EarSSR: Silent Speech Recognition via Earphones

Xue Sun, Jie Xiong, Chao Feng, Haoyu Li, Yuli Wu, Dingyi Fang, and Xiaojiang Chen

***Abstract*—As the most natural and convenient way to communicate with people, speech is always preferred in Human-Computer Interactions. However, voice-based interaction still has several limitations. It raises privacy concerns in some circumstances and the accuracy severely degrades in noisy environments. To address these limitations, silent speech recognition (SSR) has been proposed, which leverages the inaudible information (e.g., lip movements and throat vibration) to recognize the speech. In this paper, we present EarSSR, an earphone-based silent speech recognition system to enable interaction without a need of vocalization. The key insight is that when people are speaking, their ear canals exhibit unique deformation patterns and the corresponding deformation patterns are related to words/letters even without any vocalization. We utilize the built-in microphone and speaker of an earphone to capture the ear canal deformation. Ultrasound signals are emitted and the reflected signals are analyzed to extract the signal features corresponding to speech-induced ear canal deformation for silent speech recognition. We design a two-channel hierarchical convolutional neural network to achieve fine-grained letter/word recognition. Our extensive experiments show that EarSSR can achieve an accuracy of 82% for single alphabetic letter recognition and an accuracy of 93% for word recognition.**

***Index Terms*—Acoustic sensing, silent speech recognition, earphone.**

Manuscript received 15 December 2022; revised 8 December 2023; accepted 28 December 2023. Date of publication 22 January 2024; date of current version 2 July 2024. This work was supported in part by NSFC A3 Foresight Program under Grant 62061146001, in part by the National Natural Science Foundation of China under Grant 62272388 and Grant 62372372, and in part by the Key Project of Shaanxi Province International Science and Technology Cooperation Program under Grant 2023-GHZD-04 and Grant 2023-GHZD-06, and in part by the Shaanxi Qinchuangyuan Program under Grant QCYRCXM-2023-103. Recommended for acceptance by K. Wu. *(Corresponding author: Xiaojiang Chen.)*

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Northwest University of China.

Xue Sun, Chao Feng, Haoyu Li, and Yuli Wu are with the Shaanxi International Joint Research Centre for the Battery-Free Internet of Things, School of Information Science and Technology, Northwest University, Xi'an 710127, China (e-mail: sunxue@stumail.nwu.edu.cn; chaofeng@nwu.edu.cn; lihaoyu@stumail.nwu.edu.cn; wuyuli@stumail.nwu.edu.cn).

Jie Xiong is with the Microsoft Research Asia, Shanghai 200000, China.

Dingyi Fang is with the Xi'an Key Laboratory of Advanced Computing and System Security, School of Information Science and Technology, Northwest University, Xi'an 710127, China (e-mail: dyf@nwu.edu.cn).

Xiaojiang Chen is with the Xi'an Key Laboratory of Advanced Computing and System Security, School of Information Science and Technology, Northwest University, Xi'an 710127, China, also with the School of Information Science and Technology, Northwest University, Xi'an 710127, China, and also with the Xi'an Advanced Battery-Free Sensing and Computing Technology International Science and Technology Cooperation Base, Xi'an 710127, China (e-mail: jxiong@cs.umass.edu).

Digital Object Identifier 10.1109/TMC.2024.3356719

I. Introduction

SPEECH interaction plays an important role in our daily lives. It is reported that the global market related to speech and voice recognition will reach 22 billion by 2026 [1]. The obvious advantage of speech communication is that it is the most convenient interaction method. Compared with other commonly-seen interaction schemes such as typing and gesture, speech can achieve higher efficiency. According to a recent study [2], speaking is on average four times faster than typing on a touch screen. Also, speech interaction is not affected by lighting conditions.

The primary carrier of speech is the human voice. However, in some scenarios, vocalization is infeasible such as in a meeting, or inefficient such as in a noisy environment. Silent speech recognition (SSR) was therefore proposed as an alternative in these scenarios. SSR allows the user to communicate with people or devices in a silent manner. The user simply mouths the utterance without actually voicing it and the speech information can still be captured through a variety of SSR technologies. Existing SSR technologies can mainly be classified into two categories, i.e., contact-based and contact-free approaches. Specifically, contact-based approaches need to attach sensors (e.g., Electroencephalogram electrodes [3] or Electromyography electrodes [4], [5]) to human body. Contact-based approaches are inconvenient in a lot of real-world scenarios. Most contact-free approaches employ cameras or wireless signals to capture the mouth and vocal tract movements to infer silent speech [6], [7], [8]. For example, EchoWhisper [7] utilizes acoustic signals transmitted from a smartphone to extract the Doppler shift induced by mouth movements to achieve silent speech recognition. However, while lip movements can be captured, the information about the tongue's movement which is also critical for recognizing silent speech can hardly be obtained. Moreover, this method requires the smartphone to be placed in front of the user's mouth to make sure the signal can be reflected from the mouth. The above limitations motivate us to propose an alternative speech recognition system. We propose to utilize the most popular wearable devices, i.e., earphones to achieve silent speech recognition. The basic principle behind this is that when a person speaks even without vocalization, his/her ear canal gets deformed and the deformation pattern is uniquely related to letters and words. Motivated by this idea, we present EarSSR, which utilizes earphones to sense the deformation of ear canal caused by silent speaking, enabling a new silent speech recognition modality.

Specifically, as shown in Fig. 1, we employ inaudible acoustic signals to sense the fine-grained ear canal deformation when a person speaks without vocalization (termed silent speaking).

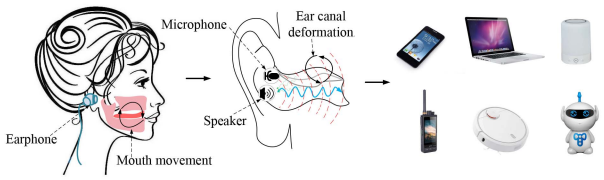


Fig. 1. Conceptual illustration of EarSSR.

The acoustic signals reflected from the ear canal are captured by the microphone[1] and analyzed to obtain the deformation of ear canal during the silent speaking process. Although promising progress has been achieved in acoustic sensing, fine-grained ear canal deformation sensing is still a challenging task due to the following reasons.

- *Extremely subtle deformation of ear canal:* The length of a person's ear canal is only 2–3 cm and the cross-sectional area of the ear canal usually does not exceed 0.7 cm^2. The ear canal deformation caused by silent speech is on the scale of sub-millimeter level. This subtle deformation-caused signal variation is very small and can be easily buried in noise without being detected.
- *The diversity of ear canal structure:* Studies have shown that the geometry of each individual's ear canal is unique. For the same speech, the reflected signals from the ear canal could be different among users. Moreover, people may speak at different speeds with different mouth movement amplitudes for the same speech, leading to differences in ear canal deformation. It is challenging to collect a large amount of silent speech samples to train a deep learning network for silent speech recognition.
- *Non-speech induced ear canal deformation interference:* The third challenge is the interference from non-speech induced ear canal deformation. In addition to speech, non-speech activities such as head movements and mouth movements also cause ear canal deformation.
- *Adapt to new words:* In real-world scenarios, there are new words that our model did not learn. Training the model for each unseen word incurs a high cost. If the trained model can adapt to new words at a low cost, it will greatly facilitate the applicability of silent speech recognition.

To solve the first challenge, we observe that acoustic signals experience complex reflections when traveling inside the ear canal. We discover that the ear canal deformations caused by similar speeches exhibit highly consistent patterns while deformations caused by different speeches are distinct. The reflections which contain the deformation information can thus be utilized for silent speech recognition.

To tackle the second challenge, we quantitatively model the relationship between signal variations and the ear canal deformation caused by speech. We extract the signal variation feature during the process of speech, and apply the Continuous Wavelet Transform (CWT) to obtain the time-frequency spectrum of signal in both time and frequency domains. We further design a data augmentation scheme to expand the size of training data set, which significantly reduces the data collection load and enriches the data diversity (i.e., varying speech speeds and mouth movement amplitudes). Then, we elaborate a two-channel hierarchical neural network, *SsrNet*, to aggregate signal features to extract the difference between speeches for speech recognition.

To deal with the third challenge, based on our analysis, we discover that the extracted speech-induced signal variations exhibit unique patterns compared with the non-speech (e.g., head motions and mouth motions) interference patterns. This motivates us to differentiate the speech-induced and other non-speech induced signal variation patterns by identifying unique features only related to speech. We then extract twelve signal features and utilize the Support Vector Domain Description (SVDD) scheme to find a minimum hypersphere to identify speech-induced signal variations.

To address the issue of recognizing new words, we develop an incremental learning method to adapt the trained network for new words without losing the accuracy of recognizing old words. Instead of retraining the whole network, we inherit the parameters from the original network and fine tune the model to achieve a high accuracy in recognizing both new words and old words at low cost.

The main contributions of EarSSR are summarized as follows.

- We demonstrate the feasibility of using acoustic signals collected from low-cost earphones to obtain ear canal deformation information for silent speech recognition. The proposed system can generalize across users, words and environments.
- We design a two-channel hierarchical network, *SSrNet*, that can effectively extract intra-modality information and fully utilize the complementarity between multiple modalities to achieve good performance in terms of robustness and accuracy at the same time.
- We conduct extensive experiments with 50 volunteers to demonstrate the effectiveness and robustness of the proposed system. Experiment results show that EarSSR can obtain an accuracy of 82% for recognizing a single alphabet letter, and an accuracy of 93% in recognizing words and phrases. The proposed system can still work well in the presence of interference (e.g., when the user is walking).

The rest of this paper is organized as follows. Section II introduces the related work and Section III introduces the background knowledge. Section IV presents the system overview followed by the detailed system design in Section V. Section VI presents the implementation and evaluation of our system. Sections VII and VIII discuss the limitation and conclude the paper.

