EarSE: Bringing Robust Speech Enhancement to COTS Headphones

DI DUAN, City University of Hong Kong, China
YONGLIANG CHEN, City University of Hong Kong, China
WEITAO XU, City University of Hong Kong, China
TIANXING LI, Michigan State University, USA

Speech enhancement is regarded as the key to the quality of digital communication and is gaining increasing attention in the research field of audio processing. In this paper, we present *EarSE*, the first robust, hands-free, multi-modal speech enhancement solution using commercial off-the-shelf headphones. The key idea of *EarSE* is a novel hardware setting—leveraging the form factor of headphones equipped with a boom microphone to establish a stable acoustic sensing field across the user's face. Furthermore, we designed a sensing methodology based on Frequency-Modulated Continuous-Wave, which is an ultrasonic modality sensitive to capture subtle facial articulatory gestures of users when speaking. Moreover, we design a fully attention-based deep neural network to self-adaptively solve the user diversity problem by introducing the Vision Transformer network. We enhance the collaboration between the speech and ultrasonic modalities using a multi-head attention mechanism and a Factorized Bilinear Pooling gate. Extensive experiments demonstrate that *EarSE* achieves remarkable performance as increasing SiSDR by 14.61 dB and reducing the word error rate of user speech recognition by 22.45–66.41% in real-world application. *EarSE* not only outperforms seven baselines by 38.0% in SiSNR, 12.4% in STOI, and 20.5% in PESQ on average but also maintains practicality.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing systems and tools**.

Additional Key Words and Phrases: speech enhancement, COTS device, acoustic sensing, multi-modality fusion, deep learning

ACM Reference Format:
Di Duan, Yongliang Chen, Weitao Xu, and Tianxing Li. 2023. EarSE: Bringing Robust Speech Enhancement to COTS Headphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 158 (December 2023), 33 pages. https://doi.org/10.1145/3631447

1 INTRODUCTION

Recent years have witnessed a surge in the use of digital communications in human society. Unlike the born auditory system of human that can separate out the target audio source of interest, it is challenging for machines to extract the clean source from the mixture of interfering components (*i.e.*, competing speech and background noise). Therefore, speech enhancement, an audio processing technique, has been developed to improve the quality of speech signals. Due to its promising performance, it has been widely adopted in a range of applications, such as telecommunication [85], speech recognition [13], and hearing aids [45].

Existing speech enhancement systems can be categorized into software-based or hardware-based solutions (Figure 1). Software-based solutions involve single-channel methods and are only related to speech modality.

Authors' addresses: Di Duan, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong SAR, Hong Kong, China, dduan5-c@my.cityu.edu.hk; Yongliang Chen, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong SAR, Hong Kong, China, cs.ylchen@my.cityu.edu.hk; Weitao Xu, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong SAR, Hong Kong, China, weitaoxu@cityu.edu.hk; Tianxing Li, Michigan State University, 428 S Shaw Ln, East Lansing, Michigan, USA, 48824, litianx2@msu.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2023 Copyright held by the owner/author(s). Publication rights licensed to ACM.
2474-9567/2023/12-ART158 $15.00
https://doi.org/10.1145/3631447

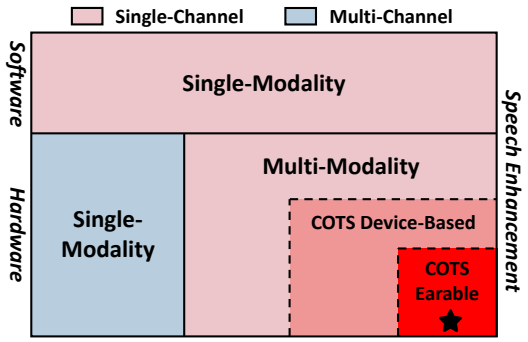


Fig. 1. A high-level overview of speech enhancement. *EarSE* is represented by the black star.

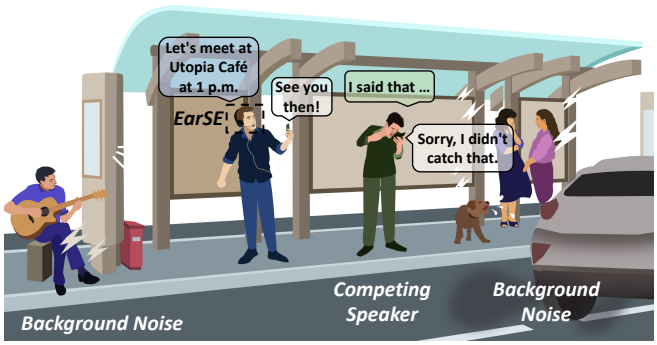


Fig. 2. Illustration of a usage scenario of *EarSE* at a noisy bus stop.

