. Similarly, the spectrograms obtained by the Anti-noise TEAS exhibited higher consistency with those from the quiet environment, while those from the microphone were significantly impacted by ambient noise, which manifested pronounced background noise and distortion (Supplementary Fig. S25). Therefore, compared with the microphone, the Anti-noise TEAS could effectively detect the acoustic signals in noisy environments.

Fig. 5 | Anti-noise and speech recognition performance of the Anti-noise TEAS-DLM. a Acoustic signals of help me in (I) a quiet environment, and in different noisy environments, such as (II) gusty wind, (III) heavy rain, (IV) train noise, (V) gunshot noise, and (VI) crowd noise. The red curves were the acoustic signals recorded by the Anti-noise TEAS, and the black curves were the acoustic signals captured by the microphone. **b** Symbols of the seven action commands, three object commands, and two noise environments involved in the human-machine collaborative tasks. **c** Waveforms of the seven action commands, three object commands captured by the Anti-noise TEAS and microphone in quiet, natural noise, and anthropogenic noise environments, respectively. Among them, different symbols represented different command signals, while different color backgrounds represented the signals of the Anti-noise TEAS and the microphone, respectively. **d** Waveforms of natural noise and anthropogenic noise. **e** Confusion matrices of command signal recognition results of the Anti-noise TEAS-DLM in (I) quiet, (II) natural noise, and (III) anthropogenic noise environment. Confusion matrices of command signal recognition results of the MIC-DLM in (IV) quiet, (V) natural noise, and (VI) anthropogenic noise environment. **f** T-SNE cluster plots of (I–II) original waveforms and (III–IV) output fully-connected layers of the Anti-noise TEAS-DLM (sample size is 1400) and the MIC-DLM (sample size is 2067). Among them, different symbols represented various command categories, and different colors represented various environments.

The advantages of the Anti-noise TEAS for the acoustic signal detection in noisy environments were discussed comprehensively in Supplementary Note S4. In specific, the Anti-noise TEAS was optimized through the structure design and parameter selection to achieve physical-denoising, which was versatile for buffering unknown types of noise in open and harsh noise scenarios. In contrast, algorithmic denoising or AI-model denoising remained challenging in variable open scenarios to solve the unknown noise interference, and imposed certain requirements on the computing capacity of hardware devices. Second, as a flexible contact acoustic sensor, the Anti-noise TEAS could directly acquire the acoustic signals from the mixed-mode weak vibration of the larynx. The Anti-noise TEAS could selectively buffer the interference of the noisy environments, while remaining sensitive to record target acoustic signals from the larynx, due to the anisotropic sensing characteristics of the Anti-noise TEAS. Conventional rigid microphones were mainly designed to detect airborne acoustic signals, where the transduction of the target signal was not inherently different from the noise signals, and thus the noise could not be selectively buffered. Moreover, the Anti-noise TEAS was also available for patients with laryngeal dysfunction, by detecting the vibrations of their laryngeal muscles. Third, the Anti-noise TEAS possessed a unique air-layer structure, where the middle air layer was compressed to change the distance between the two friction layers during the contact sensing. Due to the low modulus of elasticity of the air layer, the inner diaphragm was able to highly compress it during the deformation process, and generate friction initiation and electrostatic induction with the outer friction layer, thus responding to the acoustic signals sensitively. The interference vibrations of external ambient noises mainly affected the deformation of the outer friction layer and the compression of the middle air layer. The Anti-noise TEAS could utilize the inner diaphragm to sense the acoustic vibrations of the larynx, while the outer friction layer isolated the interference of ambient noise. The specific selection and sensing of target throat signals and external noise signals allowed the sensor to resist the noise interference without reducing the detection sensitivity. In addition, due to the low modulus of elasticity of the middle air layer, the inner diaphragm and outer friction layer were in low-resistance free deformation states. Thus, the Anti-noise TEAS demonstrated a high degree of sensitivity to weak acoustic signals of laryngeal vibrations, merely by being comfortably attached to the surface of the larynx. In contrast, a conventional flexible PVDF acoustic sensor recorded a vibration signal from the larynx through the compressive deformation-induced polarization of the PVDF layer. Differing from the Anti-noise TEAS, PVDF piezoelectric acoustic sensors did not possess the anisotropic sensing characteristics. With the adjustment of the PVDF layer, the detection sensitivity of the PVDF acoustic sensor and the resistance to ambient noise were simultaneously diminished. The PVDF piezoelectric sensors also required a high pre-pressure to be attached tightly to the throat surface to achieve sufficient PVDF material deformation for acoustic sensing. Conventional flexible piezoresistive acoustic sensors mainly rely on external pressure to induce a change in contact resistance, typically exhibiting lower dynamic mechanical detection sensitivity. The piezoresistive acoustic sensors also did not possess anisotropic sensing characteristics, which was infeasible for selectively buffering noise.

The feasibility and accuracy of the Anti-noise TEAS-DLM were also validated by a complex human-machine collaborative tasks in noisy environments. In the complex human-machine collaboration (Fig. 5b), there were seven types of action commands, including move on (MO), come back (CB), turn left (TL), turn right (TR), follow me (FM), hold on (HO), be patrolling (BP), and three object commands, comprised the drone (TD), the nurse (TN), the spider (TS). There were three types of different environmental sounds in complex scenarios, including quiet, natural noise (such as rain, thunder), and anthropogenic noise (containing crowd sounds, or gunshot sounds) environment. These ten types of command signals were collected by the Anti-noise TEAS or microphone in different environments, respectively. Specifically, the Anti-noise TEAS recorded 800 sets of command signals in the quiet environment, and 300 sets each in the natural and anthropogenic noise environments, respectively. The microphone detected 800 sets of command signals in the quiet environment, 629 sets in the natural noise environment, and 638 sets in the anthropogenic noise environment. Figure 5c exhibited the representative command signals obtained by both the Anti-noise TEAS and the microphone across three distinct environments. The waveform contours of the command signals recorded by the Anti-noise TEAS exhibited notable consistency and repeatability across different environments, which were less affected by the interference of environmental noise. Conversely, the waveform contours of the command signals detected by the microphones in quiet and noisy environments displayed considerable variability, which were significantly interfered with and distorted. Meanwhile, as shown in the spectrograms (Supplementary Figs. S26 and S27), the command signals acquired by the Anti-noise TEAS in different environments possessed better consistency and repeatability than those of the microphone.

To implement complex human-machine collaborative tasks utilizing the Anti-noise TEAS-DLM in noisy environments, a CNN-based DLM for ASR was constructed. Analogous to the construction of the DLM for multi-label ASR, the command signals were truncated, padded, and normalized to eliminate undesirable effects due to anomalous samples during the training process. After frame-splitting, windowing, STFT, Mel filtering, and logarithmic operations, the command signals were converted into 128×128-dimensional Log-Mel-spectrograms, which were more suitable for DLM training. Furthermore, the command samples collected by the Anti-noise TEAS in the quiet environment were divided into a train dataset and a testing dataset following a 5:3 ratio. Specifically, the train dataset contained 10 types of command samples, with 50 sets of each signal type, totaling 500 sets of samples. The test dataset included 10 types of command samples, with 30 sets of each signal type, totaling 300 sets of samples. The samples of the train dataset were utilized to train the CNN-based DLM (Supplementary Fig. S28), thereby obtaining the DLM for ASR. Notably, in this work, the action and object commands were specified as two different types of signals. Thus, the classification outputs of the model were specified accordingly, i.e., action commands could only be recognized as action labels, while object commands could only be

recognized as object labels (Supplementary Figs. S29 and S30). To compare the performance of the Anti-noise TEAS-DLM, the microphone coupled with the DLM for ASR (MIC-DLM) was also constructed. The model was trained utilizing the training dataset of acoustic signal samples collected by the microphone in a quiet environment (Supplementary Fig. S31). The detailed parameters of the Anti-noise TEAS-DLM and MIC-DLM models were exhibited in Supplementary Tables S5 and S6.

The test dataset of the acoustic signals recorded by the Anti-noise TEAS and the microphone in different environments was fed into the DLM, respectively. The recognition results of these acoustic signals were obtained, and separately plotted the corresponding confusion matrices (Fig. 5e). In the quiet environment, both the Anti-noise TEAS-DLM and the MIC-DLM demonstrated remarkable speech recognition accuracy, which possessed an overall command recognition accuracy of 99.3% and 99.7%, respectively. In the natural and anthropogenic noise environments, the accuracy of the Anti-noise TEAS-DLM remained unaffected, maintaining an overall accuracy of 99.7% and 99.0%. Conversely, the accuracy of the MIC-DLM notably declined due to noise interference, which decreased to 46.1% and 33.1%, respectively. Compared with the MIC-DLM, the Anti-noise TEAS-DLM preserved desirable command recognition efficacy across various environments. Such anti-noise capability was pivotal for facilitating human-machine collaborative tasks in noisy scenarios. Additionally, the ROC curves of the 10 command recognitions were performed for both the Anti-noise TEAS-DLM and the MIC-DLM across various environments, and the AUC values of each ROC curve were also calculated. For the Anti-noise TEAS-DLM (Supplementary Figs. S32–S34), the average AUC value of the object commands in a quiet environment was up to 1.00, while that of the action command was as high as 0.99. Similarly, the average AUC value in natural and anthropogenic noise environments was all close to 1.00, which indicated the desirable recognition performance of the Anti-noise TEAS-DLM. For the MIC-DLM (Supplementary Figs. S35–S37), the average AUC values of the object and the action commands in the quiet environment both reached 1.00. However, the average AUC value of the object and action commands in the natural noise environments was reduced to 0.59 and 0.65, while that in the anthropogenic noise environments was reduced to 0.60 and 0.58. These results indicated the inferior recognition performance of the MIC-DLM, which demonstrated its incapability to distinguish command signals in noisy environments effectively.