II. Related Work

In this section, we introduce the literature related to silent speech recognition and wearable based sensing.

A. Silent Speech Recognition

The basic idea of SSR is to enable speech recognition using inaudible information such as EEG [3], EMG [4], [5], mouth

[1] A lot of earphones now have built-in microphone for noise cancellation purposes.

movements [6], [7], [8], [9], and vocal folds [10]. According to the characteristics, existing SSR technologies can mainly be classified into two categories, i.e., contact-based and contact-free approaches.

1) Contact-Based Approaches: Contact-based approaches recognize silent speech via attaching sensors to the target. For example, previous works propose to obtain the cerebral cortex activities via a brain–computer interface (BCI) to predict intended speech [3], [11]. Robin et al. [12] utilize an in-mouth magnetic bead to obtain the tongue and mouth movements for word-level silent speech recognition. JawSense [13] uses the inertial sensors built into smart earphones to detect jaw movement caused by speaking. SottoVoce [14] utilizes ultrasonic imaging sensors attached under the jaw to obtain the images of the jaw to recognize silent speeches. The performance of these sensors is highly location-dependent. Moreover, the sensors could affect the daily activities and may cause skin irritations in long-term use.

2) Contact-Free Approaches: Contact-free approaches either utilize cameras to capture the images of joint movements or leverage wireless signals to sense the movements of lips, throat and chin caused by speaking [7], [8], [18], [19], [20], [21], [22], [23], [24], [25], [26] to achieve silent speech recognition. For example, SpeeChin [18] leverages a customized infrared camera mounted on a necklace to capture the images of neck and face to recognize silent speech. C-Face [22] designs an earphone with two ear-mounted cameras to obtain the images of the face to achieve silent speech recognition. However, camera-based methods require monitoring the target's face, which raises privacy concerns. RFTattoo [23] utilizes RF signals to achieve silent speech recognition, but it requires attaching RFID tags to the target's face. WiHear [24] exploits Wi-Fi signals to perceive the mouth movement-induced changes on the Channel State Information (CSI) to recognize speech. WaveEar [25] employs millimeter waves to sense the vibration of vocal cord when people speak to enhance recognition of speech in noisy environments. EchoWhisper [7] utilizes acoustic signals transmitted from smartphones to extract the Doppler shift induced by mouth movements to achieve silent speech recognition. SoundLip [8] also utilizes the speaker of the smartphone to transmit acoustic signals and leverages the microphone to receive the reflected signals from mouth. EarCommand [27] is the first to implement acoustic sensing on earphones for silent speech recognition. It utilizes the IMU sensor attached on the earphone to detect the motion interference (e.g., head motions). When new users use EarCommand, the model needs to be fine-tuned with a small number of samples.

B. Wearable Based Sensing

Recently, wearable sensing has received a lot of attention [28]. For example, BioFace-3D [29] proposes a 2D facial landmark tracking and 3D facial reconstruction system through in-ear biosensors. OESense [30] utilizes a self-made earphone with the in-ear microphone to realize three applications including step counting, daily activity sensing and hand-to-face gesture recognition. EarGate [31] utilizes the self-made earphone to obtain the bone-conduction gait sound to reliably detect the user's gait for user identification. EBP [32] utilizes an in-ear device to measure blood pressure from the ear canal. WAKE [33] designs a novel behind-the-ear wearable device to detect the bio-signals from the brain, eye movements, facial muscle contractions, and sweat gland activities. CanalScan [34] proposes a tongue-jaw movement recognition system via ear canal deformation sensing using smartphones. It can identify six tongue-jaw movements based on a random forest classifier. Takashi et al. [35] utilize earphones with in-ear speaker and microphone to build a facial expression recognition system. It recognizes 21 facial expressions based on the ear canal features utilizing the support vector machine classifier. EarDynamic [16] proposes an ear canal-based user identification system, which divides phonemes into 6 categories according to the pronunciation positions and each category is linked to one kind of ear canal deformation.

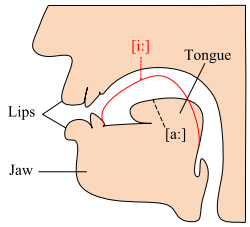


Fig. 2. Tongue and mouth positions of English vowels [15].

III. Background and Principle

In this section, we present the background knowledge about human speech, and analyze the relationship between speech and ear canal deformation. Finally, we model the process of acoustic signal propagation in the ear canal and investigate the feasibility of utilizing acoustic signals to sense the ear canal deformation for silent speech recognition.

A. Human Speech Production and Phonemes

Human speech production system involves three vital physiological organs, i.e., lung, vocal cord, and vocal tract. Sound is resonated and reshaped by the vocal tract which consists of multiple organs, such as nose, mouth (tongue, teeth, and lips), and throat.

The phoneme is the smallest distinctive unit of a language. Vowels and consonants are the two major phoneme categories. Specifically, vowels are produced when vocal cords constrict airflow and the vocal tract is open. Tongue position is the most important physical characteristic that distinguishes one vowel from another [36]. For example, when the tongue moves to the lower right corner and the mouth opens, the vowel /a:/ can be pronounced, and when the tongue moves to the upper left corner and backward, the vowel /i:/ can be pronounced, as shown in Fig. 2 [15]. More generally, we show the vowel chart which includes two dimensions of mouth and tongue movements. Extending or retracting the tongue forward or backward towards the teeth produces a more front or back vowel sound, and lowering or raising the tongue towards the lower jaw or towards the roof

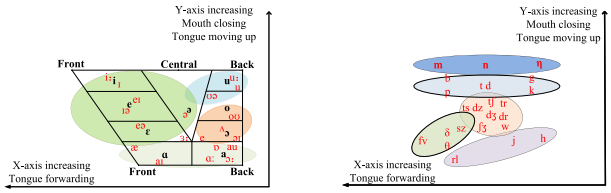


Fig. 3. Vowels and consonants categories based on the mouth and tongue positions [16].

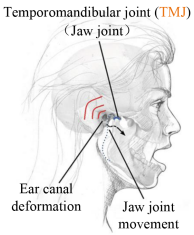


Fig. 4. Temporomandibular joints of the human skull [17].

of the mouth produces a more open or close vowel, as shown in Fig. 3 [16]. The x-axis shows the information of the depth of the tongue forward and backward, and y-axis displays the information of the height of the tongue up and down. Similarly, consonants are also the interaction of the speech organs (i.e., the tongue, lips, and palate). As shown in Fig. 3(b), we also classify the consonants based on the position of the mouth and tongue. From Fig. 3, we can see that the position of the mouth and tongue is an important factor in phoneme production, and each phoneme corresponds to unique and consistent organ movements.

B. The Relationship Between Speech and Ear Canal Deformation

We next describe the relationship between speech and ear canal deformation. Specifically, ear canal deformation is caused by the movement of the temporomandibular joints (*TMJ*) that connect the jaw with the skull as shown in Fig. 4 [17]. When a person is speaking, the mouth movements drive the *TMJ* to move, shifting the ear canal wall. The tongue movements also lead to changes of the ear canal. Researchers [37], [38] discover that when *TMJ* moves, the ear canal volume is changed by almost 20 mm^3, and similar *TMJ* movements cause similar ear canal deformation. Motivated by this observation, we utilize the ear canal deformation to realize silent speech recognition.

C. Sensing Ear Canal Deformation

Measuring the ear canal deformation caused by speech directly is challenging due to the structural complexity and invisibility of the ear canal. In EarSSR, we discover that when acoustic signals are sent into the ear canal, the ear canal deformation causes variations on the reflected signals. This motivated us to explore the feasibility of using acoustic signals to sense the ear canal deformation.

We first model the propagation of acoustic signals in the ear canal. Specifically, when an acoustic wave propagates in a medium without reflective boundary, the incident acoustic wave P_i can be defined as:

$$P_i = P_0 \cos(\omega t - kx + \varphi), \tag{1}$$

where P_0 is acoustic pressure amplitude, ω is angular frequency, k is wave number associated with wavelength, x is the propagation distance, and φ is initial phase. When acoustic wave propagates in ear canal, and is reflected by the ear canal wall, the reflected acoustic wave P_r can be expressed as:

$$P_r = R * P_i exp(-\alpha * d), \tag{2}$$

where R is the reflectance, α indicates the attenuation parameter of acoustic wave in the ear canal, and d is the propagation distance. The attenuation coefficient α includes the thermal attenuation α_t and viscous attenuation α_v [39], [40], which can be calculated as:

$$\alpha_t = \frac{\omega^2}{2\rho_0 c^3}\left(\frac{3}{4}\eta' + \eta''\right), \tag{3}$$

$$\alpha_v = \frac{\omega^2 \chi}{2\rho_0 c^3}\left(\frac{1}{c_v} - \frac{1}{c_p}\right), \tag{4}$$

$$\alpha = \alpha_t + \alpha_v = A(\omega, \tau), \tag{5}$$

η', η'', χ, c_v, c_p, ρ_0 are related to the propagation media and c is sound speed. The reflected acoustic wave P_r is affected by the attenuation coefficient α. Moreover, when a user wears an earphone, the earphone, ear canal, and eardrum would couple together to form a closed space that is extremely sensitive to acoustic pressure changes [41]. We thus can leverage the multipath fading and amplitude variations of the received signals to sense the subtle ear canal deformation.

To investigate the feasibility of utilizing acoustic signals to sense ear canal deformation for silent speech recognition, we conduct experiments to extract signals in different sessions. In each session, we take off the earphones and put them back before our experiment. The results are shown in Fig. 5. These subgraphs represent each of these five vowels (i.e., /a :/, /o :/, /3 :/, /i :/, and /u :/). From the figures, we can see that although there are changes across different sessions, the signal patterns of each phoneme are still very similar. This demonstrates the feasibility of utilizing ear canal deformation to recognize silent speech.