These methods manipulate the noisy speech directly with signal processing [2, 41, 51, 63], using algorithms to filter out noise, enhance certain frequencies, or model and subtract environmental noise for cleaner target speech. However, they rely heavily on prior knowledge of the encountered noise, which is usually unavailable in actual scenarios. Advanced deep learning methods [54, 64, 83, 102] have been proposed to exempt the reliance on prior knowledge. Furthermore, some online industry solutions, such as Krisp [1], were also developed for speech enhancement based on deep learning. However, these solutions cannot work offline and may aggravate the privacy leakage problem. Moreover, all these software-based solutions suffer from the label permutation problem (*i.e.*, mistakenly select competing speech as output). To solve this problem, researchers focused on hardware support, utilizing multiple microphones (*i.e.*, multi-channel) or sensors from other modalities (*i.e.*, multi-modal) to assist in selection [11, 14, 31, 62, 76, 85, 105]. As shown in Table 1, Ozturk *et al.* [62] proposed to use mmWave radar to detect the vibration of the user's vocal fold to enhance the user's speech. However, this system should be steadily placed on a table, which is not portable and practical. In addition, it suffers from the synchronization problem between auxiliary modality and speech modality, which is crucial in multi-modal speech enhancement. For instance, the sychronization between visual cues (*e.g.*, lip movements, natural facial motion) and speech is an important challenge [100] in recent cutting-edge technique—talking face generation. Although some specially designed neural networks [42] can solve the synchronization problem to some extent, they cannot achieve perfect alignment between multiple modalities, unless these modalities are collected by the same sensor, sharing the same clock. If the different modalities (*e.g.*, speech, and vibration) corresponding to the same articulatory gesture cannot be synchronized in terms of timestamps, it hinders the neural network's ability to learn the natural relationship between the features of the auxiliary modality and user speech, resulting performance degradation in speech enhancement. Recent work [14, 85, 105] shifted the focus to commercial off-the-shelf (COTS) devices, such as smartphones. However, these methods must be used in a hand-held mode and thus are vulnerable to hand tremors and common human activities (*e.g.*, walking). Additionally, they are impractical in some hands-free scenarios (*e.g.*, driving and typing). Another line of work has explored solutions using dedicated earbuds to achieve stable speech enhancement [11, 76]. However, the solutions either cannot leverage auxiliary modalities or employ the sensors that are rarely found in COTS devices (*e.g.*, bone conduction microphones [76]) to improve the performance. Furthermore, the synchronization challenge still persists in the collaboration between heterogeneous modalities, and using a dedicated device significantly introduces deployment overhead and limits its wide use. To address the synchronization problem, recent work [11] leverage a pair of dedicated earbuds to measure the time difference of arrival (TDoA) and distinguish the direction of the sound source and suppress the noise from the side. Furthermore, there are also some industry products (*e.g.*, Airpods Pro, Galaxy Buds2 Pro) that use beamforming techniques

Table 1. Comparison of hardware-based speech enhancement methods (❍–Low, ◗–Medium, ●–High).

Solution	Hardware Platform	Multi-Modal Auxiliary	COTS Device	Stability	Hands-Free	Attention Mechanism
Ozturk *et al.* [62]	Mic, mmW Radar	●	◗	❍	❍	❍
UltraSE [85]	Smartphone	●	●	❍	❍	◗
UltraSpeech [14]	Smartphone	●	●	❍	❍	❍
Sensing to Hear [105]	Smartphone	●	●	❍	❍	❍
ClearBuds [11]	Earbuds	❍	❍	●	●	❍
Schilk *et al.* [76]	Earbuds	●	❍	●	●	❍
EarSE	Headphone	●	●	●	●	●

and measure TDoA to enhance the user's speech. However, these solutions cannot address the noise directly ahead, as noise from this direction does not produce TDoA at the binaural microphones. Despite the inherent advantages of earables devices (*e.g.*, portability and hands-free use), there is no robust, hands-free, multi-modal speech enhancement solution based on them, leaving a research gap in the field of speech enhancement.

To fill this unexplored research gap, we propose *EarSE*—the first earable speech enhancement solution based on slightly modified COTS headphones. *EarSE* is a robust, hands-free, multi-modal speech enhancement system that provides remarkable speech enhancement performance. Figure 2 illustrates an application scenario in which a user, wearing *EarSE*, calls his friend at a noisy bus stop. By leveraging ultrasonic signals as an auxiliary modality, *EarSE* can focus on the user's facial articulatory gestures and extract clean speech from the complex, noisy environment. The key idea of *EarSE* is leveraging the ultrasonic waves leaked from an ear pad of headphones and the opposite-side boom/modular microphone to establish a stable acoustic sensing field across the user's face. It can actively capture the subtle facial articulatory gestures when speaking (*i.e.*, lip motions, tongue protrusions, jaw movements, and skin deformation), which can serve as the auxiliary modality for speech enhancement. However, it is nontrivial to instantiate our idea in practice with three main challenges:

- *Extremely low Signal-to-Noise Ratio (SNR).* Due to the limitation of COTS headphones (*i.e.*, the side sensing view, obstruction of the user's cheek, soundproof materials of the ear pads, and volume control for health considerations), SNR of ultrasonic signals is extremely low, degrading *EarSE*'s sensing granularity.
- *The lack of a self-adaptive method for extracting informative features that is sensitive to subtle deformation.* Extracting effective features with consistency and reproducibility to detect the user's subtle articulatory gestures (typically involve < 5 cm moving distance) is a challenging task. Furthermore, user diversity (the size and shape of the head, facial adiposity, and hair) directly influences the position and characteristics of informative areas in the Channel Impulse Response (CIR) profiles generated by the acoustic sensing field.
- *The shortage of auxiliary modalities for accurate speech separation and effective multi-modal fusion methods.* Owing to the complex structure of the human brain, which relies on visual cues (*i.e.*, lip-reading) and effective modality fusion mechanisms, it continually gives humans an edge over machines in distinguishing target speech in noisy environments. The shortage and ineffectiveness of multi-modal auxiliary in machine hearing results in unsatisfactory speech enhancement performance in complex auditory environments.