To visually exhibit the command recognition capability of the Anti-noise TEAS-DLM and the MIC-DLM, the t-SNE cluster maps were performed on various components, including original speech waveform signals, feature spectrograms, the fourth convolution layers, and the third fully-connected layers. The silhouette coefficients of different t-SNE cluster maps were also performed (Supplementary Figs. S38 and S39). In the t-SNE cluster maps, ten different symbols represented ten types of command signals, while three different colors corresponded to the three types of environments, respectively. As presented in Fig. 5f-I, II, the t-SNE cluster maps of original waveform signals, which were collected by the Anti-noise TEAS and the microphone, appeared largely mixed without distinct demarcation or clustering phenomena. The silhouette coefficients of the quiet, natural noise, and anthropogenic noise environments were −0.1244, −0.0715, −0.1204, and −0.1609, −0.1060, −0.1122, respectively. Upon transforming the original waveforms into Log-Mel-spectrograms, the t-SNE cluster maps of the Anti-noise TEAS-DLM partial separation and clustering (Supplementary Fig. S40). The silhouette coefficients were improved to 0.6062, 0.7238, and 0.6644, respectively, which demonstrated the effectiveness of feature extraction within the Anti-noise TEAS-DLM. After Log-Mel-spectrograms transformation, the t-SNE cluster maps of the MIC-DLM showed a clustering phenomenon of samples from different environments, but there was still overlap between samples of various command types (Supplementary Fig. S41). The silhouette coefficients were −0.0080, 0.2852, and 0.0134, respectively, which indicated that the feature extraction of the MIC-DLM could differentiate environmental information effectively, but not the command types. Further, the test datasets of the Anti-noise-DLM and the MIC-DLM were fed into the DLM, and the output values of the fourth convolution layer and the third fully-connected layer were intercepted to perform the t-SNE cluster operation. The t-SNE cluster maps of the fourth convolution layers of the Anti-noise TEAS-DLM exhibited a significant clustering effect, which could enhance the discrimination of command types across various environments. However, the clustering effect of the fourth convolution layers of the MIC-DLM showed marginal improvement in a quiet environment, while the clustering remained poor in natural or anthropogenic noise environments. Additionally, the t-SNE cluster maps of the third fully-connected layer of the Anti-noise TEAS-DLM exhibited robust clustering effects (Fig. 5f-III), which the silhouette coefficients improved to 0.7096, 0.6859, and 0.6686, respectively. The same type of command signals across three different environments were clustered together, indicating that the clustering effect of the Anti-noise TEAS-DLM was not affected by environmental noise. The t-SNE cluster maps of the third fully-connected layer of the MIC-DLM showed clearer clustering phenomena in the quiet environment (Fig. 5f-IV), with the silhouette coefficient improving to 0.7960. However, the clustering effect of MIC-DLM in noisy environments remained poor, which displayed significant overlap, with silhouette coefficients of only −0.0010 and −0.1149. This indicated that the microphone was susceptible to environmental noise in complex scenarios, which led to a notable absence of discernible signal features. Therefore, compared with the MIC-DLM, the Anti-noise TEAS-DLM possessed desirable noise immunity, which could effectively differentiate command signals in various noisy environments.

We then evaluated the feasibility and accuracy of the Anti-noise TEAS-DLM by performing the complex human-machine collaborative tasks in different harsh scenarios. Three different scenarios were designed in the virtual world, including an open field, a complex field#1, and a complex field#2, to test the feasibility of the Anti-noise TEAS-DLM in human-machine collaboration. In the open field, the objects were free to move around, while in complex field#1 and complex filed#2, the objects were required to perform rescue tasks. As the controller, the fireman leader was the subject of sending commands by voice in various scenarios, where the voice signals were detected by the Anti-noise TEAS. These acoustic signals were then recognized and parsed into the task's commands by the DLM. Subsequently, the commands were wirelessly transmitted via the communication device to three objects (drone, spider robot, and nurse), who collaborated to accomplish the tasks. In each field, there were three types of environments, including quiet, natural noise, and anthropogenic noise environments. The leader's acoustic signals were significantly affected by that noise, which made it challenging to accurately transmit the acoustic signal using a conventional microphone. With the action commands, the leader controlled different objects to execute specific human-machine collaborative tasks according to defined routes. As the control group, the command signals issued by the leader were simultaneously obtained by the microphone. These signals were then converted into command signals to be transmitted to the drone, robot and nurse via the MIC-DLM, which were significantly interfered with by the noisy environments. In order to simulate the complex interactions between multiple human and machine entities in real-life scenarios, the nurse (real person) also acted as an information-receiving entity. The nurse judged the types of command signals recorded by the microphone through hearing perception (MIC-Hearing), and performed the specific human-machine collaborative tasks according to defined routes (Supplementary Fig. S42). The recorded tracks of different objects in response to Anti-noise TEAS-DLM (Supplementary Fig. S43), the MIC-DLM

https://doi.org/10.1038/s41467-025-59523-6

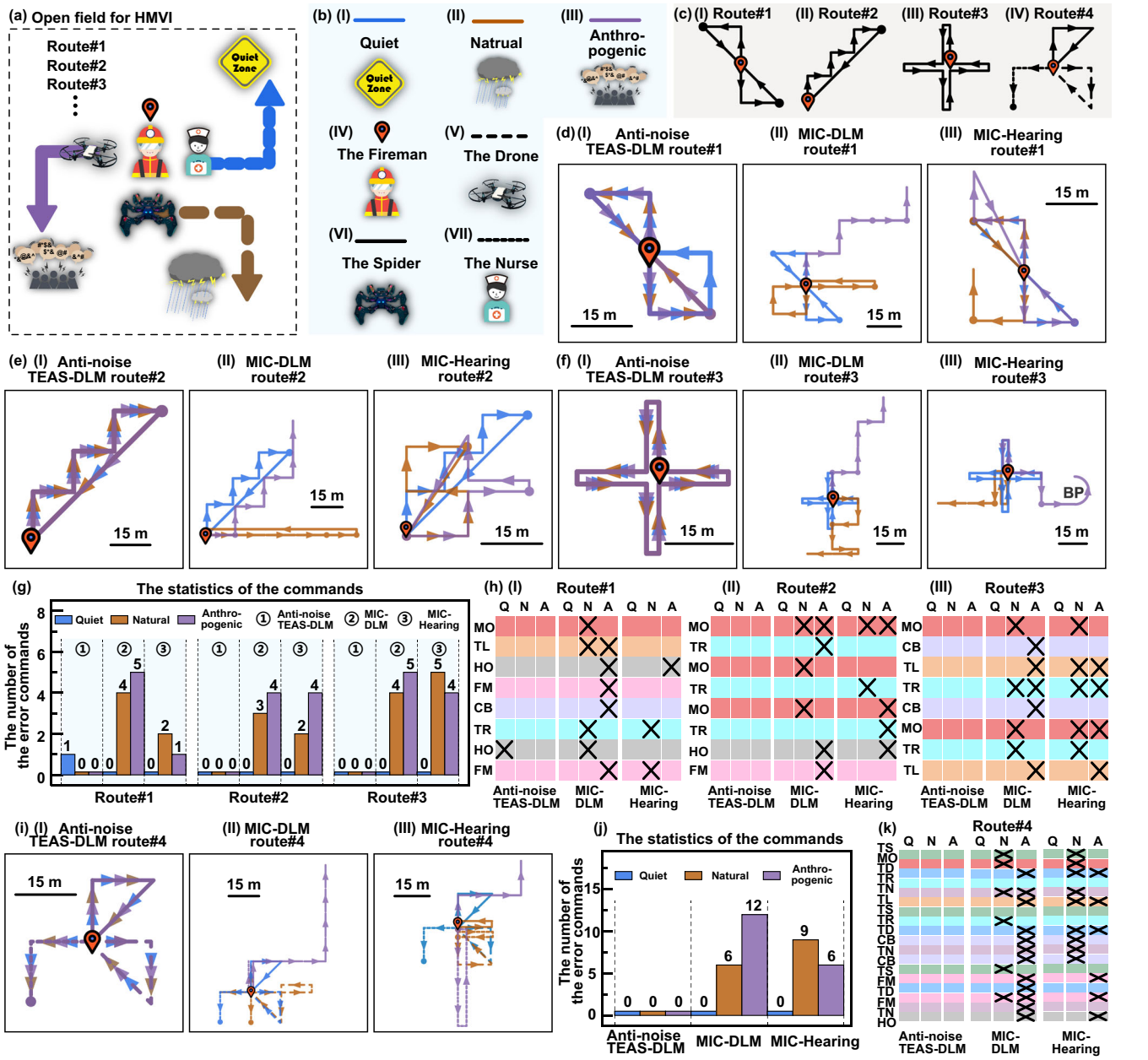


Fig. 6 | Performance of the Anti-noise TEAS for human-machine collaboration in virtual open fields. a Schematic diagram of controlling the robot, drone, and nurse to execute a specified route in the open field via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing. **b** Different colors, lines, and symbols were used to represent the environments, objects, and action tracks in the human-machine collaborative tasks. **c** Four kinds of theoretical routes executed in the open field, where (I) route#1, (II) route#2, and (III) route#3 were single-object HMVI tasks, while (IV) route#4 was a multi-objects HMVI task. **d** Recorded tracks of the route #1 executed in a quiet environment, a natural noise environment, and an anthropogenic noise environment via (I) the Anti-noise TEAS-DLM, (II) the MIC-DLM, and (III) the MIC-Hearing. **e** Recorded tracks of the route#2 executed in various environments via (I) the Anti-noise TEAS-DLM, (II) the MIC-DLM, and (III) the MIC-Hearing. **f** Recorded tracks of the route#3 executed in various environments via (I) the Anti-noise TEAS-DLM, (II) the MIC-DLM, and (III) the MIC-Hearing. **g** Statistical analysis of the number of erroneous commands for the execution of route#1, route#2, and route#3 via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing in various environments. **h** Color maps of the commands for the execution of (I) the route#1, (II) the route#2, and (III) the route#3 via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing in various environments. **i** Recorded tracks of the route#4 executed in various environments via (I) the Anti-noise TEAS-DLM, (II) the MIC-DLM, and (III) the MIC-Hearing. **j** Statistical analysis of the number of erroneous commands for the execution of route#4 via different HMVIs in various environments. **k** Color maps of the commands for the execution of route#4 via different HMVIs in various environments.