IV. System Overview

The system framework and processing flow are shown in Fig. 6, which consist of the following modules:

- *Signal Design and Processing Modul:* We utilize the in-ear speaker to transmit acoustic signals. The built-in microphone captures the reflected signals from the ear canal. We then synchronize the received signals with transmitted signals using cross-correlation. After synchronization, we remove the direct-path interference signal from the speaker to the microphone to obtain clean ear canal reflection signals.

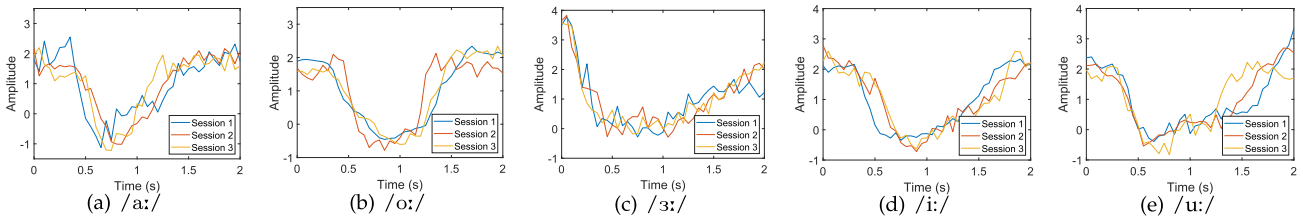


Fig. 5. Extracted ear canal deformation features of the different phonemes.

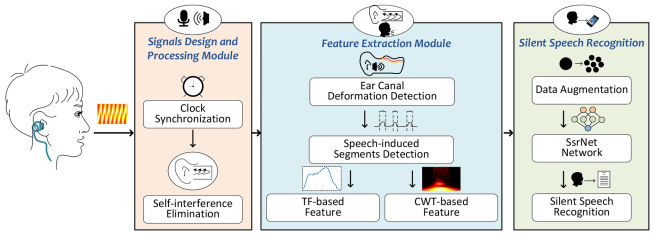


Fig. 6. System overview of EarSSR.

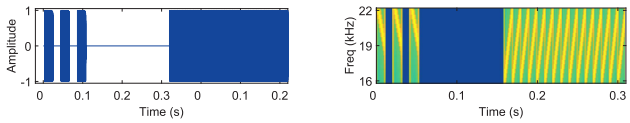


Fig. 7. Transmitted signals in the time and frequency domains.

- *Feature Extraction Module:* In this module, we first design a segmentation algorithm to detect the ear canal deformation events in the reflected signals. We then utilize the SVDD method to find a minimum hypersphere to distinguish between speech-induced and non-speech-induced ear canal deformation events. Finally, we extract the transfer function and CWT features from ear canal reflection signals for silent speech recognition.
- *Silent Speech Recognition:* In this module, we first design a data augmentation scheme to enrich the limited training samples and reduce the cost of data collection. We elaborate a two-channel hierarchical neural network, *SsrNet*, to aggregate features. Finally, we develop an incremental learning method to adapt the trained network for new words.

V. System Design

In this section, we present the design details of each module.

A. Signal Design

In EarSSR, the transmitted signals are designed as FMCW chirps with frequency sweeping from 16 kHz to 22 kHz, and a pseudo-noise (PN) preamble is added at the beginning of the transmitted signals for synchronization, as shown in Fig. 7. As the length of the ear canal is only 2–3 cm, a long chirp may cause the reflection signals to collide with the transmission of the next chirp, and a too short chirp will result in a low SNR, we thus set the chirp length as 10 ms.

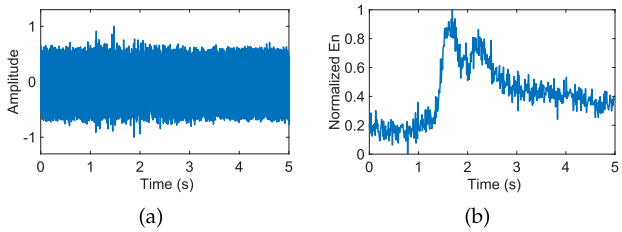


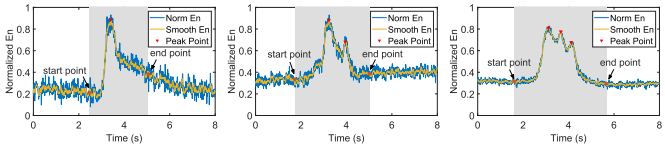
Fig. 8. Example of an event with silent speech: (a) Raw received signals and (b) Short-term energy spectrum.

B. Deformation Event Detection and Segmentation

To effectively recognize the silent speech event in the received signals, EarSSR performs ear canal deformation event detection and segmentation first. Fig. 8(a) plots the received signals when a user speaks. We can see that it is challenging to detect ear canal deformation events in the time domain. We then calculate the energy spectrum of the received signal, as shown in Fig. 8(b). We can see that signal energy varies dramatically in the ear canal deformation region, while in other regions it is relatively flat. Therefore, we can extract the ear canal deformation events based on the short-term energy variation. To segment the signal variation part, we design an ear canal deformation event detection and segmentation algorithm. Specifically, we first calculate the energy (EN) of all the samples within a selected time window as:

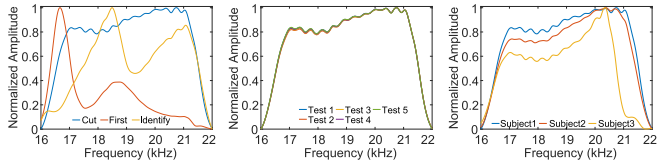
$$EN(t) = \sum_{m=0}^{M} (A^2(t + m * \triangle t)), \tag{6}$$

where $A(t)$ denotes the amplitude reading at time t, m is the sample index, and $\triangle t$ is the sample interval. M denotes the number of samples within the sliding window. We set the sliding window size as 20 ms based on the length of the fundamental speech tone [42]. We then apply a peak detection algorithm [43] on EN to identify peaks. If the detected peak is larger than a predefined threshold, an ear canal deformation event is considered to be detected. The detected peak is denoted as an event peak. We then form an event time window of 20 ms with the event peak as the center point. To determine the start and end points of this event, we first identify the maximum peak within the event



(a) One (b) Answer (c) Information

Fig. 9. Signal variation patterns of different words.



(a) (b) (c)

Fig. 10. TF feature of the ear canal deformation caused by speech: (a) different words; (b) same word; and (c) same word from different subjects.

window. Note that the maximum peak and the event peak may not be the same one. We then take the maximum peak as the initial position to search the start and end points of the event on the left and right sides of the maximum peak respectively. The start and end points are determined when two adjacent peaks have an amplitude difference smaller than a predefined threshold (empirically set as 6% in our implementation) which indicates a relatively stable noise region without speech motion. Note that two adjacent peaks may still cause false positives. We therefore adopt three adjacent peaks and ensure three adjacent peaks are close enough to reduce the chance of early stop. We then perform a linear interpolation [44] operation to make sure each event is the same length.

Fig. 9 shows the detection and segmentation results of the proposed algorithm on 3 words (i.e., "one", "answer", and "information"). The proposed algorithm precisely detects and segments ear canal deformation intervals. In addition, we can also find that the numbers of signal peaks induced by the three words are different because peaks correspond to the word's syllables. While "one" consists of one syllable, "answer" consists of two syllables.

C. Feature Extraction

In this section, our goal is to extract unique and stable speech-induced ear canal deformation features for silent speech identification.

1) Transfer Function-Based Feature Extraction: Transfer function (TF) of the reflected signals can depict the variations of the signal caused by ear canal deformation. We define the transfer function $H(f)$ as:

$$H(f) = \frac{psd(y(t))}{psd(x(t))}, \tag{7}$$

where $psd(y(t)), psd(x(t))$ are the power spectral density of the received signals $y(t)$, and transmitted signals $x(t)$, respectively. We estimate $H(f)$ using Welch's averaged and modified periodogram. We extract the TF feature with 2400 Discrete Fourier Transform (DFT) points and a 20 ms hamming sliding window with 50% overlapping.

Fig. 10(a) illustrates the extracted TF features of different words. We can discover that the TF features of different words exhibit significantly different profiles. Fig. 10(b) shows the TF features of the same word extracted in different rounds. We can see that the TF features of the same word exhibit highly consistent patterns. These results show that the TF feature can be used for silent speech recognition. However, the geometry of each individual's ear canal is unique [41], which causes TF feature differences among different users for the same speech. As shown in Fig. 10(c), we can see that the TF features of the same word are slightly different for different subjects.

2) CWT-Based Feature Extraction: While the TF feature is unique for different words, it varies among subjects. To enhance the robustness, we would like to capture signal feature independent of subjects. In EarSSR, the transmitted FMCW signal $x(t)$ can be represented as:

$$x(t) = \cos\left(2\pi\left(f_0 t + \frac{Bt^2}{2T}\right)\right), \tag{8}$$

where f_0 is the starting frequency, B is the bandwidth and T is the sweep time interval. The received signal is a superposition of all the reflected paths from the ear canal, and each path is an attenuated version of the transmitted signal. Thus, the received signal $y(t)$ can be represented as:

$$y(t) = \sum_{i=1}^{n} \alpha_i \cos\left(2\pi(f_0(t-\tau_i) + \frac{B}{2T}(t-\tau_i)^2\right), \tag{9}$$

where α_i and τ_i are the signal attenuation and the time delay of $i\text{-}th$ path signal.