To address the three challenges, we propose three countermeasures in *EarSE*: (1) We place two skin-friendly auxiliary spacers on a headphone's ear pad, creating a gap for ultrasonic waves to "escape"; the asymmetrical distribution of the spacers intentionally makes the gap towards the boom mic, which increases the SNR of ultrasonic signals significantly. (2) We modulate an inaudible Frequency-Modulated Continuous-Wave (FMCW) sensitive to subtle skin deformation. By proposing a method to identify the arrival point of the direct echo, the generated CIR profiles are consistent and reproducible. Furthermore, we introduce the Vision Transformer

(ViT) [15], treating the adaptive extraction of informative areas in CIR profiles among different users as a vision task. (3) To effectively enhance speech modality by ultrasonic modality, we design a Multi-modal Fusion module that combines the Multi-Head Attention (MHA) and Factorized Bilinear Pooling (FBP) gate to fuse the features of two modalities and selectively use auxiliary information.

We implement *EarSE* on three COTS headphones and conduct comprehensive evaluation against seven baselines. In general, *EarSE* achieves remarkable 19.48 dB SiSNR and 3.32 PESQ in speech enhancement and outperforms the state-of-the-art solutions by 2.34 dB–7.83 dB and 9.9–48.9%, respectively. *EarSE* is robust (>16 dB in SiSNR) to user diversity (18 users from 11 countries) and effective (>13 dB in SiSNR) in extremely noisy environments (three competing speakers). Furthermore, *EarSE* can significantly reduce the word error rate (WER) by 22.45%–66.41% in real-world applications.

The contributions of this paper can be summarized as follows:

- We propose *EarSE*, the first robust, hands-free, multi-modal speech enhancement solution based on slightly modified COTS headphones without any modification on internal circuitry. *EarSE* fills the research gap in COTS earable-based speech enhancement.
- We address three key challenges which hinder the implementation of the prototype of *EarSE* by proposing a method to enhance the SNR of leaked ultrasonic waves, applying FMCW sensing techniques that are sensitive to skin deformation from the side sensing view, and designing a fully attention-based DNN to obtain the selective attention ability, proposing a novel multi-modal fusion method that employs multi-head attention and gating method to enhance speech modality by ultrasonic modality. Furthermore, we are the first to deploy the proposed multi-modal fusion method and a fully attention-based system on mobile devices, and we design a companion application for the system on mobile platforms.
- We conduct comprehensive evaluations on a sizeable self-collected dataset with diverse participants and devices (21 participants from 11 countries, 19 hours using three devices). The experimental result shows that *EarSE* outperforms the seven baselines in all five evaluation metrics and is robust to extremely complex environments while maintaining practicality.

The rest of the paper is organized as follows. We first present a comprehensive related work in Section 2. Then, we introduce the background and motivation of *EarSE* in Section 3. Next, we provide an overview of *EarSE* in Section 4. In Section 5, Section 6, and Section 7, we describe the hardware design, signal processing, and DNN model design. Then, we evaluate the performance of *EarSE* in Section 8. Finally, we discuss the limitations of *EarSE* and indicate the future work in Section 9 before concluding the paper in Section 10.

2 RELATED WORK

2.1 Speech Enhancement

Despite the extensive study of speech enhancement, persistent challenges such as the label permutation problem in speech separation and the cocktail party problem [12] in auditory perception remain unresolved. Traditional methods for speech enhancement, including MMSE [16], Kalman filtering [63], and spectral subtraction [41], rely on the prior knowledge of noise, thus limiting their performance and practicality. Recently, deep learning has exhibited significant potential across numerous fields, inclusive of speech enhancement. Numerous deep learning-based methods have been proposed for speech enhancement [14, 53, 54, 85, 103]. These methods can be classified according to their utilized modalities, specifically, single-modal and multi-modal methods.

2.1.1 Single-Modal Speech Enhancement. The methods which merely leverage speech modality for speech separation and enhancement, can be further subdivided based on the domains they operate in. More specifically, these methods can be categorized as either time-frequency (T-F) domain methods or time (T) domain methods.

Classical Time-Frequency (T-F) domain methods [32, 58, 97] aim to recover clean target speech by applying a learned spectrogram mask to the noisy speech spectrogram, supplemented by the original noisy phase. However, due to imprecise phase information, their performance is limited. Advanced deep learning techniques [18, 59, 102] have emerged to address this, but challenges persist, including handling long-duration noise and the absence of an effective attention mechanism for source separation. Furthermore, T-F domain methods increase computational complexity and are sensitive to parameter settings, potentially introducing distortions during the spectrogram to audio transformation, such as iSTFT. To mitigate these issues, some researchers have shifted their focus to the time (T) domain, processing the waveform directly. Techniques such as generative methods [65, 74] and encoder-decoder architectures like TCNN [64] and Conv-TasNet [54] have been proposed. Nevertheless, they have limitations in context understanding and inference speeds. Recently, attention mechanism, which is deeply explored in natural language processing, has also been applied on speech enhancement. For instance, Subakan *et al.* proposed SepFormer [83], an innovative transformer-based network for speech separation. Despite their potential, all of these methods encounter a crucial but challenging problem—the label permutation problem (*i.e.*, output competing speech as the target speech), which is a lion in the way to achieve outstanding performance.