(Supplementary Fig. S44), and the MIC-Hearing (Supplementary Fig. S45) in various harsh scenarios were monitored, by counting the indicators of the total number of commands, the command accuracy rate, and the total time for task completion.

The accuracy of the three types of HMVIs in the open field was verified by employing combined command signals to control object actions. In the open field, the fireman leader would control the drone, robot, and nurse to move along specific routes, while the scene environment had no object collision barriers to their movements (Fig. 6a). To facilitate demonstration of object tracks within the tasks in the plots, the robot's recorded tracks were denoted with solid lines, the nurse's tracks were represented with dense dashed lines, and the drone's tracks were loose depicted with dashed lines (Fig. 6b). Additionally, blue, brown, and purple colors indicated the moving tracks in

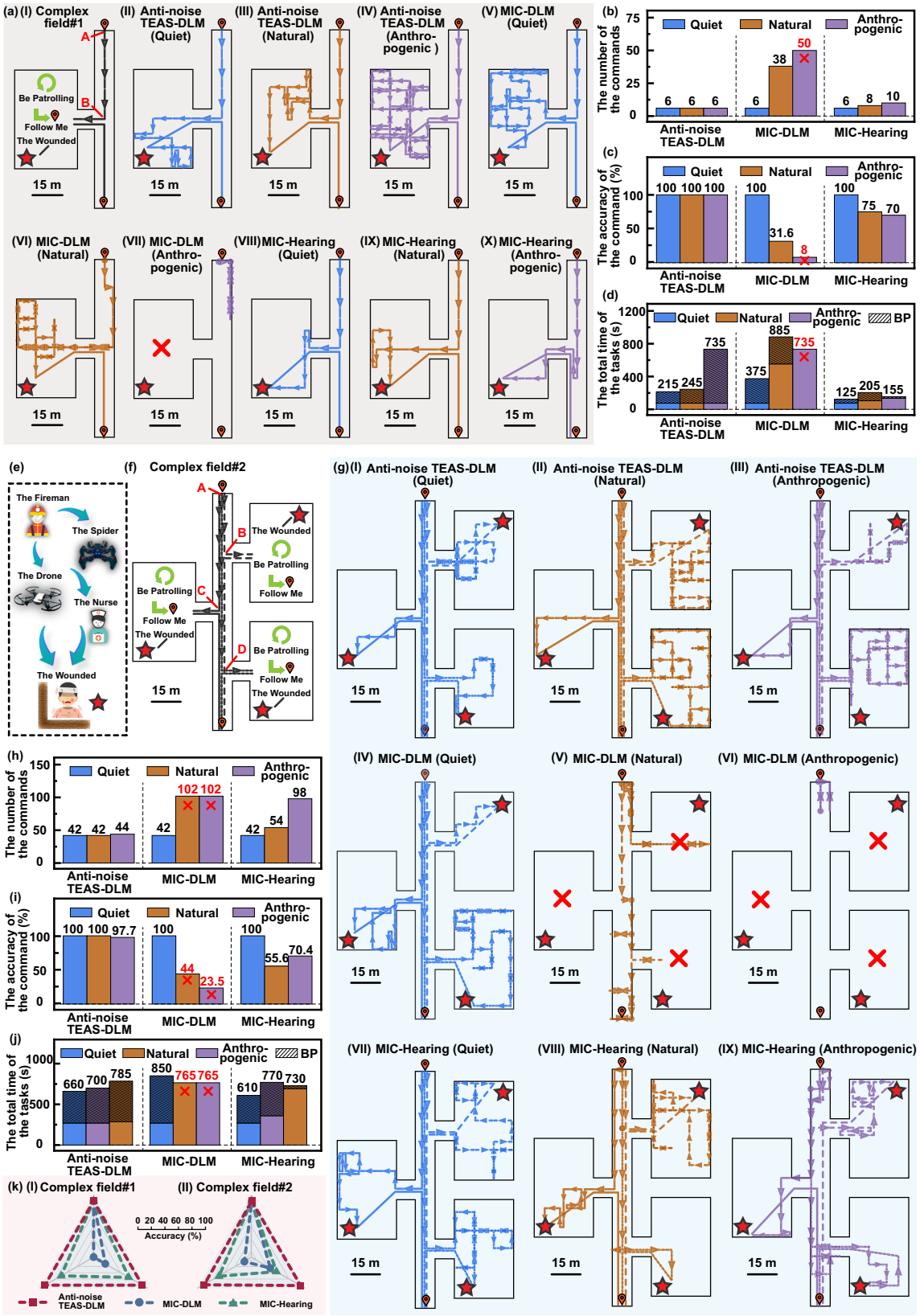
the quiet, natural noise, and anthropogenic noise environments, respectively, while the position of the leader was indicated by orange coordinates. Figures 6c-I–III illustrated the theoretical routes of three specific routes in the open field, where route#1 encompassed MO, TL, HO, FM, CB, TR, HO, and FM; route#2 followed MO, TR, MO, TR, MO, MO, TR, HO, and FM; and route#3 involved MO, CB, TL, TR, CB, MO, TR, and TL. Utilizing the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing, the route#1, the route#2, and the route#3 were executed in the open field (Supplementary Figs. S46–S48), respectively. In the route#1, the Anti-noise TEAS-DLM demonstrated commendable performance across various environments, with only 1 error occurring in a command intended for HO. However, while the MIC-DLM exhibited proficiency in quiet environments, its performance varied significantly in the noisy scenarios, leading to significant deviations in recorded tracks and an increased occurrence of command errors. Similarly, the MIC-Hearing performed effectively in quiet environments, but some command errors were observed in the natural and anthropogenic noise environments, resulting in some deviations of the recorded tracks from the theoretical routes. In the route#2 (Fig. 6e) and route#3 (Fig. 6f), the Anti-noise TEAS-DLM could also execute the tasks well in various environments, and the recorded tracks of the tasks were consistent with the theoretical routes. Conversely, the MIC-DLM and the MIC-Hearing successfully executed the specified routes only in quiet environments, but there were more command errors in the two noisy environments, resulting in a large deviation of the recorded tracks from the theoretical routes. The number of incorrect commands of 3 routes across different environments were counted. As exhibited in Fig. 6g, the Anti-noise TEAS-DLM had only 1 (1.4%) command error in the route#1, while MIC-DLM and MIC-Hearing possessed multiple command errors in noisy environments, which totaled up to 25 (34.7%) and 18 (25%) errors, respectively. To provide a more intuitive representation of the performance of these HMVI approaches across varied environments, the command executions for the three routes were presented with color maps (Fig. 6h). Different colors denoted various command categories, while error symbols above the colors indicated error levels in command execution. The Anti-noise TEAS-DLM demonstrated desirable performance across diverse environments and routes, whereas the MIC-DLM and the MIC-Hearing exhibited better execution proficiency solely in quiet environment, with a higher occurrence of command errors in noisy environments.

Subsequently, the performance of these HMVIs were further assessed by the collaborative tasks of simultaneously controlling multiple objects within an open field (Supplementary Fig. S49). As illustrated in Fig. 6b-IV, the fireman leader simultaneously controlled multiple objects to move along the theoretical route#4, via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing, respectively. The robot was required to execute MO, TL, and FM, the drone was controlled to execute TR, CB, and FM, and the nurse was ordered to execute TL, CB, and HO. The fireman leader must specify the object to be controlled with the object command signals before issuing the action command each time. For instance, when controlling the robot to move forward, the command signals TS + MO were issued, and when directed the nurse to turn left, the command TN + TL were issued. Figure 6i-I–III exhibited the recorded track diagrams of the route#4 via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing, respectively. Across the three various environments, the recorded tracks of the Anti-noise TEAS-DLM exhibited a high level of consistency with the theoretical route#4. However, the recorded tracks of the MIC-DLM and the MIC-Hearing aligned with the theoretical route#4 solely in the quiet environment, while the recorded tracks deviated from the theoretical route in both natural and anthropogenic noise environments. Additionally, the erroneous commands of the these three HMVIs during the execution of route #4 were counted (Fig. 6j). The Anti-noise TEAS-DLM recorded zero erroneous commands while performing the multi-objects task across diverse environments. Conversely, the MIC-DLM and the MIC-Hearing experienced numerous command errors in noisy environments, totaling 18 (75%) and 15 (62.5%) instances. The Anti-noise TEAS-DLM demonstrated desirable performance across various environments, whereas the MIC-DLM and the MIC-Hearing showed better execution solely in quiet environment, and displayed more errors in noisy environments (Fig. 6k). Thus, compared to the MIC-DLM and the MIC-Hearing, the Anti-noise TEAS-DLM possessed better performance on multi-object human-machine collaboration in harsh noisy environments.

Furthermore, the performance of the Anti-noise TEAS was also evaluated by the human-machine collaborations within complex field#1 and complex field#2. The complex field#1 was simulated in a scenario where a fireman leader utilized the HMVI to control the robot for a wounded rescue task in a disaster environment. As illustrated in Supplementary Fig. S50, the complex field#1 was simulated in a scenario, where a fireman leader controlled the robot for a wounded rescue task in a disaster environment. In complex field#1, the fireman directed the robot from the starting point (point A) to the entrance of the blank area (point B) at a speed of 1 m/s. Subsequently, the robot was required to enter the blank area and execute a BP command to search for the wounded (indicated by a pentagram). Upon locating the wounded, the leader would control the robot to return to point B by an FM command, and the task of rescuing was accomplished. During the execution of the rescue tasks, if the robot deviated from the theoretical route due to incorrect action commands, the leader should provide the correct commands based on the recorded situations. This provision ensured the successful implementation of the wounded rescue task. If the wounded rescue task was unable to be completed within 50 commands, the task was judged to have failed, and the subsequent actions were terminated. Figure 7a-I exhibited the schematic diagram of the theoretical route of controlling the robot to realize the wounded rescue tasks in complex field#1 through different HMVIs.