Then the received signal is multiplied by the transmitted signal to obtain the mixed signal. The mixed signal is passed through a low pass filtering and can be represented as:

$$m(t) = \sum_{i=1}^{n} \alpha_i \cos\left(2\pi\left(\frac{B}{T}\tau_i t + f_0\tau_i - \frac{B}{2T}\tau_i^2\right)\right) \tag{10}$$

In order to obtain the signal variations caused by the ear canal deformation and eliminate static path related to the ear canal structure, we focus on the dynamic vector of the mixed signal, which can be represented as:

$$H_d = \gamma \cos\left(2\pi\left(\frac{B}{T}\tau t + f_0\tau - \frac{B}{2T}\tau^2\right)\right), \tag{11}$$

where γ, τ are the amplitude and time delay of the dynamic path.

Traditional methods utilize the phase to extract the fine-grained variation. Since the speech-induced ear canal deformation is sub-mm level [38], the phase variation can be calculated as: $\varphi = 2\pi f_0 \tau = \frac{2\pi * 16000 * 2 * 0.0001}{343} = 0.02\pi$. Theoretically, the fluctuation of 0.02π is sufficient to be detected. However, we discover that phase information is disturbed by body motions (i.e., respiration, heartbeat, and walking), as shown in Fig. 11(a). We can see that the speech-induced phase variation is just slightly larger than the phase noise. Thus, traditional phase-based methods do not work well for detecting speech-induced

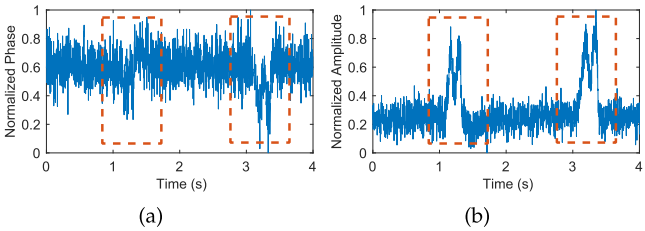


Fig. 11. Extracted features from the mixed signals: (a) phase variation and (b) amplitude variation.

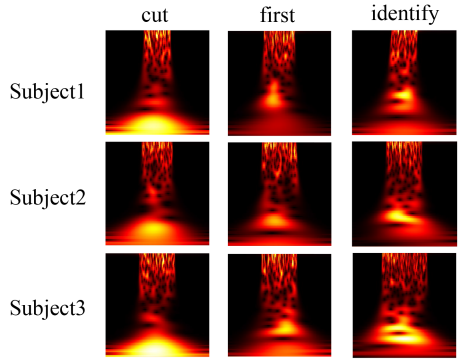


Fig. 12. CWT-based feature from different subjects with different words.

signal variation. Fortunately, the amplitude variations induced by speech are more obvious. As shown in Fig. 11(b), the ear canal deformation caused by speech (inside of the red box) can be clearly observed in the amplitude variations.

To enhance the feature representation, we further apply the Continuous Wavelet Transform (CWT) to the amplitude curve to obtain the spectrum in both time and frequency domains. Specifically, CWT can be calculated as:

$$CWT(f,\phi,\tau) = \{f(t), W_{\phi,\tau(t)}\} = \phi^{-1/2}\int_R f(t)W\left(\frac{t-\tau}{\phi}\right)dt, \quad (12)$$

where $CWT(f,\phi,\tau)$ is the extracted wavelet spectrum, $f(t)$ is the signal function, and $W(\phi,\tau(t))$ is the wavelet base function with ϕ and τ representing the resolution of frequency and time domain respectively. The wavelet base function affects the result of the wavelet transform. In EarSSR, we apply *generalized morse wavelets* as the wavelet base function due to their high resolution in the low-frequency and time domain. Fig. 12 shows the CWT-based features of three words for three subjects. We can see that the CWT features of the same word are similar for different subjects. It shows that the extracted features is not affected by subject diversity.

The Difference Between CWT and TF Features: The TF feature characterizes the channel response of different frequencies, i.e., the signal power distribution over different frequencies. It depends on the geometry of the ear canal and can, therefore, reflect the deformation of the ear canal caused by speeches. CWT feature describes the power of the signal at different timestamps and at different frequencies. In our design, we perform CWT

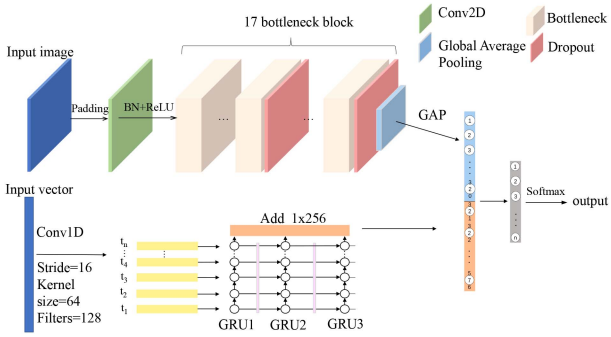


Fig. 13. Architectural diagram of *SsrNet*.

operation after removing the carrier frequency and therefore the obtained CWT feature characterizes the vibration property of the ear canal induced by speech which is dependent on the speech and independent of the ear canal structure.

D. SsrNet *Network Model*

In this subsection, our goal is to fuse the above two features to achieve high robustness and high accuracy for silent speech recognition. The above two features have different dimensions, i.e., *TF* feature is a 1D vector while *CWT* feature is a 3D image. The two features can be used to describe the same word and provide complementary information.

We propose *SsrNet*, a two-channel hierarchical model to effectively fuse the two features. We apply Recurrent Neural Network (RNN) [45], [46], [47] on the TF feature to capture the long-term dependencies in sequences. We then apply Convolutional Neural Network (CNN) [48], [49] to extract information from the CWT image feature. Finally, the features from the two channels are concatenated for recognition.

The model architecture is shown in Fig. 13. For the first channel, we design the network structure based up MobileNet V3 [50]. It is a lightweight network that can be deployed on mobile devices. The input of the first channel is the 224 $\times$ 224 $\times$ 3 spectrogram image. We add 17 residual bottlenecks after the Conv2D layer, apply the attention block in each bottleneck, and add the max pooling weights on the attention block, which can automatically focus more attention on the useful feature regions. To enhance the capability of the network to extract fine-grained features from the image, we add two channel-attention blocks on the bottleneck blocks. The global average pooling layer (GAP) is employed to output the extracted features. For the second channel, it is a recursive network and the input of the second channel is the 6000 $\times$ 1 vector sequence. The gated recurrent unit (GRU) is chosen as the basic unit, including the update gate and reset gate used to capture the dependencies in the sequence, and outputs the extracted features.

We add batch normalization (BN) at the convolutional layer and apply dropout after the fully connected layer to avoid overfitting. We then add a global average pooling layer and a fully connected layer to map the feature representation F into a new space W. Finally, the probability of different word categories $\hat{y}$ could be calculated from the softmax layer, and we utilize a

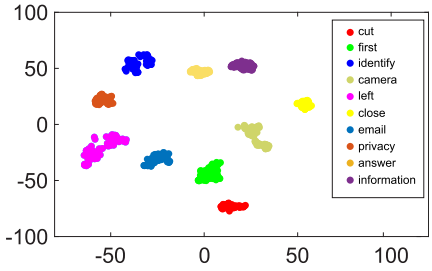


Fig. 14. t-SNE visualization of two combined features.

loss function to compute the similarity between the predicted label $\hat{y_i}$ and the ground-truth y_i. The loss function $L(W)$ can be calculated as:

$$L(W) = -\frac{1}{N_b}\sum_{i=1}^{N_b} y_i * log(\hat{y_i}) + k * R, \tag{13}$$

where N_b is the size of the batch, R is a regularization process and k is the hyper-parameter.

We train the model on the training set with batch size 80, and we utilize the adam optimizer with 0.001 initial learning rate to minimize the loss function. By applying the proposed method, we visualize the learned features for ten words from three subjects with t-SNE [51], as shown in Fig. 14. We can see that different words from different subjects have distinct feature distributions. It illustrates that *SsrNet* can effectively fuse features from different modalities for silent speech recognition.

E. Discriminating Speech and Non-Speech Induced Ear Canal Deformation

We oberve that head and mouth motions also lead to ear canal deformations, resulting in signal variations. It is thus necessary to recognize whether the ear canal deformation is speech-induced or non-speech-induced.

As shown in Fig. 15, we ask three volunteers to perform the following experiments: (1) opening and closing the mouth; (2) shaking the head left and right; (3) speaking without vocalization. We then extract the amplitude feature from the mixed signals, as introduced in Section V-C2. We discover that these activities also cause variations of the signal. But a key observation is that the extracted speech-induced signal variations exhibit unique patterns, such as more peaks and valleys, compared with the non-speech interference patterns caused by head motions and mouth motions. This motivates us to differentiate the speech-induced and other non-speech-induced signal variation patterns by identifying unique features related to speech. We then extract twelve signal features, such as *peak to peak value, the number of peaks, and the length*, as shown in Table I. We then utilize the SVDD algorithm [52] to find a minimum hypersphere to identify speech-induced signal variations. When an ear canal deformation event is in the hypersphere, EarSSR determines that the signal variation is speech-induced.

TABLE I
TWELVE KINDS OF STATISTICAL FEATURES

Name	Feature
Statistical features	(1) Peak-to-peak value, (2) The number of peaks, (3) Length, (4) Standard deviation, (5) Mean absolute deviation, (6) Root mean square, (7) Interquartile range, (8) Entrop, (9) Skewness, (10) Kurtosis, (11) Impulse factor, (12) Form factor

F. Data Augmentation

Because the speaking behaviors (e.g., speech speed, and mouth movement amplitude) are user-dependent, we thus design a data augmentation scheme to expand the size of training data set. This process significantly enriches the data diversity and reduces the data collection load. Inspired by the data augmentation method in the speech recognition domain [53], we design the data augmentation scheme for our two different forms of features (vector feature and image feature). For the TF feature (vector) X_f, we apply the *jitter* method, which adds different categories of noise to the original data to enhance the ability of the model against noise. For the CWT feature, we apply the augmentation scheme proposed by Park et al. [54], which consists of three key steps, i.e., time warping, time masking and time-frequency masking.