2.1.2 Multi-Modal Speech Enhancement. To accurately select target speech amidst multiple sources, auxiliary modalities have been explored for speech enhancement, including bone conduction modality [31, 76], visual modality [17], wireless communication modality [62], and audio modality [14, 85, 105].

Ephrat *et al.* [17] and Rahimi *et al.* [71] employ visual cues from video as an auxiliary speech enhancement modality. Ozturk *et al.* [62] utilize mmWave measurements of vocal fold vibrations to improve speech separation and enhancement. However, privacy concerns arise with front-facing cameras, and mmWave radars lack portability. Consequently, research has shifted to audio-only, multi-modal speech enhancement solutions [14, 85, 105] utilizing low-frequency ultrasonic modalities. These smartphone-based solutions are susceptible to human activities and hand tremors. More recently, Schilk *et al.* [76] developed customized earbuds using bone-conduction audio as an auxiliary modality, potentially mitigating these issues. Nevertheless, MEMS bone-conduction microphones are rarely found in commercial off-the-shelf (COTS) earbuds. Despite various attempts, a multi-modal speech enhancement solution based on COTS head-mounted devices remains unexplored, presenting a significant research gap. To fill this gap, we propose *EarSE*, the first stable, hands-free, multi-modal speech enhancement solution capable of effectively resolving the label permutation problem via multi-modality fusion. This is due to its unique acoustic sensing field, merely focusing the intended user's face.

2.2 Earable Sensing

Earable devices have emerged as a new sensing platform for intelligent applications such as face reconstruction [47, 98] and health monitoring [7, 8].

2.2.1 Dedicated or Remoulded Device. Several researchers have fabricated custom prototypes for various applications such as blood pressure measurement [7], face touch detection [40], microsleep events detection [67], facial reconstruction [47, 98], and unvoiced commands recognition [80]. Notably, Chatterjee *et al.* developed ClearBuds [11], a pair of earbuds with synchronized timestamps forming a dual-microphone array for speech enhancement. However, they fail to distinguish user speech from frontal interference due to lack of the time difference of arrival (TDoA). Turning such prototypes into cost-effective, practical devices remains challenging, limiting their widespread use. Alternatively, some studies [26, 37, 38, 49, 69, 94, 99] modify existing earbuds by adding sensors. The in-ear microphone is a popular addition, enabling applications like ear disease detection [38], teeth gesture-based interaction [69], user authentication [25, 26, 94, 95, 99], gait-based user identification [22], activity recognition [55], silent command recognition [39], and on-face interaction [101]. Other sensor additions

have also been explored, such as magnetic coils [49] and IMUs [37]. However, these devices rely on dedicated systems or sensors rarely found in COTS devices.

2.2.2 COTS Device. Given the constraints of remodeled and dedicated earable devices, researchers have turned to COTS headphones and earphones. Cao *et al.* [10] presented EarphoneTrack, an innovative acoustic motion tracking approach using earphones. It employs earphone speakers and screen or phone microphones to create a new human-computer interaction mode. Subsequently, Fan *et al.* [21] proposed HeadFi, which leverages a Wheatstone Bridge to measure the imbalance between two earphones, enabling applications like user identification, heart rate monitoring, and gesture recognition. Recently, Wang *et al.* [93] proposed FaceOri, which uses the three microphones in typical active noise cancellation headphones, along with a smartphone, for head pose estimation. To our best knowledge, no COTS headphone-based system has been proposed for speech enhancement. We present *EarSE*, the first multi-modal speech enhancement solution based on COTS headphones. It utilizes the form factor of the headphones equipped with a boom microphone, and the ultrasonic wave "escaped" from the gap created by auxiliary spacers, establishing a stable sensing field for detecting subtle articulatory gestures.

3 BACKGROUND & MOTIVATION

3.1 Human Speech Articulatory Gestures

Articulatory gestures refer to the vocal tract movements, such as the lips, tongue, jaw, velum, and larynx movements [29]. These gestures create specific acoustic characteristics during speech production. Overall, sensing articulatory gestures requires particular attention to the lips, jaw, and tongue when using acoustic sensing methods. These facial regions play a crucial role in speech production and provide valuable information for speech enhancement applications [24, 44, 81].

Lips: The lips play a significant role in shaping the airflow during the production of bilabial sounds (*e.g.*, /p/, /b/, and /m/) and labiodental sounds (*e.g.*, /f/ and /v/). In addition, lip rounding and spreading can alter the resonance properties of the vocal tract, affecting vowel quality.

Jaw: The jaw's position influences the tongue's position and shape, affecting the vocal tract's configuration. Jaw movements are crucial for the production of consonants (*e.g.*, /k/ and /g/) and vowels with different tongue heights (*e.g.*, /i/, /u/, and /a/).