The leader directed the robot to execute the rescue tasks in various environments via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing (Supplementary Fig. S51). As shown in Fig. 7a-II–IV, the Anti-noise TEAS-DLM effectively executed the wounded rescue tasks across various environments to accomplish wounded rescue tasks. While the MIC-DLM successfully executed tasks in quiet environments, significant interference from noise environments led to numerous command errors, which necessitated repeated corrections by the leader and even resulted in task failure in anthropogenic noise environments. Similarly, although MIC-Hearing successfully completed the tasks, more obvious command execution errors occurred in noisy environments, which also required the leader to continually adjust the robot's actions. The number of commands, the accuracy of commands, and the overall time of the rescue tasks performed by these HMVIs in different environments were summarized. The Anti-noise TEAS-DLM required a total of 6 commands to complete the rescue tasks in various environments (Fig. 7b), and the accuracy of the commands were all 100% (Fig. 7c). The overall time for the rescue tasks varied, being 215 s, 245 s, and 735 s, respectively (Fig. 7d). Discrepancies in the overall time of the rescue tasks were primarily due to the differences in the length of randomized patrols during the execution of BP commands, whereas the actual duration time caused by other commands was 75 s across all scenarios. The MIC-DLM required a total of six commands (100%) for the rescue task in the quiet environment, with an overall time of 375 s (actual time 75 s and BP time 300 s). However, in the natural noise environment, the MIC-DLM required 38 commands (31.6%), with an overall time of 885 s (555 s + 330 s). In the anthropogenic noise environment, despite issuing over 50 commands, the MIC-DLM failed to effectively accomplish the rescue tasks with a low accuracy rate of 8%, which resulted in rescue task failure. Meanwhile, the MIC-Hearing executed 6 commands

https://doi.org/10.1038/s41467-025-59523-6



in the quiet environment (100%), 8 commands in the natural noise environment (75%), and 10 commands (70%) in the anthropogenic noise environment. The overall rescue times were 125 s (75 s + 50 s), 205 s (105 s + 100 s), and 155 s (135 s + 20 s), respectively. The Anti-noise TEAS-DLM effectively resisted extreme noise interference in various noisy scenarios, which facilitated accurate and efficient human-machine collaborative tasks. Conversely, the MIC-DLM and the MIC-Hearing were highly susceptible to noise interference, which made them unsuitable for complex human-machine collaborative tasks in noisy scenarios.

Furthermore, the Anti-noise TEAS-DLM was also validated by complex multi-object human-machine collaboration in complex

Fig. 7 | Performance of the Anti-noise TEAS for multi-object human-machine collaboration in virtual complex fields. a Schematic diagrams of wounded rescue tasks in the complex field#1 with the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing across various environments. Among them, (I) was the theoretical route of the complex field#1, (II–IV) were the recorded tracks of the Anti-noise TEAS-DLM, (V–VII) were the recorded tracks of the MIC-DLM, and (VIII–X) were the recorded tracks of the MIC-Hearing. Point A represented the starting point of the tasks, while point B represented the entrance to the blank areas. Different colors, lines, and symbols were used to represent the environments, the objects, and the action tracks in the human-machine collaborative tasks. Statistical analysis of (**b**) the number of commands, **c** the accuracy of commands, and **d** the overall time of the wounded rescue tasks utilizing the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing in the complex field#1. **e** Illustration of the fireman leader controlling the robot, drone and nurse to perform wounded rescue tasks in complex field#2 via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing across various environments. **f** Schematic diagram of the theoretical route for performing a wounded rescue task in complex field#2 via multi-object collaboration. Point A represented the starting point of the tasks, while points B, C, and D represented the entrance to the blank areas. **g** Recorded tracks of the wounded rescue tasks in complex field#2 using the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing in various environments. Among them, (I–III) were the recorded tracks of the Anti-noise TEAS-DLM, (IV–VI) were the recorded tracks of the MIC-DLM, and (VII–IX) were the recorded tracks of the MIC-Hearing. Statistical analysis of (**h**) the number of commands, **i** the accuracy of commands, and **j** the overall time of the rescue tasks using the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing in complex field#2. **k** Radar plots of command accuracy for rescue tasks performed by different HMVIs in (I) complex field#1 and (II) complex field#2. The larger area of the plot indicated the higher command accuracy, while the symmetric feature of the plot represented the result uniformity applied to different HMVIs.

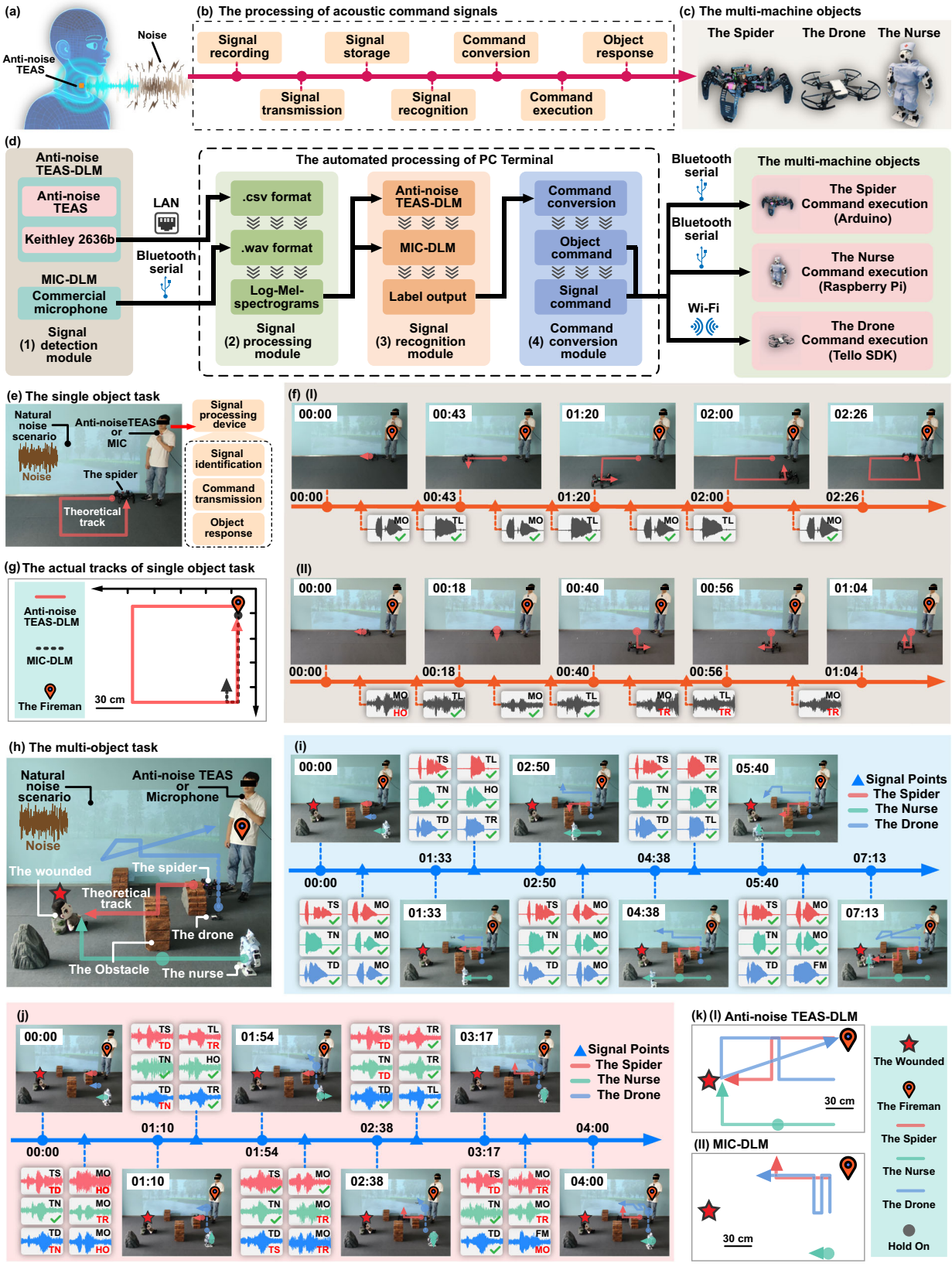
field#2 (Fig. 7e). In complex field#2 (Supplementary Fig. S52), the leader simultaneously controlled 3 objects (including the robot, drone, and nurse) via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing. Three objects were executed at a speed of 1 m/s from the starting point A to the entrance points B, C, and D of three blank areas. Subsequently, after reaching the entrances, the objects were required to enter the blank areas and search for the wounded. Upon locating the wounded, the leader would then control the objects to return to point B with an FM command, thereby achieving the rescue tasks. If the wounded rescue tasks were unable to be completed within 100 commands, the tasks of complex field#2 were judged to have failed, and the subsequent actions were terminated. Figure 7f illustrates the schematic diagram of the theoretical route of the rescue tasks in complex field#2.

The fireman leader directed the robot, drone, and nurse to execute the rescue tasks in three different environments via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing (Supplementary Fig. S53). The Anti-noise TEAS-DLM demonstrated effective control over three objects to execute the rescue tasks across three environments, with its recorded tracks aligning perfectly with the theoretical route (Fig. 7g). Although the MIC-DLM successfully executed the rescue tasks in the quiet environment, it exhibited significant command errors in the noisy environments. Despite repeated correction attempts, the leader still failed to complete the rescue tasks within 100 commands, resulting in the rescue tasks failure. Similarly, while the MIC-Hearing managed to accomplish the tasks across various environments, there were a lot of errors in the commands, which required the leader to repeatedly correct the commands. The number of commands, the command accuracy rate, and the overall rescue time t accomplish the rescue tasks in different environments were statistically analyzed (Fig. 7h–j). The total commands used by the Anti-noise TEAS-DLM to complete the rescue tasks across different environments was 42 (100%), 42 (100%), and 44 (97.7%), with the overall rescue time of 600 s, 700 s, and 785 s, respectively. Similarly, the differences in the overall rescue times were mainly due to the randomness in the execution of the BP commands, while the actual times due to the Anti-noise TEAS-DLM were 270 s, 270 s, and 285 s. In the case of the MIC-DLM, 42 commands (100%) were needed to accomplish the rescue task in a quiet environment, with an overall rescue time of 850 s (270 s + 580 s). However, in the natural and anthropogenic noise environments, the MIC-DLM failed to complete the rescue tasks within 100 commands, with the accuracy rates of only 44% and 23.5%. Meanwhile, the total number of commands for the MIC-Hearing in different environments were 42 (100%), 54 (55.6%) and 98 (70.4%), and the total time of the rescue task were 610 s (270 s + 340 s), 770 s (360 s + 410 s) and 730 s (360 s + 410 s), respectively. The accuracies of the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing in complex field#1 and complex field#2 were also summarized via radar plots. As exhibited in Fig. 7k, the command accuracy of the Anti-noise TEAS-DLM demonstrated a nearly symmetrical distribution across a large area in the radar plots, which indicated minimal susceptibility to variations across different scenarios. Conversely, the command accuracies of the MIC-DLM and the MIC-Hearing exhibited asymmetric and smaller area distributions in the radar plots, which suggested varying degrees of interference when deployed in noisy environments.