Time Warping: Given the CWT spectrum with a time length of T, we view it as an image where the horizontal axis indicates time and the vertical axis indicates frequency. We then stretch the spectrum with the time length of $(W, T-W)$ to a larger time length $(W-d, T-W+d)$ as shown in Fig. 16(b) with $(0 < d < W)$ and $(W < T/2)$.

Time Masking: Time masking is applied to mask a time window $[t_i, t_i + \triangle t]$ with $\triangle t$ as the masked time interval and $(0 < t_i < T - \triangle t)$, as shown in Fig. 16(c).

Time-Frequency Masking: Time-frequency masking means that time masking and frequency masking occur at the same time. Frequency masking is applied to the frequency range $[f_i, f_i + \triangle f]$, with $\triangle f$ as the masked frequency range, F as the total frequency range and $(0 < f_i < F - \triangle f)$, as shown in Fig. 16(d).

G. Accommodating New Words

To accommodate more words, a common method is to retrain the whole model including the old words and new words. However, retraining comes with a large time cost. To solve the problem, we employ an incremental learning method [55], where the input data is continuously used to extend the existing model's knowledge. Instead of retraining the whole network, we inherit the parameters from the original network and fine tune the model under the old-model constraints to achieve high accuracy in recognizing both new words and old words at low cost.

Algorithm 1 illustrates how this process works in detail. First, we compute Y_o, which is the output of the original network for new words on the old parameters (defined by θ_s, and θ_o). Next, we train the model to minimize loss function for all words and regularize R using stochastic gradient descent. We freeze θ_s

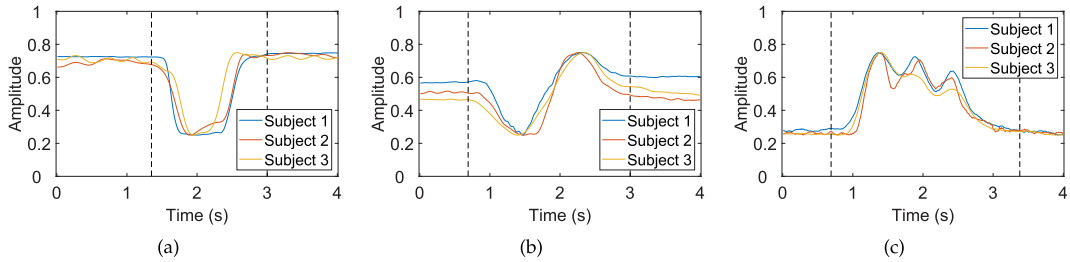


Fig. 15. Example of non-speech-induced signal variations and speech-induced signal variations: (a) Opening and closing mouth; (b) shaking the head from left and right; and (c) speaking.

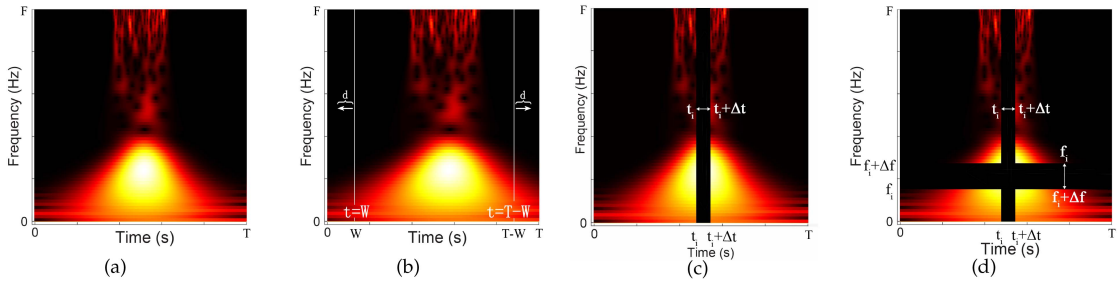


Fig. 16. Data augmentation. (a) Raw; (b) time warping; (c) time masking; and (d) time-frequency masking.

Algorithm 1: Accommodating New Words.

Start with:

θ_s : shared parameters

θ_o: specific parameters for old model

X_n: training samples of the new words

Y_n: labels of the new words

Initialize:

$Y_o \leftarrow SsrNet(X_n, \theta_s, \theta_o)$ // compute output of old model for new word

$\theta_n \leftarrow randinit(|\theta_n|)$ // randomly initialize new parameters

Train:

Define $\hat{Y_o} \equiv SsrNet(X_n, \hat{\theta_s}, \hat{\theta_o})$ // old model output

Define $\hat{Y_n} \equiv SsrNet(X_n, \hat{\theta_s}, \hat{\theta_n})$ //new model output

$\theta_s^*, \theta_o^*, \theta_n^* \leftarrow argmin_{\hat{\theta_s}, \hat{\theta_o}, \hat{\theta_n}} (\lambda_o L_{old}(Y_o, \hat{Y_o}) + L_{new}(Y_n, \hat{Y_n}) + R(\hat{\theta_s}, \hat{\theta_o}, \hat{\theta_n}))$

and θ_o and train θ_n to convergence. Then we jointly train all weights θ_s, θ_o, and θ_n until convergence. For new words, the loss encourages predictions $\hat{Y_n}$ to be consistent with the ground truth Y_n. For old words, we want the output probabilities for each word to be close to the recorded output from the original model.

VI. Evaluation

In this section, we first introduce the experiment implementation and dataset. Then a series of experiments are conducted to evaluate the generalization and robustness of EarSSR. Finally, we evaluate the effect of sensing signal on user's comfortableness.

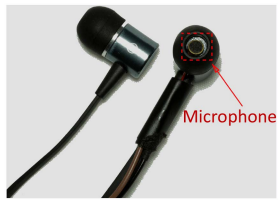


Fig. 17. Prototype earphone in EarSSR.

A. Experiment Implementation

There are several commercial earphones equipped with in-ear microphones (e.g., Apple Airpods Pro [56] and Bose QuietComfort [57]). Due to the hardware restriction, we can not obtain received signals from the in-ear microphone. Thus, we design EarSSR by attaching a cheap low-end microphone chip (10 cents) to the front side of the earphone's speaker, as shown in Fig. 17. In addition, we equip the earphones with sponge earplugs to ensure that the earphone can fit snugly in the ear canal. We connect the earphones and devices (i.e., smartphone and laptop) via a 3.5 mm audio interface to control the signal transmission and recording.

B. Data Collection

We recruit a total of 50 volunteers (36 males and 14 females in the range of 20 to 57 years old) and spend more than three months collecting the data. Among these volunteers, three volunteers are

TABLE II
COMMANDS SET: ENGLISH COMMAND SET

Category	Command
Digits	(1) Zero, (2) One, (3) Two, (4) Three, (5) Four, (6) Five, (7) Six, (8) Seven, (9) Eight, (10) Nine
Interactive Commands	(11) Answer, (12) Camera, (13) Close, (14) Copy, (15) Cut, (16) Help, (17) Keyboard, (18) Next, (19) Open, (20) Skype, (21) Undo
Voice Assistant	(22) Alexa, (23) Ok Google, (24) Hey Siri
Navigation	(25) Left, (26) Right, (27) Up, (28) Down, (29) East, (30) South, (31) West, (32) North
Privacy	(33) Secret, (34) Information, (35) Email, (36) Privacy, (37) Identify, (38) Money
High Frequency	(39) Please, (40) What, (41) Nice, (42) How, (43) Are, (44) Enjoy, (45) They, (46) Meet, (47) Thanks, (48) Really, (49) Ours, (50) You

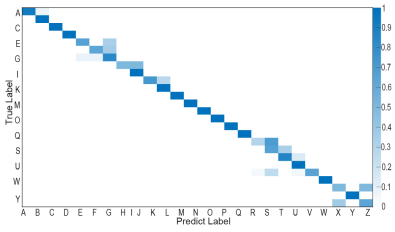


Fig. 18. Confusion matrix of letter-level recognition in EarSSR.

native English speakers, five volunteers major in broadcasting and received professional pronunciation training with fluent oral English, three volunteers do not use English frequently, and the rest are non-native English speakers who can speak English fluently. During our experiment process, we ask the users to wear the earphones tightly to mitigate the issue of earphone movements. The experiments were IRB-approved by the host university.

We thus collect two silent speech datasets: a letter-level dataset and a word-level dataset. The letter-level dataset contains 26 alphabets and the word-level dataset contains 50 words and phrases which can be categorized into six groups (i.e., number 1–10, 11 interactive commands, 3 voice assistant commands, 8 navigation commands, 6 privacy-related words, and 12 high-frequency words), as shown in Table II. These volunteers speak each word/phrase 10 times in one group and 10 groups of data are collected in total. In addition, volunteers are asked to take off and put back on the earphones between groups to simulate the behavior of wearing earphones in daily life. We divide the collected data into training set, validation set, and test set according to the percentage of 60%, 20%, and 20%. We perform 10× data augmentation on training sets. We adopt accuracy, precision, recall, and F1-score as the evaluation metrics.

C. Overall Performance

We first evaluate our system's performance on recognizing letters and words, and then verify the effectiveness of the feature fusion and network designs.

1) Accuracy of Letter-Level Silent Speech Recognition: We first evaluate the ability of our system to recognize the 26 alphabet letters. The confusion matrix is shown in Fig. 18. The overall recognition accuracy is about 82%. It is challenging to achieve letter-level recognition since letters have fewer syllables and some letters have similar mouth movements (i.e., "e" and "g", "s" and "x", etc).