Tongue: The tongue is the most versatile articulator, capable of altering its shape, position, and stiffness. Different tongue positions and shapes contribute to the production of various speech sounds, such as alveolar (*e.g.*, /t/, /d/, /s/, and /z/), palatal (*e.g.*, /ʃ/ and /ʒ/), and velar (*e.g.*, /k/ and /g/) consonants, as well as different vowel qualities.

Velum: The velum controls the airflow between the oral and nasal cavities. By raising or lowering the velum, speakers can produce nasal (*e.g.*, /m/, /n/, and /ŋ/) or oral sounds.

Larynx: The larynx houses the vocal folds, responsible for producing voiced sounds when they vibrate. Adjusting the tension and position of the vocal folds can modify the pitch and intensity of the voice and produce speech harmonics.

3.2 Motivation

In real-world usage scenarios, pervasive inevitable noises, such as background noise (*e.g.*, construction sounds, traffic noise, pet sounds, music in a bar, and instrument sounds) and the interfering speech from competing speakers near the user, deteriorate the quality of the recorded user speech. These noises are recorded alongside the user's speech by a microphone, resulting in low SNR, low intelligibility, and poor quality. Existing speech enhancement solution solutions, however, either require hand-held devices (*e.g.*, smartphone) [14, 85] or dedicated

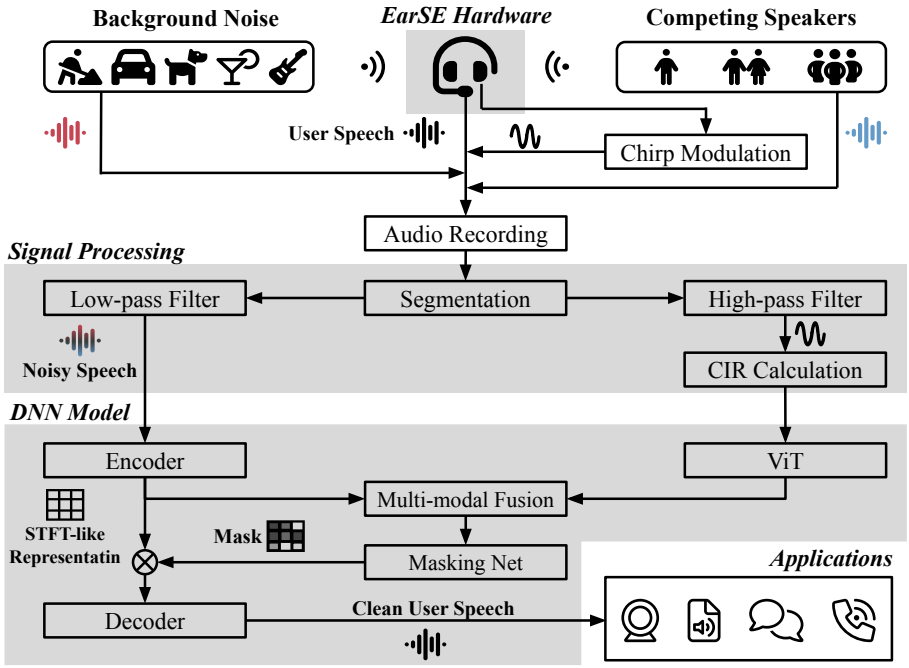


Fig. 3. *EarSE* system overview. It illustrates the pipeline from polluted speech to clean user speech, achieved through the integration of three main components: hardware, signal processing, and DNN model.

earbuds [11]. The former suffers from performance degradation due to hand tremors[1], fatigue, or activities like walking and it is not suitable for scenarios like driving. The latter introduces significant deployment overhead and is not compatible to various earbud models. Therefore, developing *EarSE* lies in the need for a robust, ubiquitous, and hands-free speech enhancement system that overcomes the limitations of current SSE solutions and can be applied in various scenarios, including driving, walking, and exercising.

4 SYSTEM OVERVIEW

We propose *EarSE*, the first hands-free single-channel multi-modal speech enhancement solution using COTS headphones. The key idea is utilizing the modulated chirp signal "escaped" from the gap created by skin-friendly auxiliary spacers to extract the informative features of articulatory gestures and separate the user speech from ambient noise. Figure 3 illustrates the overview of *EarSE*, including three main components. First, *EarSE* leverages the speaker opposite the boom microphone to emit FMCW signals(*i.e.*, chirp signals). With the auxiliary spacers, the "escaped" chirp signal from the speaker can be recorded by the boom microphone to monitor articulatory gestures. Second, the signal processing module segments the recorded audio into clips, filtering each clip with a high-pass filter and a low-pass filter to separate the noisy speech within the chirp signal. The user's articulatory gestures perturb the chirp-formed sensing field, reflecting at the direct propagation path or nearby positions. Thus, we calculate the Channel Impulse Response (CIR) profiles, serving as auxiliary information to aid speech enhancement. Third, the noisy speech and CIR profiles are fed into *EarSE*'s transformer-based DNN model.