In addition, we also evaluated the feasibility of the Anti-noise TEAS-DLM by performing the complex human-machine collaborative tasks in real-life noisy scenarios with multiple machine objects (Fig. 8a). As exhibited in Fig. 8b, d, an automated real-time HMVI for automated commands processing of acoustic signals was constructed, which contained a signal detection module, a signal processing module, a signal recognition module, and a command conversion module. As a real person, the leader could send the commands by voice in various scenarios, which were detected and recorded by the Anti-noise TEAS or microphone. After signal transmission and storage, the terminal device would perform data processing and DLM signal recognition in real time. Finally, with wireless communication, the swarm of machine objects (Fig. 8c), including the spider robot, nurse robot, and drone, could be controlled to execute specific commands. In the automated HMVI process, which was built in the laboratory environment, the duration of the acoustic signal acquisition was ~2 s, the acoustic signal loading and transmission by a commercial device was ~8 s, as well as the model recognition and command sending for ~2 s. Thus, the execution process of each commanded action required a duration of ~12 s. Further optimization of automated signal acquisition and processing could effectively reduce the total duration required for command execution, resulting in an even faster response (Supplementary Figs. S54 and S55).

The feasibility of the Anti-noise TEAS-DLM and the MIC-DLM for complex HMVI tasks in harsh noisy scenarios was verified in quiet and natural noisy environments, respectively. In real-world testing, a real-life scenario of harsh natural noise was simulated by a harsh rainstorm background, which was played using a projector, and a high-decibel natural noise interference, which was generated by a high-power speaker. In an open field, the leader should control an individual spider robot to move according to a specific track in quiet or harsh natural noise environments, respectively (Fig. 8e). After the leader sent acoustic commands with the Anti-noise TEAS or microphone, the signal processing device would perform the automated data processing and signal recognition, followed by sending commands to control the movements of the spider robot. In this scenario, the leader controlled the spider robot to execute MO, TL, MO, TL, MO, TL, MO, TL, MO, respectively. The video screenshots of controlling the action of the spider robot in the real world were demonstrated, where each image exhibited the position map and track map of the spider robot at different moments during the task execution, respectively. In addition, the acoustic signal waveforms, the signal truth labels, and the signal recognition results at different moments were also shown along with the timeline. In the harsh

https://doi.org/10.1038/s41467-025-59523-6



noise scenario, the acoustic signals collected by the Anti-noise TEAS were basically not disturbed by the noise, where the waveform contours of the signals were consistent with those in the quiet environment (Supplementary Fig. S56), and possessed a recognition accuracy of 100%. After signal recording, transmission, storage, recognition, and command transmission, the leader was able to accurately control the spider robot to accomplish the assigned commands in real time (Fig. 8f-l). As a control group, the MIC-DLM was also utilized to perform the same task in a harsh scenario. Similarly, the leader controlled the spider robot to execute MO, TL, MO, TL, TL, MO, TL, TL, MO, respectively. Due to the interference from the harsh noise, the acoustic signals detected by the microphone presented obviously distortion and

Fig. 8 | Performance of the Anti-noise TEAS for multi-object complex human-machine collaboration in real-life noisy environments. a The Anti-noise TEAS detected the acoustic command signals for complex human-machine collaborative tasks. **b** Processing of acoustic command signals, which comprises signal recording, signal transmission, signal storage, signal recognition, command transmission, command execution, and object response. **c** A swarm of machines, including the spider robot, nurse robot, and drone, which could be controlled to execute specific commands in real time. **d** Logical schematic of the automatic processing of acoustic command signals, which comprised a signal detection module, a signal processing module, a signal recognition module, and a command conversion module. **e** Real-life noisy scenario constructed on an open field, which contained a harsh rainstorm background and a high-decibel natural noise interference. **f** Video screenshots of (I) the Anti-noise TEAS-DLM and (II) the MIC-DLM of controlling the spider robot performing the task in the real-life noisy scenario, where the object positions, actual tracks, acoustic signal waveforms, signal truth labels, and signal recognition results at different moments were exhibited, respectively. **g** Actual tracks of the spider robot performing the human-machine collaborative task via the Anti-noise TEAS-DLM and the MIC-DLM. **h** Real-life noisy scenario constructed on a complex field, which contained a harsh rainstorm background, a high-decibel natural noise interference, and multiple obstacles. Video screenshots of (**i**) the Anti-noise TEAS-DLM and **j** the MIC-DLM of controlling the swarm of machines performing the task in the real-life scenario, where the object positions, actual tracks, acoustic signal waveforms, signal truth labels, and signal recognition results at different moments were exhibited, respectively. **k** Actual tracks of the swarm of machines performing complex human-machine collaborative tasks via (I) the Anti-noise TEAS-DLM and (II) the MIC-DLM.

aberration, which were different from the signals in the quiet environment. The acoustic commands were recognized as HO, TL, MO, TL, TR, TR, TR, TR, respectively, where the recognition accuracy of the acoustic command signals was only 42.9%. Thus, the leader could not accurately control the spider robot to accomplish the task with MIC-DLM (Fig. 8f-II). Comparing the actual track with the theoretical track (Fig. 8g), the actual track of the Anti-noise TEAS-DLM was highly consistent with the theoretical track, while the MIC-DLMs presented a large deviation.

Further, a more complex real-life scenario was constructed to verify the effectiveness of the Anti-noise TEAS-DLM in performing complex HMVI tasks with multi-machine objects in noisy environments (Fig. 8h). In this actual complex scenario, the leader was required to use the Anti-noise TEAS-DLM or the MIC-DLM to control the spider robot, nurse robot, and drone to accomplish the casualty rescue task in the natural noise environment simultaneously (Movie S1). To find and rescue the wounded, the swarm of machines needed to avoid the obstacles, where the spider robot needed to execute MO, TL, MO, TR, MO, the nurse robot needed to execute MO, HO, MO, TR, MO, and the drone needed to execute MO, TR, MO, TL, FM, respectively. Due to the high maneuverability of the drone, it defaulted to move directly to the left and right when executing the TL and TR commands, whereas the spider robot and the nurse robot only perform a turn to the left or the right. Similarly, the video screenshots of controlling the actions of the swarm of machines in the real world were demonstrated, where each image exhibited the position map and track map of the multiple machines at different moments during the task execution, respectively. Additionally, the acoustic signal waveforms, the signal truth labels, and the signal recognition results at different moments were also displayed along with the timeline. As exhibited in Fig. 8i, the acoustic signals detected by the Anti-noise TEAS were virtually undisturbed by the harsh noise, where the signals were consistent with those in the quiet environment (Supplementary Fig. S57), and possessed a recognition accuracy of 100%. After signal recording, transmission, storage, recognition, and command transmission, the leader was capable of accurately controlling the swarm of machines to perform the complex task in the real world. Thus, the actual tracks of the multiple machines of the Anti-noise TEAS-DLM were consistent with the theoretical tracks (Fig. 8k-I). However, the object or action command signals detected by the microphone suffered from significant distortion due to the harsh noise, which were various significantly from the signals in the quiet environment (Fig. 8j). The recognition accuracy of the object commands was 46.6%, whereas the recognition accuracy of the action commands was 40%, and the execution accuracy of the object and action commands correctly at the same time was only 20%. Thus, the erroneous execution of the commands caused the swarm of machines to follow the wrong instruction, resulting in a large deviation of the actual tracks of the MIC-DLM from the theoretical tracks (Fig. 8k-II). The above real-life scenario experiments also further supported the previous HMVI experiments in virtual scenarios, verifying that the Anti-noise TEAS-DLM was able to effectively control a swarm of machines to perform complex human-machine collaborative tasks in noisy environments. Therefore, compared to conventional HMVI, the Anti-noise TEAS-DLM exhibited robust resistance against noisy environments, which enabled effective command execution for multi-object human-machine collaboration in complex scenarios.

Discussion

In this work, we developed an Anti-noise TEAS based on flexible nanopillar substrates, and integrated it with a CNN-based DLM to identify acoustic signals for complex human-machine collaboration in noisy environments. With the high-sensitivity, high-stability, and a wide response frequency range, the Anti-noise TEAS could effectively detect the acoustic fundamental frequency signals from the laryngeal mixed-mode vibrations through contact sensing. This enables the detection of multi-type acoustic signals, including music, animal sounds, and speeches. Owing to its optimal structure, the Anti-noise TEAS possessed desirable noise immunity, while the accurate meaning of the acoustic signals was recognized with the assistance of DLM. The Anti-noise TEAS-DLM could effectively recognize complex acoustic signals in noisy environments with high accuracy, almost immune to significant noise. Additionally, visualization techniques such as t-SNE and Grad-CAM proved the effectiveness of the Anti-noise TEAS-DLM in feature extraction and information recognition for multiple acoustic signals. Furthermore, in the simulated virtual world and real-life scenarios, the Anti-noise TEAS-DLM was applied to control robots, drones, and nurses to perform complex post-disaster casualty rescue tasks across various noisy environments. While conventional microphones failed to transmit correct commands due to significant noisy interference, the Anti-noise TEAS-DLM exhibited desirable noisy interference resistance, with the rescue tasks' success rate reaching 100%. Thus, the Anti-noise TEAS-DLM exhibited great anti-noise capability, which could address the inaccuracies inherent in conventional HMVI within noisy environments. Deep learning-enhanced Anti-noise TEAS possessed high sensitivity, anti-noise, and robust stability, which provided a practical solution for human-machine collaboration in noisy scenarios, such as post-disaster rescue, collaborative operations, and wilderness exploration. The concept of the DLM-enhanced advanced noise-resistant acoustic sensors would positively promote the development of advanced HMVI, addressing the accuracy challenges prevalent in conventional HMVI under noisy interference. This highly integrated HMVI mode possessed the possibility of being extended to a variety of human-machine collaborative tasks, which offered new opportunities and directions for diversified and intelligent human-machine collaboration.