TABLE III
AVERAGE ACCURACY FOR DIFFERENT NUMBERS OF SYLLABLES

	One-syllable	Two-syllable	Multi-syllable	Total
No.words	13	19	18	50
Avg.accuracy	84%	95%	98%	93%

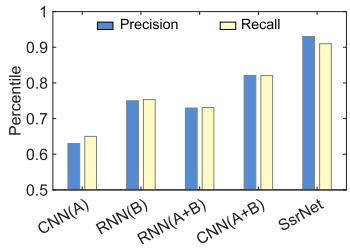


Fig. 19. Performance of feature fusion (A: CWT feature; B: TF feature).

2) Accuracy of Word-Level Silent Speech Recognition: For word-level recognition, we employ 50 commonly used words and phrases as shown in Table II. The overall recognition accuracy is about 93%. We further group the words/phrases based on the number of syllables and show the recognition accuracy in Table III. We can see we achieve a higher accuracy with multi-syllable words. This is because the multi-syllable words contain more mouth and tongue movements, which bring richer ear canal deformation features for higher recognition accuracy.

3) Performance of Feature Fusion: To show the effectiveness of the proposed feature fusion scheme, we conduct experiments to compare the recognition performance of *SsrNet* with only one feature or raw fusion of two features on the word-level dataset.

Specifically, we abbreviate the CWT feature as "A", and TF feature as "B" and design five different strategies to evaluate the performance of feature fusion: (1) apply CNN to A, denoted as "CNN(A)"; (2) apply RNN to B, denoted as "RNN(B)"; (3) raw concatenation: apply RNN to the two features, denoted as "RNN(A+B)"; (4) raw concatenation: apply CNN to the two features, denoted as "CNN(A+B)"; (5) *SsrNet*. The results are shown in Fig. 19. When only CWT feature or only TF

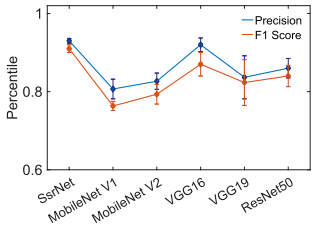


Fig. 20. Comparison of *SsrNet* with other CNN model.

TABLE IV
NUMBER OF PARAMETERS IN DIFFERENT NETWORKS

Networks	Parameters (Million)
SsrNet (ours)	2.44
MobileNet V1	3.23
MobileNet V2	2.26
VGG 16	14.71
VGG 19	20.02
ResNet 50	23.59

feature is employed, the achieved precision is about 60% and 75% respectively. It is interesting to see that the recognition performance of "RNN(A+B)" is even worse than applying one feature "RNN(B)" alone. It indicates that RNN network does not fuse the two features well. CNN (A+B) is more effective than CNN (A) but the precision is still below 85%. The proposed *SsrNet* achieves a precision close to 95% which shows that the proposed *SsrNet* can efficiently fuse the TF and CWT features.

4) Performance Over Different Network Structures: In *SsrNet*, the first channel is a CNN network. We evaluate the performance of EarSSR based on five different CNN network structures (i.e., mobileNet V1 [58], mobileNet V2 [59], VGG-16 [60], VGG-19 [61], and ResNet-50 [62]). We fine-tune the five pre-trained CNN networks using our dataset with 50 words. Each network structure is trained 3 times and the network is tuned for 200 epochs each time to ensure convergence. Fig. 20 shows that the proposed *SsrNet* outperforms the other networks. Among the 5 networks, VGG-16 also achieves an accuracy around 90%. But as shown in Table IV, *SsrNet* is trained with much fewer parameters (2.44 million) compared to VGG-16 (14.71 million).

D. System Robustness

We now evaluate the robustness and reliability of EarSSR.

1) Subject Diversity: We use data collected from 40 subjects to train the model and then apply the trained model on 10 new subjects (Person Index 41–50). As shown in Fig. 21, EarSSR achieves good performance on new users. The average accuracy is above 80% and 90% for word-level and letter-level datasets, respectively. These results show that our system generalizes well to new users and we do not need to train all the users.

2) The Impact of Data Augmentation: We now evaluate the impact of the data augmentation ratio. The ratio of augmentation varies from 0 to 20 times at a step size of 5. From Fig. 22, we can see that without data augmentation, the precision is only about 60% and 75% for letter-level and word-level recognition.

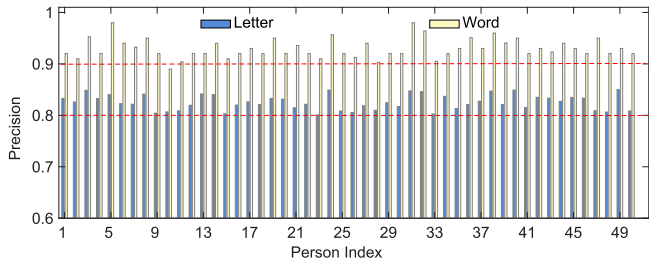


Fig. 21. Recognition performance of letters and words with different subjects.

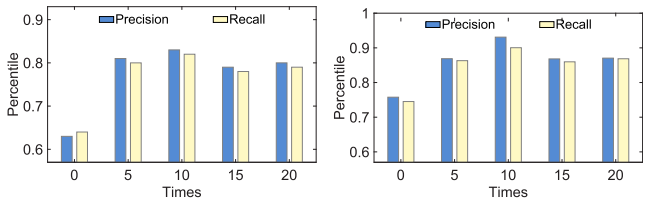


Fig. 22. Performance of data augmentation method.

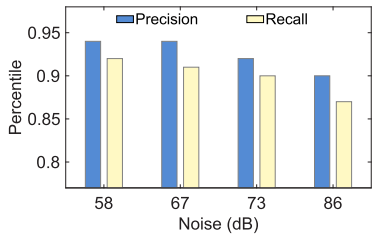


Fig. 23. Recognition accuracy under different ambient noise.

We further found that a data augmentation ratio of 10 was the most effective in improving the recognition precision. If the data augmentation ratio is further increased, the accuracy even decreases. This is because too much generated data can compromise the features contained in the data.

3) Resistance to Environmental Noise: We evaluate the performance of EarSSR in different environments. We choose four environments with different noise levels that are 58 dB, 67 dB, 73 dB and 86 dB. The result is shown in Fig. 23. We can see that noise does slightly affect the performance. However, even in the very noisy environment (86 dB), reasonably good performance (a precision higher than 90%) can still be achieved.

4) Robustness on Earphone Wearing Positions: We first conduct experiment to collect measurements when the sponge earplugs are removed and the earphones are inserted shallowly. The results are shown in Fig. 25(a). We can see that the precision decreases from 93% to 72% under shallow conditions without the sponge to constraint the earphones' movements. Note that earphone manufactures also include silicon ear tips of different sizes to fit ear canals of different sizes. Apple includes ear tips of four different sizes to ensure the earphone is always tightly and comfortably worn in our ear canal. Thus, to alleviate the impact of earphone movement during speaking, it is better to insert the earphones in a deep location and equip earphones with

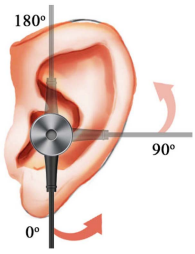


Fig. 24. Earphone wearing positions.

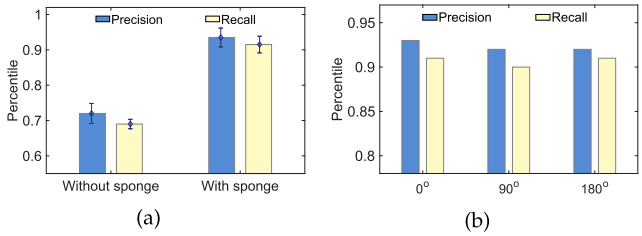


Fig. 25. Recognition performance with different earphone wearing positions: (a) Depths and (b) angles.

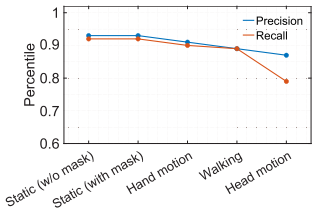


Fig. 26. Performance comparison for different body motions.

sponge/silicon earplugs. Moreover, we evaluated three earphone wearing angles (0°, 90°, 180°) as shown in Fig. 24. As shown in Fig. 25(b), EarSSR still works well under other wearing angles.

5) Reliability on Body Motions and Face Occlusion: The user's body movements introduce interference to the ear canal's reflection signals. Moreover, mask wearing has become common in public due to the pandemic of COVID-19. To evaluate the robustness of EarSSR against body motions and face occlusion, we examine EarSSR under different body states: static (without facial mask), static (with facial mask), hand motion (arms raised and swung left and right), walking (at a speed of 2 km/h on a treadmill) and head motion (shaking head left and right). The result is shown in Fig. 26, we can see that static (with facial mask), hand motion and walking have limited impacts on the performance of EarSSR. The head motion does have a slight effect on the system performance. This is because the head motion also alters the deformation of ear canal.

6) Speech-Induced Ear Canal Deformation Detection: We then evaluate the performance of our system to discriminate between speech-induced and non-speech-induced (i.e., caused by head motions and mouth motions) ear canal deformation. For collecting the non-speech-induced ear canal deformation data, we ask the volunteers to shake their heads from left to right, or open and close their mouths. Volunteers do not speak during this process. The training set is composed of 50% speech-induced data and 50% non-speech-induced data. The rest is the test set. The result is shown in Table V. We can discover that 95.08% of the speech-induced ear canal deformation events are correctly detected. It illustrates that EarSSR can effectively detect speech-induced events.