[1]Mild tremors occur in the general population, particularly in situations involving stress, anxiety, fatigue, or excessive caffeine intake [30], introducing relative displacement between a hand-held smartphone and the user's mouth

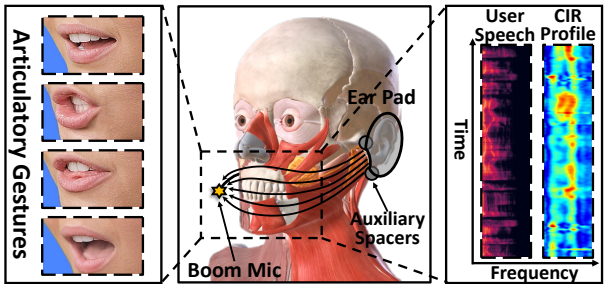


Fig. 4. The sensing principle of *EarSE*. The middle image shows the anatomy [6] of muscle groups involved in facial articulatory gestures [20], highlighted in red. A stable acoustic sensing field is established across the user's face, which captures the subtle articulatory gestures and alignment them with the user's speech using a single-channel method.

Simultaneously, the ViT module of *EarSE* extracts ultrasonic features based on attention. The features from the two modalities are then fused by a multi-modal fusion module. Then the fused features are used to generate a mask for the encoded STFT-like representation by the Masking Net module of *EarSE*. Finally, the masked representation is decoded by a decoder module, canceling noises and producing clean user speech.

5 SENSING RATIONALE & HARDWARE DESIGN

In this section, we first describe the sensing rationale that enables hands-free speech enhancement, followed by the design of *EarSE* hardware.

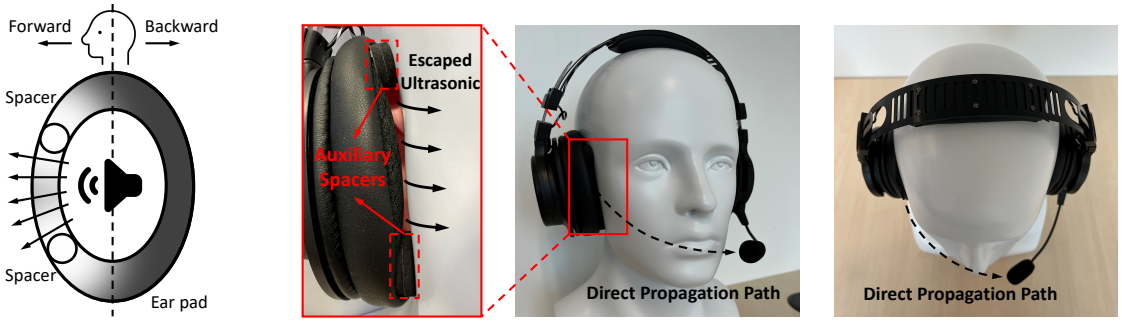
5.1 *EarSE* Sensing Rationale

As it depicted in Figure 4, the key idea of *EarSE* is leveraging the leaked ultrasonic waves from the ear pad, situated on the opposite side of the boom mic, to create a stable sensing field across the user's face. Since the attenuation rate in air is far lower than in human tissues, the direct propagation path with the least resistance runs along the user's face. Therefore, when the user performs various articulatory gestures (*i.e.*, lip motions, tongue protrusions, jaw movements, and skin deformation), these human tissues perturb the acoustic sensing field, resulting in variations in the direct propagation path or the nearby paths. Owing to the precise alignment of ultrasonic waves and the user's speech in timestamps provided by the single-channel approach, the features calculated from the ultrasonic can accurately reflect the user's articulatory gestures in real time.

However, it is not trivial to design such a hands-free speech enhancement system based on the sensing principle because the SNR of the ultrasonic leakage is extremely low in hands-free scenarios. First, the user's cheek obstructs most of the ultrasonic waves, and ultrasonic attenuates rapidly within human tissue [106]. Second, high-end headphones often employ superior soundproof materials for ear pads, resulting in near-perfect isolation. And the power of ultrasonic waves must be low for health considerations [61]. Finally, instead of sensing from the front view, *EarSE* can only utilize a side sensing view that captures the gestures which perturb the stable acoustic sensing field across the user's face. Thus, the reflection-based Doppler shift method is not applicable, and the CIR with GSM sequence-based method fails to deliver satisfactory performance in detecting skin deformation [47].

5.2 *EarSE* Hardware Design

To increase the SNR of the ultrasonic leakage, we propose using two auxiliary spacers to enhance the leaked ultrasonic waves without any modification to the COTS headphones. Also, we leverage Frequency-Modulated Continuous-Wave (FMCW) as our acoustic sensing signal, which has been demonstrated to be sensitive to skin deformation [47]. The combination of the two components helps separate the user audio from the external noise, benefiting lateral human speech enhancement.