Methods

The fabrication of the positive friction electrode

A cicada wing bio-sample was immersed in deionized water, subjected to ultrasonic cleaning for 20 min, and then placed in a dust-free vacuum environment for natural drying. Trimethylolpropane ethoxylate triacrylate (Aldrich) and 2-hydroxy-2-methylpropiophenone (Macklin) were thoroughly mixed in a ratio of 100:2 (PEGDA), and

then the solution was degassed by placing it in a vacuum desiccator (30 min) under light-avoidant conditions. The cleaned cicada wing bio-sample was cut to an appropriate size and fixed on glass slides. In a vacuum desiccator, the PEGDA was applied uniformly dropwise on the prepared cicada wing bio-sample. After the PEGDA had sufficiently covered the surface of the cicada wing, the sample was removed, and the PEGDA was cured using UV light (365 nm, 80% power, 240 s, Intelliray 600, Uvitron). After the PEGDA was sufficiently crosslinked, the samples were removed and placed in an incubator for 2 h to cool. The cooled samples underwent ultrasonication in deionized water for 6–8 h. Subsequently, the cured PEGDA negative molds were carefully peeled off from the cicada wing bio-sample and placed in a vacuum environment for natural drying. The polydimethylsiloxane base (PDMS, Sylgard 184, Dow Corning) and curing agent were combined in a 10:1 weight ratio to prepare the PDMS precursor. Next, a mixture of single-walled carbon nanotubes (CNTs, Nationwide Technology, Inc.) was added to the PDMS precursor at a mass ratio of 1:8, and through the screeding machine to fully mix and de-bubble. The CNTs/PDMS reagent mixture was uniformly spread onto a PEGDA negative mold through drop-coating in a vacuum, followed by rotation at 4000–5000 rpm for 60 s (Smart Coater 100, Best Tools). The samples were subsequently transferred to an incubator and thermal treatment at 60 °C for 8 h to ensure complete curing of the CNTs/PDMS composite. Once the CNTs/PDMS material had fully cured, the flexible CNTs/PDMS nanopillars substrate was delicately peeled off from the PEGDA negative mold. Then, a layer of Au was deposited onto the surface of the flexible CNTs/PDMS nanopillars substrate using magnetron sputtering (30 mA, 300 s, GVC-2000, GEVEE-TECH). Additionally, a layer of conductive copper tape was affixed to the substrate's backside to create a positive friction electrode. To ensure a secure connection, a thermocompression wire was applied to the back of the conductive copper tape under pressure and heat (120 °C), serving as a signal lead wire for the positive friction electrode.

The fabrication of the negative friction electrode

FEP (50 μm) films were cleaned using deionized water and placed in a dust-free vacuum environment for natural drying. The front side (friction layer interface) of the clean FEP was corona treated using a plasma corona treatment machine (3 s, ten times, Hangzhou Shangqiang Intelligent Technology Company), and then the back side (export layer interface) of the FEP was magnetron sputtered (30 mA, 300 s) with an Au layer. By applying pressure by heating (120 °C), a thermocompression lead was adhered to the back side of the conductive copper tape as a signal lead. Kapton tape (PI, 50 μm) was then utilized to adhere to the surface of the Au layer on the backside of the FEP as an insulating layer to prevent short-circuiting and contact with interfering signals, thus forming a negative friction electrode.

The assembly of the Anti-noise TEAS

The Anti-noise TEAS was assembled by affixing the negative friction electrode to the positive friction electrode, in which the FEP was used as the negative friction electrode and the CNTs/PDMS nanopillars substrate was used as the positive friction electrode. The initial spacing distance between the two friction layers was adjusted to ~300 μm by placing an appropriately sized polyethylene glycol terephthalate (PET, 300 μm) film spacer between the positive and negative friction layers. The device can be mechanically designed to feature curved bends, which could effectively increase the fit of the sensor-skin contact interface.

Material morphology characterization

The SEM images of the surface morphology of the flexible CNTs/PDMS nanopillars substrate were taken using a scanning electron microscope (SEM, Supra 60, Zeiss). Prior to SEM imaging, a layer of Au was sputtered on the surface of the substrate using a magnetron sputtering unit (KYKY-SBC-12, Kyky Technology Co. Ltd, China). The SEM image was obtained at an accelerating voltage of 10 kV and in secondary electron mode.

Optimal structure performance test of Anti-noise TEAS

In the acoustic response test experiment of the Anti-noise TEAS, a standard player (EDIFIER, G1500bar), a commercial microphone (UGREEN, CM592), a sound pressure level detector (DELIXI, DLY-2202), and a precision source/measurement unit (Keithley 2636B, SMU) were utilized to test and evaluate the sensor. To assess the acoustic response characteristics of the Anti-noise TEAS with varying spacer distances, different Anti-noise TEAS were prepared using different thicknesses of the PET materials of 300 μm, 400 μm, 500 μm, 600 μm, 800 μm, 1200 μm, and 1500 μm, respectively. These Anti-noise TEAS were then affixed onto the standard player, and evaluated by the linear standard sweep audio source with a frequency of 50–4000 Hz, a SPL of 102 dB, and a duration time of 5 s. The resulting acoustic-current response signals were captured using the SMU (sampling rate of 8 kHz). The peak value of the acoustic-current response signals measured by the 300 μm FEP was utilized as a reference for normalization. To test the acoustic response characteristics and noise immunity of the Anti-noise TEAS with different FEP thicknesses, different Anti-noise TEAS were prepared with 15 μm, 20 μm, 30 μm, 50 μm, 80 μm, 100 μm, 150 μm, 200 μm, and 250 μm FEP films, respectively. These Anti-noise TEAS were then affixed onto the standard player, and evaluated by the linear standard sweep audio source with a frequency of 50–4000 Hz, a SPL of 102 dB, and a duration time of 5 s. The resulting acoustic-current response signals were captured using the SMU (8 kHz). The peak value of the acoustic-current response signal measured by the 15 μm FEP was utilized as a reference for normalization. For testing the noise immunity of the Anti-noise TEAS, the sensors were separated from the standard player by 1 cm, white noise was played at a SPL of 100 dB for the Anti-noise TEAS with different FEP thicknesses, and the acoustic-current response signals of the sensors were captured using the SMU (8 kHz). The average value of the acoustic-current response signal measured by the 15 μm FEP was utilized as a reference for normalization.

Acoustic response characterization test of Anti-noise TEAS

To evaluate the typical acoustic response characteristics of the Anti-noise TEAS, the Anti-noise TEAS was affixed onto the standard player, and evaluated by the linear standard sweep audio source with a frequency of 50–4000 Hz, and a duration of 5 s. The resulting acoustic-current response signals were captured using the SMU (8 kHz). The peak value of the typical acoustic-current response signals was utilized as a reference for normalization. With a window of 800 ms and a moving step of 200 ms, the acoustic signal waveform curve was divided into frames and windowed, and then the STFT was performed to obtain the frequency characteristic spectrogram. To examine the acoustic response characteristics of the Anti-noise TEAS at various SPL, the Anti-noise TEAS was affixed onto the standard player, and evaluated by an acoustic signals of 2000 Hz, at the SPL of 70 dB, 77 dB, 84 dB, 90 dB, 93 dB, 96 dB, 98 dB, 101 dB, 103 dB, and 106 dB, respectively. The resulting acoustic-current response signals were captured using the SMU (8 kHz). The average value of the acoustic-current response signal measured at 106 dB was utilized as a reference for normalization. To assess the sweep frequency response characteristics of the Anti-noise TEAS at various SPL, the Anti-noise TEAS was affixed onto the standard player, and evaluated by the linear standard sweep audio source with a frequency of 50–4000 Hz, and the SPL of 84 dB, 86 dB, 90 dB, 93 dB, 96 dB, 98 dB, 101 dB, 103 dB, 105 dB, and 106 dB, respectively. The resulting acoustic-current response signals were captured using the SMU (8 kHz). The peak value of the acoustic-current response signal measured at 106 dB was utilized as a reference for normalization.

Single frequency response and durability test of Anti-noise TEAS

To evaluate the single frequency response characteristics of the Anti-noise TEAS across various frequency, the Anti-noise TEAS was affixed onto the standard player, and evaluated by the acoustic signals at frequencies of 400 Hz, 600 Hz, 800 Hz, 1000 Hz, 1200 Hz, 1400 Hz, 1600 Hz, 1800 Hz, 2000 Hz, 2200 Hz, and 2400 Hz, respectively. The resulting acoustic-current response signals were captured using the SMU (8 kHz). The acoustic-current response signals were normalized and subjected to a FFT. To evaluate the robustness and durability of the Anti-noise TEAS, the Anti-noise TEAS was affixed onto the standard player, and evaluated by the acoustic signals of 200 Hz (~7×10^5 cycles). The acoustic-current response signals of the Anti-noise TEAS were acquired using the SMU (1 kHz). The acoustic-current response signals obtained during the durability test were normalized.