TABLE V
SPEECH-INDUCED EAR CANAL DEFORMATION DETECTION

Activity	Precision	Recall
Speech-induced	95.08%	94.79%

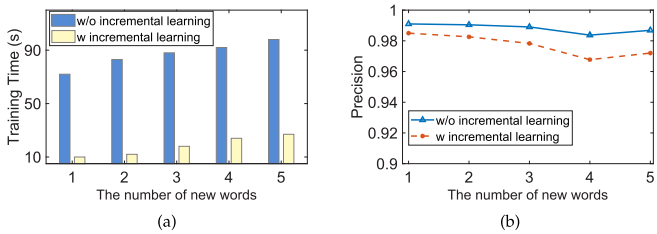


Fig. 27. Performance of incremental learning scheme: (a) training time and (b) accuracy.

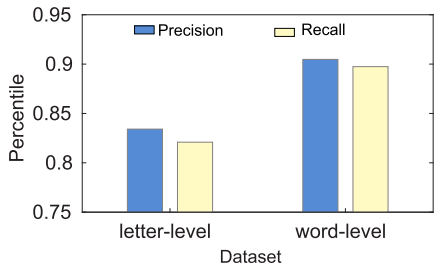


Fig. 28. Recognition performance when users listen to music.

7) Performance on New Words: We evaluate the effectiveness of the proposed incremental learning algorithm on new words. We first initialize our *SsrNet* model by training data of ten words. Then we add various numbers of new words (1 to 5) to evaluate the system performance with and without applying the incremental learning method. Fig. 27(a) presents the training time of our learning model. It manifests that the incremental learning method help reduce training time remarkably. Fig. 27(b) shows the average identification prevision for different numbers of new words. The precision with incremental learning is always above 97% and is comparable to that achieved without incremental learning.

8) Impact of Music Play: Listening to music is the key function of earphones. We now evaluate if the proposed system can recognize silent speech when users are listening to music. We ask the volunteers to conduct silent speech experiments and at the same time listen to music using the same earphones. The evaluation results are shown in Fig. 28. We can see that EarSSR has a precision above 80% on letter and over 90% on word recognition. This is basically the same as that without music listening. This is because EarSSR transmits FMCW signal on the frequency range of 16 kHz–22 kHz which is higher than

TABLE VI
RUNNING TIME OF EARSSR

Item	Time cost
Network latency	0.045 s
Data preprocessing	0.18 s
Feature extraction	0.3 s
SsrNet recognition	0.1 s
Total	0.625 s

music (below 10 kHz). The music can be removed by a bandpass filter and thus playing music in earphones does not affect the recognition performance of EarSSR.

E. System Overhead and Latency

1) System Latency: We implement an app on Samsung Galaxy S5 and employ a host PC with an AMD Ryzen 7 5800X CPU, 32 GB memory, and a Nvidia TITAN Xp GPU as the cloud server. The app is used for data collection and result output. The feature extraction module and the *SsrNet* model are implemented on the server. The system latency is defined as the time difference between the user finishing the input and EarSSR outputting the result to the user. The latency is composed of four parts, i.e., network latency, data processing, feature extraction, and *SsrNet* recognition. Table VI shows the total latency of EarSSR, i.e., 0.625 s. We notice that the feature extraction module is the most time-consuming component (0.3 s). The time of data processing is 0.18 s, and the recognition time of *SsrNet* is 0.1 s. Moreover, our *SsrNet* model is a lightweight network which can be deployed on mobile devices.

2) System Power Consumption: We connect the Samsung Galaxy S5 and our earphone via a 3.5 mm audio interface. We deploy PowerTutor [63] app on the smartphone to measure the power consumption of EarSSR. The results show that the EarSSR consumes an average of 23.1 J of energy per minute, which means that the power consumption of the EarSSR is 385 mW. This power consumption is equivalent to a voice call. Moreover, we discover that signal transmission is the most power hungry part of our system. We can thus design power control schemes, such as triggering signal transmission only when the user is speaking to save power consumption [64].

F. User Experience Survey

In this section, we evaluate the effect of sensing signal on user's comfortableness. We adopt a questionnaire-based method, which is commonly used in social and health sciences [65]. We invited 50 volunteers to wear the earphone and participate in our experiment. We asked the volunteer to rate their experience on the scale of 1 to 5 with 1 indicating "Absolutely not uncomfortable" and 5 indicating "Extremely uncomfortable". The detailed scales are shown in Table VII. We transmit the signal for one minute and chirps with a length of 10 ms are transmitted continuously. The signal volume is set as 6% of the maximum volume.

The result is also shown in Table VII, we can see that 8% of volunteers report absolutely not feeling uncomfortable and 84% of volunteers report they are basically not feeling uncomfortable. On the other hand, 6% of volunteers report they feel slightly uncomfortable and 2% of volunteers report that they feel uncomfortable. To address this issue, we embed the sensing signal into music [64] and transmit the combined signal to mitigate the negative effect of sensing signal. The new result is also presented in Table VII. We can see that when we embed the small sensing signal into music, no volunteers report uncomfortable. We believe embedding sensing signals into music is an efficient method to improve user experience with acoustic sensing.

TABLE VII
DOES CONTINUOUS OR REPEATED ACTION OF PLAYING SOUND CAUSE YOU UNCOMFORTABLE?

Score	Explanation	Sensing signals	With music
1	Absolutely not	8%	100%
2	Basically not	84%	0%
3	Slightly	6%	0%
4	Very much	2%	0%
5	Extremely	0%	0%

VII. LIMITATION

In this section, we will discuss the limitations of EarSSR and potential future work.

- *Implementation on wireless earphones:* The proposed system is currently implemented on wired earphones. This is because wireless earphones support the use of speaker and microphone at the same time only in the hands-free mode. However, in this mode, due to the transmission capability of Bluetooth, the audio sampling rate is limited to 16 kHz and therefore can not support ultrasound signal transmission.
- *Sentence-level Recognition:* This work focused on letter-level and word-level recognition. This is also the basic unit for sentence-level recognition. To achieve sentence-level recognition, the extra component is to segment the sentence into words and there is a rich literature [66], [67], [68] on this. We plan to employ advanced speech segmentation algorithms (e.g., attention-based encoder-decoder network [69]) to improve the accuracy of sentence-level recognition in our future work.

VIII. CONCLUSION

In this paper, we present EarSSR, the earphone-based silent speech recognition system. We utilize the earphone to transmit inaudible acoustic signals to obtain the fine-grained ear canal deformation for silent speech recognition. We conduct extensive experiments to demonstrate the effectiveness and robustness of the proposed system.

REFERENCES

[1] Speech and Voice Recognition Market, 2017. [Online]. Available: https://www.marketsandmarkets.com/Market-Reports/speech-voice-recognition-market-202401714.html/
[2] Advantages of Voice User Interfaces, 2022. [Online]. Available: https://www.speechly.com/blog/advantages-of-voice-user-interfaces/

[3] J. S. Brumberg, A. Nieto-Castanon, P. R. Kennedy, and F. H. Guenther, "Brain–computer interfaces for speech communication," *Speech Commun.*, vol. 52, no. 4, pp. 367–379, 2010.

[4] T. Schultz, "ICCHP keynote: Recognizing silent and weak speech based on electromyography," in *Proc. Int. Conf. Comput. Handicapped Persons*, Springer, 2010, pp. 595–604.

[5] M. Wand, T. Schultz, and J. Schmidhuber, "Domain-adversarial training for session independent EMG-based speech recognition," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 3167–3171.

[6] K. Sun, C. Yu, W. Shi, L. Liu, and Y. Shi, "Lip-Interact: Improving mobile device interaction with silent speech commands," in *Proc. 31st Annu. ACM Symp. User Interface Softw. Technol.*, 2018, pp. 581–593.

[7] Y. Gao, Y. Jin, J. Li, S. Choi, and Z. Jin, "EchoWhisper: Exploring an acoustic-based silent speech interface for smartphone users," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 3, pp. 1–27, 2020.

[8] Q. Zhang, D. Wang, R. Zhao, and Y. Yu, "SoundLip: Enabling word and sentence-level lip interaction for smart devices," in *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 1–28, 2021.

[9] H. A. C. Maruri, P. Lopez-Meyer, J. Huang, W. M. Beltman, L. Nachman, and H. Lu, "V-Speech: Noise-robust speech capturing glasses using vibration sensors," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 2, no. 4, pp. 1–23, 2018.

[10] T. Hirahara et al., "Silent-speech enhancement using body-conducted vocal-tract resonance signals," *Speech Commun.*, vol. 52, no. 4, pp. 301–313, 2010.

[11] C. Herff and T. Schultz, "Automatic speech recognition from neural signals: A focused review," *Front. Neurosci.*, vol. 10, 2016, Art. no. 429.

[12] R. Hofe et al., "Small-vocabulary speech recognition using a silent speech interface based on magnetic sensing," *Speech Commun.*, vol. 55, no. 1, pp. 22–32, 2013.

[13] P. Khanna, T. Srivastava, S. Pan, S. Jain, and P. Nguyen, "JawSense: Recognizing unvoiced sound using a low-cost ear-worn system," in *Proc. 22nd Int. Workshop Mobile Comput. Syst. Appl.*, 2021, pp. 44–49.

[14] N. Kimura, M. Kono, and J. Rekimoto, "SottoVoce: An ultrasound imaging-based silent speech interaction using deep neural networks," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2019, pp. 1–11.

[15] L. Zhang, S. Tan, J. Yang, and Y. Chen, "VoiceLive: A phoneme localization based liveness detection for voice authentication on smartphones," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 1080–1091.

[16] Z. Wang, S. Tan, L. Zhang, Y. Ren, Z. Wang, and J. Yang, "EarDynamic: An ear canal deformation based continuous user authentication using in-ear wearables," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 5, no. 1, pp. 1–27, 2021.