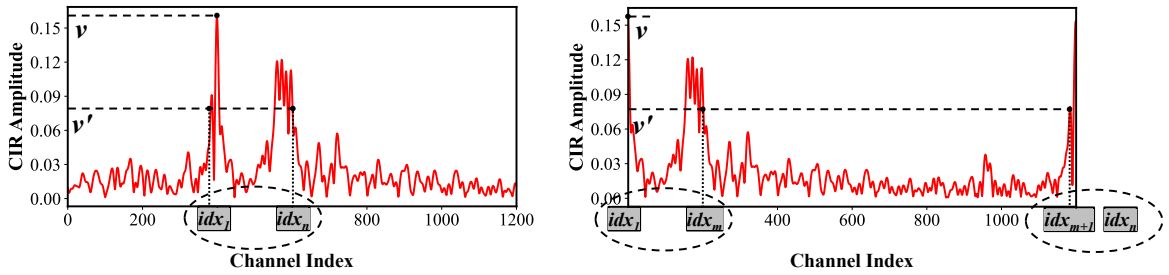
(a) Auxiliary spacers (b) *EarSE* hardware implementation

Fig. 5. *EarSE* utilizes auxiliary spacers to directionally create a gap for ultrasonic wave to "escape". These signals pass by the user's face, creating a stable acoustic sensing field. The user's articulatory gestures perturb the direct propagation path or nearby paths, resulting in variations of CIR profiles.

5.2.1 Auxiliary Spacers Design. As shown in Figure 5(a), we leverage skin-friendly material spacers placed on the same side of the ear pad to create a gap for the ultrasonic waves to "escape". Due to the asymmetrical distribution of the spacers, only one side creates a gap for the ultrasonic waves to escape, while the other side remains in close contact with the user's skin behind the ear. Then we intentionally direct the gap towards the boom/modular microphone's direction (Figure 5(b)), allowing the escaped ultrasonic waves to form a stable acoustic sensing field across the user's face.

5.2.2 FMCW Modulation. The FMCW has an ideal autocorrelation property to separate the signal propagation paths by estimating their CIR [47, 92]. Specifically, we use the speaker on the headphone and the opposite-side boom/modular microphone to transmit and receive the 15 kHz–20 kHz FMCW signal to capture the subtle displacement of the human face when speaking. This frequency band was selected based on its inherent advantages: it is above the auditory perception range of most adults (the upper limit in average adults is often closer to 15 kHz [70]), ensuring that the generated ultrasounds are inaudible to the majority of users [47] and providing sufficient bandwidth with a lower level of autocorrelation side lobes [86]. For some auditory-sensitive users, such as teenagers, we can appropriately reduce the bandwidth and use 16 kHz as the starting frequency of the chirp. Additionally, this frequency range is within the capabilities of most COTS headphones (only few headphones can emit ultrasounds above 20 kHz), making it a practical choice for implementation in existing headphone designs. Since different articulatory gestures activate different combinations of facial muscles, analyzing signals perturbed by different regions of the face (*i.e.*, signals traveling different distances) provides informative features. In *EarSE*, we set the period of the FMCW signal to be 1200 samples and the sampling rate as 48 kHz. Then, we can update the facial geometrical features 40 times (48000/1200) per second. With the 0.708 cm (34000/48000) minimal distinguishable length difference of adjacent ultrasonic propagation paths, the sensing resolution is sufficient for us to capture facial articulatory gestures effectively.

We also investigate the health implications of *EarSE*. In this experiment, we measure the sound pressure level (SPL) of the ultrasounds emitted and received by *EarSE* at the external auditory meatus of the human head model and boom microphone. The SPL of the emitted and received ultrasonic are about 65 dB and 33 dB, respectively. The World Health Organization (WHO) recommends that noise exposure levels should not exceed 70 dB over a 24-hour period, and 85 dB over a 1-hour period to avoid hearing impairment [23]. *EarSE* has a 5 dB (*i.e.*, 3.16 times) margin from the level that can affect human health, which is safe enough for the user's health.



(a) Filtered indexes are within one cluster

(b) Filtered indexes fall in two clusters

Fig. 6. Locating the direct propagation channel. Only the first and last indexes inside a cluster are shown. In (a), there is only one cluster, so the first index (idx_1) is selected. In (b), there are two clusters, and the first index of the second cluster (idx_{m+1}) is selected.

6 SIGNAL PROCESSING

In this section, we elaborate on the design details of the signal processing steps for speech and ultrasonic modalities, respectively. The resulting processed audio and CIR profiles serve as the inputs of *EarSE*'s DNN model.

6.1 Speech Modality Signal Processing

We first apply a low-pass Butterworth filter with a cut-off frequency of 8 kHz to obtain the noisy speech. The study [57] demonstrates that signals above 8 kHz have minimal impact on speech intelligibility and human perception. According to the Nyquist–Shannon sampling theorem [60, 78], a sampling rate of 16 kHz for speech signals is deemed sufficient. Therefore, to reduce the computational burden on the DNN model, we resample the filtered speech signal to 16 kHz, resulting in 1×16000 scalars per second. These scalars will serve as the speech modality input for the DNN model of *EarSE*.

6.2 Ultrasonic Modality Signal Processing

After obtaining the vocal component from the received audio, we derive auxiliary information by extracting articulation-involved features from the recorded ultrasonic signals. Specifically, we apply a high-pass Butterworth filter with a cut-off frequency of 15 kHz on the audio clip to filter out those irrelevant audible components, including human voices and ambient noises, and only keep the chirp signals occupying the ultra band.