Detection of the fundamental acoustic signals with Anti-noise TEAS

The Anti-noise TEAS was attached to the subject's throat with medical breathable tape (Jiangsu Hons Bioengineering Group., Ltd.). In a quiet environment, the acoustic-current response signals of the sensor were tested and recorded using an SMU without adding a bias voltage (sampling rate of 8 kHz). To simulate an extreme noise ring, a standard player was placed at a distance of 10–20 cm in front of the subject, and natural and anthropogenic environmental noise were played at a sound intensity of 106 dB, respectively. Then, the fundamental acoustic signals of the human larynx in the noise environment were captured from the subject using the Anti-noise TEAS. Similarly, a commercial microphone was placed close to the subject's throat, and voice command signals were captured in a quiet environment, a natural noise environment, and an anthropogenic noise environment, respectively.

The DLMs for ASR

A high-pass filter (55 Hz) was utilized to remove the interference of industrial frequency noise and motion artifacts from the Anti-noise TEAS acoustic waveform signal. The acoustic signals with different lengths were unified into 3.2 s by truncation or padding, and all the data samples were normalized. A 128 × 128-dimensional Log-Mel-spectrogram of the acoustic signals was obtained by splitting frames, adding windows, FFT, Mel filtering, and logarithmic operations on the acoustic waveform signals with 800 ms as the window and 200 ms as the step size. By associating the Log-Mel-spectrogram with the labels of the acoustic signals, the input feature spectrogram for the DLM was obtained. The DLM for ASR consisted of convolution layers, max-pooling layers, drop-out layers, flatten layers, fully-connected layers, and sigmoid layers/softmax layers. Among them, the convolution operation of the convolution layers could extract and retain the deep nonlinear feature information of the Log-Mel-spectrogram from shallow to deep, and the Max-pooling layers could downsample the feature spectrogram output from the convolution layers to reduce the number of model parameters and prevent the model from overfitting. Drop-out layers could randomly discard part of the neuron outputs of the model to reduce the number of model parameters, which further prevented the model from overfitting during the training process. Flattening layers could unfold the multi-dimensional feature maps output from the convolution layers into one-dimensional feature data. Fully-connected layers could further perform feature extraction and dimensionality reduction on the one-dimensional feature data to extract and retain the most important information components in the acoustic signals. Finally, sigmoid layers/softmax layers were used as activation function layers that enable label classification of different classes of signals. The DLM was trained using the divided training dataset. During the model training process, Adam function was used as the optimizer of the model, cross-entropy as the loss function of the model, accuracy as the monitoring of the model, and a 10-fold cross-validation method was used for the appropriate selection and validation of the hyperparameters of the model structure. The DLM for the acoustic signal detected by the microphone used the same construction process, but the choice of model parameters differed slightly depending on the data samples.

Visualization of deep learning data

The t-SNE clustering method was utilized for data downscaling and visualization clustering operations. The original speech waveform signals, feature spectrograms, the output features of the fourth layer of the convolution layers, and the output features of the third layer of the full-connected layer were used as input data, which were down-scaled to 2-dimensions by the t-SNE clustering method, and visualized and plotted according to the sample labels. Visualization of deep learning neural networks using the Grad-CAM technique. Add the Grad-CAM visualization technique to the trained DLM, which could reverse the decision-making results of the DLM, and obtain the attention score of the last layer of feature output spectrogram of the model. The score heatmap was inversely mapped to the original feature spectrogram to obtain the visualization heatmap of the original feature spectrogram.

The human-machine collaborative tasks in an open field

An open field scenario was designed in the virtual world to test the performance of the Anti-noise TEAS-DLM in human-machine collaborative tasks. In the open field, the fireman leader (a real person) would use voice to send commands through the Anti-noise TEAS, which were recognized and parsed as task commands by the DLM, including action commands (MO, CB, TL, TR, FM, HO, and BP) and object commands (TD, TS, and TN). The commands were wirelessly remote-controlled via communication devices to collaborate on three execution objects, including a drone, a robot, and a nurse (a real person) to accomplish the tasks. In each field, there were three types of sound environments, including quiet, natural noise, and anthropogenic noise environments. As a control group, the commands given by the leader were simultaneously recorded using a commercial microphone and then converted into control signals to be transmitted to the execution objects via MIC-DLM. To simulate the complex interactions between multi-object collaborative tasks in real situations, the nurse also acted as one of the information-receiving entities, and judged the types of command signals captured by the microphone through hearing perception (MIC-Hearing), followed by performing the tasks. A total of four specified routes were included in the open field. Route#1 included MO, TL, HO, FM, CB, TR, HO, and FM. Route#2 included MO, TR, MO, TR, MO, TR, HO, and FM. Route#3 included MO, CB, TL, TR, CB, MO, TR, and TL. Route#4 was a multi-object collaborative task, in which the robot executed MO, TL, FM, the drone executed TR, CB, FM, and the nurse executed TL, CB, and HO. The fireman must specify the object to be controlled by using the object command signals before issuing the action command each time. For instance, when controlling the robot to move forward, the command signals TS + MO were issued, and when directed the nurse to turn left, the commands TN + TL were issued. Additionally, blue, brown, and purple colors indicated the moving tracks in the quiet, natural noise, and anthropogenic noise environments, respectively, while the position of the leader was indicated by orange coordinates.

The human-machine collaborative tasks in complex field#1 and complex field#2

The complex field#1 and complex field#2 were designed in the virtual world to test the performance of the Anti-noise TEAS-DLM in human-machine collaborative tasks. In the complex field#1 and complex field#2, the fireman leader would use voice to send commands through the Anti-noise TEAS, which were recognized and parsed as task commands by the DLM. In complex field#1, the fireman directed the robot from the starting point (point A) to the entrance of the blank area (point B) at a speed of 1 m/s. Subsequently, the robot was required

to enter the blank area and execute a BP command to search for the wounded. Upon locating the wounded (indicated by a pentagram), the leader would then control the robot to return to point B by an FM command, and the task of rescuing the wounded was accomplished. If the wounded rescue task was unable to be completed within 50 commands, the task of complex field#1 was judged to have failed, and the subsequent actions were terminated. In complex field#2, the leader simultaneously controlled three objects (including the robot, the drone, and the nurse) via the Anti-noise TEAS-DLM, the MIC-DLM, and the MIC-Hearing. If the wounded rescue tasks were unable to be completed within 100 commands, the tasks of complex field#2 were judged to have failed, and the subsequent actions were terminated. The fireman must specify the object to be controlled by using the object command signals before issuing the action command each time. During the execution of the rescue tasks, if the objects deviated from the theoretical route due to incorrect action commands, the leader would provide the correct commands based on the actual situation. This procedure ensured the successful correction of actions aimed at the wounded rescue tasks. If the BP command was triggered incorrectly during the rescue tasks, the leader was required to abort the BP action by the HO and re-issue the correct command according to the actual situation.

Statistics on the number of commands and the total time of rescue tasks

In the process of rescue tasks, the command execution duration time of MO, CB, TL, TR, and HO was 15 s, while the duration time of BP was decided according to the actual situation of the patrol. During the execution of the BP, the controlled object would randomly change its direction of action every 10 s. When the manipulation object located the specified wound or received an HO command, the patrol terminated, and the next command was executed. The statistics for the total number of rescue tasks were contained within all command types, including action commands and object commands. The statistics for the total time of rescue tasks need to include the total duration of the BP commands. From the start of the tasks until the FM command was received after locating the wound, the rescue tasks were regarded as complete, and the timer for the rescue mission ended.

The manipulation of multi-machine robot objects

The spider robot (Hiwonder, Spiderbot, Shenzhen Mirage Technology Co., Ltd.) used the Arduino control platform, which could control the collaborative motion of each joint servo of the spider robot. Based on the secondary development of the serial communication protocol, the PC terminal could directly call the Bluetooth serial port through a Python program, thus controlling the spider robot's movement. The nurse robot (Hiwonder, TonyPi Pro, Shenzhen Mirage Technology Co., Ltd.) utilized the Raspberry Pi as the core component, which could control the collaborative motion of each joint servo of the nurse robot. Based on the encoding and decoding of the serial communication protocol under the Linux system, the PC terminal could directly call the Bluetooth serial port through a Python program, which could control the movement of the nurse robot. The drone (RYZE, Tello, Shenzhen Rui Ki Technology Co., Ltd) used the Tello SDK for secondary development, programming control of the drone. The PC terminal can directly call Wi-Fi through a Python program to control the drone's movement. In the automated real-time HMVI, the acoustic command signals were detected by the Anti-noise TEAS or microphone and transmitted to the PC terminal through the network cable or Bluetooth serial port. The PC terminal converted the original acoustic data into Log-Mel-spectrograms through the Python program, and then recognized the signal types through the DLM. Finally, the signal types were converted into command signals to directly control the spider robot, the nurse robot, or the drone to execute specific commands with Bluetooth serial communication or Wi-Fi communication.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All relevant data supporting the key findings of this study are available within the article and its Supplementary Information files. Source data is provided as a Source Data file. Source data are provided with this paper.

Code availability

The code that supports the funding of this study is available on GitHub with the identifier https://github.com/VTminY/Anti-noise-TEAS.