[17] T. D. T. Devices, 2020. [Online]. Available: https://www.fda.gov/medical-devices/dental-devices/temporomandibular-disorders-tmd-devices/

[18] R. Zhang et al., "SpeeChin: A smart necklace for silent speech recognition," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 5, no. 4, pp. 1–23, 2021.

[19] Y. M. Assael, B. Shillingford, S. Whiteson, and N. De Freitas, "LipNet: End-to-end sentence-level lipreading," 2016, *arXiv:1611.01599*.

[20] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An audio-visual corpus for speech perception and automatic speech recognition," *J. Acoustical Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424, 2006.

[21] J. S. Chung and A. Zisserman, "Lip reading in the wild," in *Proc. Asian Conf. Comput. Vis.*, Springer, 2016, pp. 87–103.

[22] T. Chen, B. Steeper, K. Alsheikh, S. Tao, F. Guimbretière, and C. Zhang, "C-Face: Continuously reconstructing facial expressions by deep learning contours of the face with ear-mounted miniature cameras," in *Proc. 33rd Annu. ACM Symp. User Interface Softw. Technol.*, 2020, pp. 112–125.

[23] J. Wang et al., "RFID Tattoo: A wireless platform for speech recognition," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 3, no. 4, pp. 1–24, 2019.

[24] G. Wang, Y. Zou, Z. Zhou, K. Wu, and L. M. Ni, "We can hear you with Wi-Fi!," *IEEE Trans. Mobile Comput.*, vol. 15, no. 11, pp. 2907–2920, Nov. 2016.

[25] C. Xu et al., "WaveEar: Exploring a mmWave-based noise-resistant speech sensing for voice-user interface," in *Proc. 17th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2019, pp. 14–26.

[26] Y. Zhang et al., "Endophasia: Utilizing acoustic-based imaging for issuing contact-free silent speech commands," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 4, no. 1, pp. 1–26, 2020.

[27] Y. Jin et al., "EarCommand: "Hearing" your silent speech commands in ear," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 6, no. 2, pp. 1–28, 2022.

[28] T. Röddiger et al., "Sensing with earables: A systematic literature review and taxonomy of phenomena," in *Proc. ACM Interactive Mobile Wearable Ubiquitous Technol.*, vol. 6, no. 3, Sep. 2022, Art. no. 135. [Online]. Available: https://doi.org/10.1145/3550314

[29] Y. Wu, V. Kakaraparthi, Z. Li, T. Pham, J. Liu, and P. Nguyen, "BioFace-3D: Continuous 3D facial reconstruction through lightweight single-ear biosensors," in *Proc. 27th Annu. Int. Conf. Mobile Comput. Netw.*, 2021, pp. 350–363.

[30] D. Ma, A. Ferlini, and C. Mascolo, "OESense: Employing occlusion effect for in-ear human sensing," in *Proc. 19th Annu. Int. Conf. Mobile Syst. Appl. Serv.*, 2021, pp. 175–187.

[31] A. Ferlini, D. Ma, R. Harle, and C. Mascolo, "EarGate: Gait-based user identification with in-ear microphones," 2021, *arXiv:2108.12305*.

[32] N. Bui et al., "eBP: A wearable system for frequent and comfortable blood pressure monitoring from user's ear," in *Proc. 25th Annu. Int. Conf. Mobile Comput. Netw.*, 2019, pp. 1–17.

[33] N. Pham et al., "WAKE: A behind-the-ear wearable system for microsleep detection," in *Proc. 18th Int. Conf. Mobile Syst. Appl. Serv.*, 2020, pp. 404–418.

[34] Y. Cao, H. Chen, F. Li, and Y. Wang, "CanalScan: Tongue-Jaw movement recognition via ear canal deformation sensing," in *Proc. IEEE Conf. Comput. Commun.*, 2021, pp. 1–10.

[35] T. Amesaka, H. Watanabe, and M. Sugimoto, "Facial expression recognition using ear canal transfer function," in *Proc. 23rd Int. Symp. Wearable Comput.*, 2019, pp. 1–9.

[36] J. P. Olive, A. Greenwood, and J. Coleman, *Acoustics of American English Speech: A Dynamic Approach*, Berlin, Germany: Springer, 1993.

[37] M. J. Grenness, J. Osborn, and W. L. Weller, "Mapping ear canal movement using area-based surface matching," *J. Acoustical Soc. Amer.*, vol. 111, no. 2, pp. 960–971, 2002.

[38] C. Pirzanski and B. Berge, "Ear canal dynamics: Facts versus perception," *Hear. J.*, vol. 58, no. 10, pp. 50–52, 2005.

[39] S. L. Garrett, *Understanding Acoustics: An Experimentalist's View of Sound and Vibration*, Berlin, Germany: Springer, 2020.

[40] L. E. Kinsler, A. R. Frey, A. B. Coppens, and J. V. Sanders, *Fundamentals of Acoustics*, Hoboken, NJ, USA: Wiley, 2000.

[41] X. Fan et al., "HeadFi: Bringing intelligence to all headphones," in *Proc. 27th Annu. Int. Conf. Mobile Comput. Netw.*, 2021, pp. 147–159.

[42] L. Rabiner and B.-H. Juang, *Fundamentals of Speech Recognition*, Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[43] F. Scholkmann, J. Boss, and M. Wolf, "An efficient algorithm for automatic peak detection in noisy periodic and quasi-periodic signals," *Algorithms*, vol. 5, no. 4, pp. 588–603, 2012.

[44] L. Erup, F. M. Gardner, and R. A. Harris, "Interpolation in digital modems. II. Implementation and performance," *IEEE Trans. Commun.*, vol. 41, no. 6, pp. 998–1008, Jun. 1993.

[45] K. Cho et al., "Learning phrase representations using RNN encoder–decoder for statistical machine translation," 2014, *arXiv:1406.1078*.

[46] K. Sohn, W. Shang, and H. Lee, "Improved multimodal deep learning with variation of information," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2141–2149.

[47] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena*, vol. 404, 2020, Art. no. 132306.

[48] K. O'Shea and R. Nash, "An introduction to convolutional neural networks," 2015, *arXiv:1511.08458*.

[49] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn.*, 2011, pp. 689–696.

[50] A. Howard et al., "Searching for MobileNetV3," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 1314–1324.

[51] L. Van der Maaten and G. Hinton, "Visualizing data using T-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[52] D. M. Tax and R. P. Duin, "Support vector domain description," *Pattern Recognit. Lett.*, vol. 20, no. 11–13, pp. 1191–1199, 1999.

[53] T. Ko, V. Peddinti, D. Povey, and S. Khudanpur, "Audio augmentation for speech recognition," in *Proc. 16th Annu. Conf. Int. Speech Commun. Assoc.*, 2015, pp. 3586–3589.

[54] D. S. Park et al., "SpecAugment: A simple data augmentation method for automatic speech recognition," 2019, *arXiv: 1904.08779*.

[55] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2935–2947, Dec. 2018.

[56] Apple, 2020. [Online]. Available: https://www.apple.com/airpods-pro/

[57] Amazon, 2020. [Online]. Available: https://www.amazon.com/Echo-Buds/dp/B07F6VM1S3/
[58] A. G. Howard et al., "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv: 1704.04861*.
[59] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.
[60] H. Qassim, A. Verma, and D. Feinzimer, "Compressed residual-VGG16 CNN model for big data places image recognition," in *Proc. IEEE 8th Annu. Comput. Commun. Workshop Conf.*, 2018, pp. 169–175.
[61] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
[63] Powertutor, 2009. [Online]. Available: http://ziyang.eecs.umich.edu/projects/powertutor/
[64] D. Li, S. Cao, S. I. Lee, and J. Xiong, "Experience: Practical problems for acoustic sensing," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, 2022, pp. 381–390.
[65] B. Gillham, *Developing a Questionnaire*, London, U.K.: A&C Black, 2008.
[66] G. Hinton et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.
[67] S. Alaparthi and M. Mishra, "Bidirectional encoder representations from transformers (BERT): A sentiment analysis odyssey," 2020, *arXiv: 2007.01127*.
[68] D. M. Korngiebel and S. D. Mooney, "Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery," *NPJ Digit. Med.*, vol. 4, no. 1, 2021, Art. no. 93.
[69] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 4960–4964.

Xue Sun is currently working toward the PhD degree majoring in software engineering with the School of Information Science and Technology, Northwest University. Her current research interests include ubiquitous computing and wireless sensing.

Jie Xiong received the BEng degree from Nanyang Technological University, the MS degree from Duke University, and the PhD degree from the University College London. He is currently a principal researcher with Microsoft Research Asia and an Associate Professor with the College of Information and Computer Sciences, University of Massachusetts Amherst. His current research interests include wireless sensing, mobile health, and smart IoT.

Chao Feng received the PhD degree in computer software and theory from Northwest University, Xi'an, China, in 2022. He is a lecturer with the School of Information Science and Technology, Northwest University. His current research interests include ubiquitous computing and wireless.

Haoyu Li is currently working toward the postgraduate degree majoring in software engineering with the School of Information Science and Technology, Northwest University. His current research interests include ubiquitous computing and wireless sensing.

Yuli Wu is currently working toward the undergraduate degree majoring in computer science with the School of Information Science and Technology, Northwest University. Her current research interests include ubiquitous computing and wireless sensing.

Dingyi Fang received the PhD degree in computer science from Northwestern Poly-Technical University, in 2001. He is a professor with the School of Information Science and Technology, Northwest University. His current research interests include Internet of Things, mobile and wireless computing, and information security.

Xiaojiang Chen received the PhD degree in computer software and theory from Northwest University, Xi'an, China, in 2010. He is a professor with the School of Information Science and Technology, Northwest University. His current research interests include RF-based sensing and performance issues in Internet of Things.