References

1. Yin, R., Wang, D., Zhao, S., Lou, Z. & Shen, G. Wearable sensors-enabled human–machine interaction systems: from design to application. *Adv. Funct. Mater.* **31**, 2008936 (2021).
2. Heng, W., Solomon, S. & Gao, W. Flexible electronics and devices as human–machine interfaces for medical robotics. *Adv. Mater.* **34**, 2107902 (2022).
3. Hu, Z. et al. Machine learning for tactile perception: advancements, challenges, and opportunities. *Adv. Intell. Syst.* **5**, 2200371 (2023).
4. Tao, K. et al. Deep-learning enabled active biomimetic multifunctional hydrogel electronic skin. *ACS Nano* **17**, 16160–16173 (2023).
5. Qin, K. et al. Magnetic array assisted triboelectric nanogenerator sensor for real-time gesture interaction. *Nanomicro Lett.* **13**, 9 (2021).
6. Sun, T. et al. Machine learning-coupled vertical graphene triboelectric pressure sensors array as artificial tactile receptor for finger action recognition. *Nano Energy* **123**, 109395 (2024).
7. Kim, T. et al. Ultra-stable and tough bioinspired crack-based tactile sensor for small legged robots. *npj Flex. Electron.* **7**, 22 (2023).
8. Kim, K. K. et al. A substrate-less nanomesh receptor with meta-learning for rapid hand task recognition. *Nat. Electron.* **6**, 64–75 (2023).
9. Griggs, W. S. et al. Decoding motor plans using a closed-loop ultrasonic brain–machine interface. *Nat. Neurosci.* **27**, 196–207 (2023).
10. Zhao, Y., Chen, C., Lu, B., Zhu, X. & Gu, G. All 3D-printed soft high-density surface electromyography electrode arrays for accurate muscle activation mapping and decomposition. *Adv. Funct. Mater.* **34**, 2312480 (2023).
11. Yang, S. et al. Stretchable surface electromyography electrode array patch for tendon location and muscle injury prevention. *Nat. Commun.* **14**, 6494 (2023).
12. Renton, A. I., Mattingley, J. B. & Painter, D. R. Optimising non-invasive brain-computer interface systems for free communication between naïve human participants. *Sci. Rep.* **9**, 18705 (2019).
13. Fan, J., Vargas, L., Kamper, D. G. & Hu, X. Robust neural decoding for dexterous control of robotic hand kinematics. *Comput. Biol. Med.* **162**, 107139 (2023).
14. Yu, Y. et al. All-printed soft human-machine interface for robotic physicochemical sensing. *Sci. Robot* **7**, eabn0495 (2022).
15. Nassif, A. B., Shahin, I., Attili, I., Azzeh, M. & Shaalan, K. Speech recognition using deep neural networks: a systematic review. *IEEE Access* **7**, 19143–19165 (2019).
16. Xiao, X., Fang, Y., Xiao, X., Xu, J. & Chen, J. Machine-learning-aided self-powered assistive physical therapy devices. *ACS Nano* **15**, 18633–18646 (2021).

17. Wang, Y. et al. Machine learning-enhanced flexible mechanical sensing. *Nanomicro Lett.* **15**, 55 (2023).
18. Bai, Y. et al. Acoustic-based sensing and applications: a survey. *Comput. Netw.* **181**, 107447 (2020).
19. Peng, T.-h & Huang, J. H. Electroacoustic analysis of a directional moving-coil microphone using combined lumped-parameter models and BEM. *Sens. Actuators A Phys.* **332**, 112914 (2021).
20. Tao, L.-Q. et al. An intelligent artificial throat with sound-sensing ability based on laser induced graphene. *Nat. Commun.* **8**, 14579 (2017).
21. Gong, S. et al. Hierarchically resistive skins as specific and multimetric on-throat wearable biosensors. *Nat. Nanotechnol.* **18**, 889 (2023).
22. 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).
23. Shao, H. et al. High-performance voice recognition based on piezoelectric polyacrylonitrile nanofibers. *Adv. Electron. Mater.* **7**, 2100206 (2021).
24. Zhao, J. et al. Anti-interference self-powered acoustic fabric for complex acoustic environments. *Nano Energy* **113**, 108534 (2023).
25. Sun, H. et al. Graphene-based dual-function acoustic transducers for machine learning-assisted human–robot interfaces. *InfoMat.* **5**, e12385 (2023).
26. Lin, Z. et al. A personalized acoustic interface for wearable human–machine interaction. *Adv. Funct. Mater.* **32**, 2109430 (2022).
27. Jiang, Y. et al. Ultrathin eardrum-inspired self-powered acoustic sensor for vocal synchronization recognition with the assistance of machine learning. *Small* **18**, 2106960 (2022).
28. Guo, H. et al. A highly sensitive, self-powered triboelectric auditory sensor for social robotics and hearing aids. *Sci. Robot.* **3**, eaat2516 (2018).
29. Zhao, J. et al. Self-powered speech recognition system for deaf users. *Cell Rep. Phys. Sci.* **3**, 101168 (2022).
30. Jang, J., Lee, J., Jang, J. H. & Choi, H. A triboelectric-based artificial basilar membrane to mimic cochlear tonotopy. *Adv. Healthc. Mater.* **5**, 2481–2487 (2016).
31. Zhu, G. et al. A multi-hole resonator enhanced acoustic energy harvester for ultra-high electrical output and machine-learning-assisted intelligent voice sensing. *Nano Energy* **108**, 108237 (2023).
32. Xu, W. et al. Laminated triboelectric acoustic energy harvester based on electrospun nanofiber towards real-time noise decibel monitoring. *Nano Energy* **99**, 107348 (2022).
33. Fang, Y. et al. Ambulatory cardiovascular monitoring via a machine-learning-assisted textile triboelectric sensor. *Adv. Mater.* **33**, 2104178 (2021).
34. Zhou, H. et al. Bionic ultra-sensitive self-powered electromechanical sensor for muscle-triggered communication application. *Adv. Sci.* **8**, 2101020 (2021).
35. Tao, K. et al. Ultra-sensitive, deformable, and transparent triboelectric tactile sensor based on micro-pyramid patterned ionic hydrogel for interactive human–machine interfaces. *Adv. Sci.* **9**, 2104168 (2022).
36. Che, Z. et al. Speaking without vocal folds using a machine-learning-assisted wearable sensing-actuation system. *Nat. Commun.* **15**, 1873 (2024).
37. Kim, K. K. et al. A deep-learned skin sensor decoding the epicentral human motions. *Nat. Commun.* **11**, 2149 (2020).
38. Yang, Q. et al. Mixed-modality speech recognition and interaction using a wearable artificial throat. *Nat. Mach. Intell.* **5**, 169–180 (2023).
39. Zhao, H. et al. Intelligent recognition using ultralight multifunctional nano-layered carbon aerogel sensors with human-like tactile perception. *Nanomicro Lett.* **16**, 11 (2023).
40. Qu, X. C. et al. Artificial tactile perception smart finger for material identification based on triboelectric sensing. *Sci. Adv.* **8**, eabq2521 (2022).
41. Niu, H. et al. Micropyramid array Bimodal electronic skin for intelligent material and surface shape perception based on capacitive sensing. *Adv. Sci.* **11**, 2305528 (2024).
42. Jin, T. et al. Triboelectric nanogenerator sensors for soft robotics aiming at digital twin applications. *Nat. Commun.* **11**, 5381 (2020).
43. Tian, H. et al. Bioinspired dual-channel speech recognition using graphene-based electromyographic and mechanical sensors. *Cell Rep. Phys. Sci.* **3**, 101075 (2022).
44. Liu, H. et al. An epidermal sEMG tattoo-like patch as a new human–machine interface for patients with loss of voice. *Microsyst. Nanoeng.* **6**, 16 (2020).
45. Zheng, C. et al. Dual-path tr8k]iansformer-based network with equalization-generation components prediction for flexible vibrational sensor speech enhancement in the time domain. *J. Acoust. Soc. Am.* **151**, 2814–2825 (2022).
46. Zou, H. et al. Quantifying the triboelectric series. *Nat. Commun.* **10**, 1427 (2019).
47. Deng, W. et al. Ternary electrification layered architecture for high-performance triboelectric nanogenerators. *ACS Nano* **14**, 9050–9058 (2020).
48. Tan, D. & Xu, B. Advanced interfacial design for electronic skins with customizable functionalities and wearability. *Adv. Funct. Mater.* **33**, 2306793 (2023).

Acknowledgements

The authors would like to acknowledge financial support from the National Natural Science Foundation of China (grant no. T2225010 and 32171399 to X.X., 32171456 to H.-j.C., and 32401202 to S.H.), the Guangdong Basic and Applied Basic Research Foundation (grant no. 2023A1515011267 to X.X.), the Science and Technology Program of Guangzhou, China (grant no. 2024B03J0121 to X.X. and 2024B03J1284 to H.-j.C.), the Shenzhen Science and Technology Program (grant no. RCBS20231211090558093 to S.H.), the China Postdoctoral Science Foundation (2023TQ0386 to S.H.), and the Fundamental Research Funds for the Central Universities, Sun Yat-sen University (no. 24xkjc011 to X.X.).

Author contributions

C.Y. and X.X. conceived the concept, designed the work, analyzed data, and wrote the manuscript. C.Y., S.L., T.S. and M.H. performed the experiments. C.Y., X.X., and S.L. performed the theoretical calculations. C.Y., S.L., S.H., H.-j.C. and X.X. performed statistical analyses of datasets and aided in the preparation of displays communicating datasets. Z. Liu, S.H., G.X., H.O., Y.T., Y.Q., M.L., Z. Li, P.S. and H.-j.C. provided suggestions and assisted in wearable device experiments. X.X. supervised the study. All authors discussed the results and assisted in the preparation of the manuscript.

Competing interests

The authors declare no competing interests.

Ethics approval

Three participants participated (one female and two males, aged from 18- to 30-years-old) in this work to evaluate the performance of the Anti-noise TEAS-DLM. The informed consent of all participants was obtained prior to inclusion in this study. The study was approved by the Institutional Review Board of the Seventh Affiliated Hospital and Third Affiliated Hospital of Sun Yat-sen University (KY-2024-037-02 and A2023-582-01), as well as by the Committee on Ethics of Beijing Institute of Nanoenergy and Nanosystems (R-2023002).

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-59523-6.

Correspondence and requests for materials should be addressed to Xi Xie.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]State Key Laboratory of Optoelectronic Materials and Technologies, School of Electronics and Information Technology, Sun Yat-Sen University, Guangzhou, China. [2]Guangdong Province Key Laboratory of Display Material and Technology, Sun Yat-Sen University, Guangzhou, China. [3]School of Biomedical Engineering, Sun Yat-Sen University, Shenzhen, China. [4]The Third Affiliated Hospital of Sun Yat-Sen University, Sun Yat-Sen University, Guangzhou, China. [5]School of Nanoscience and Engineering, University of Chinese Academy of Sciences, Beijing, China. [6]Beijing Institute of Nanoenergy and Nanosystems, Chinese Academy of Sciences, Beijing, China. [7]Department of Biomedical Engineering, The City University of Hong Kong, Kowloon, China. ✉e-mail: xiexi27@mail.sysu.edu.cn