6.2.1 CIR Profile Calculation. In *EarSE*, we employ the cross-correlation analysis to determine the distance variations of multi-path signal propagation, reflecting the articulatory gesture-induced patterns. Given the transmitted chirp clip and the received high-pass filtered signals, we calculate their cross-correlation values, and the obtained complex sequence can approximate the CIR of the echo propagation channels [9]. However, estimating the CIR value of one acoustic channel using cross-correlation requires 1,200 multiplications and additions, and most of the channels are irrelevant to our purpose (*e.g.*, signals rebounding from surroundings or reflecting from other body parts). Therefore, to reduce computational overhead in estimating CIR values such that the acoustic frames of interest can be obtained in real-time, it is necessary to accurately identify the arrival point of the direct echo and capture the portions of signals perturbed by articulatory gestures only. Specifically, we calculate 1,200 consecutive CIR values with the early audio samples collected in the initial state of *EarSE* and apply an adaptive threshold-based method to identify the target components. This is because choosing the length of one complete chirp sequence can cover all measurable channels in our acoustic sensing setting. Due to signal superposition in the multi-path propagation scenario, the signal traversing through the direct path is not

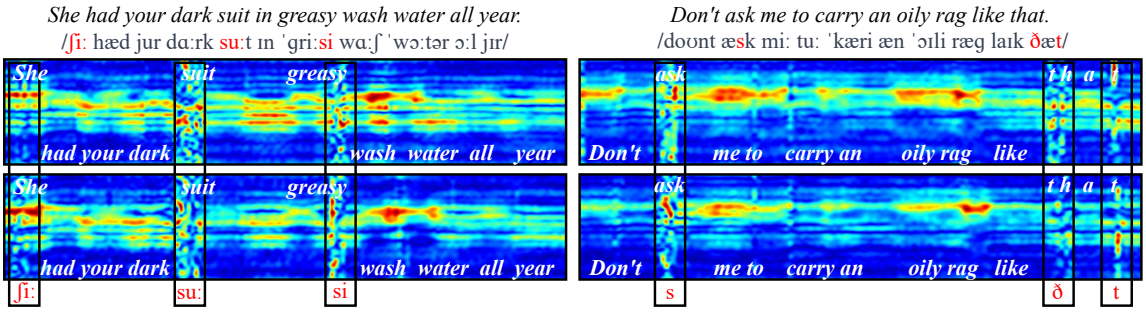


Fig. 7. The CIR profiles generated from four sentences. *EarSE* is capable of providing consistent and reproducible CIR profiles when the same user repeats a sentence, while different speech produces significantly different profiles. By analyzing the profiles and using airflows in the sentence as landmarks, the generated CIR profiles become explainable and can be aligned with the user's speech.

the strongest [9]. Therefore, we filter out all the channels with CIR amplitudes greater than half of the largest amplitude value in the estimated 1200-length CIR sequence.

Denoting the indexes of those channels satisfying the condition as $(idx_1, idx_2, ..., idx_n)$, we apply the "DBSCAN" algorithm [19] to cluster these indices into clusters, in which the distance between neighboring indexes is smaller than a predetermined value (*i.e.*, 600 in *EarSE*). Typically, there should only be one resultant cluster (Figure 6(a)), which indicates that the estimated audio clip contains only the echoes of interest from one chirp sequence. These are the ultrasounds that sweep across the lip and cheek area with sufficiently strong power. However, it is possible to obtain two separated clusters (Figure 6(b)) is possible because the audio clip may cut parts of the signals of interest into two consecutive chirps. In *EarSE*, to accurately approximate the channel index of the direct channel, we select the smallest index when there is only one cluster. In the case of two clusters, we select the smallest index in the second cluster because the direct channel in the first cluster is missing in the estimated audio clip, and only the direct channel in the second cluster is captured.

In practice, we discover that the head sizes and shapes vary among users, and the informative CIR range affected by articulatory gestures may differ. To cover the channels of interest thoroughly for all users, we intentionally choose a large channel number to segment the informative CIR range in a coarse grain and further employ a Vision Transformer (ViT) neural network (Section 7.2) to precisely capture the informative CIR ranges of individual users. Then, we set the informative CIR range to 30 channels before and 170 channels after the direct channel, resulting in a 200-length acoustic frame generated by one chirp sequence. Once the first frame is determined, the lateral frames can be directly located by adding 1,200 delays since the chirp sequences are transmitted consecutively, and the computational burden is reduced by only estimating 200 CIR values for each chirp.

6.2.2 Profile Analysis. Figure 7 displays the informative areas of CIR profiles for four sentences. The left two figures are the CIR profiles when a user speaks "She had your dark suit in greasy wash water all year", while the right two figures are the CIR profiles when the same user speaks "Don't ask me to carry an oily rag like that". We observe that the head-mounted property of *EarSE* provides a stable acoustic sensing field over time (the horizontal axis). The user's articulatory gestures perturb the paths nearby the direct propagation path, resulting in variations of each CIR profile.

Figure 7 shows the features (*i.e.*, CIR profiles) for sensing subtle articulatory gestures. We observe that when the user utters certain fricatives, such as /ʃ/, /s/, /ð/, and /t/, the generated airflow is captured by the microphone near the corner of the user's mouth. These airflows are reflected in the spectrogram and extend far beyond the frequency range of human speech (300 Hz–3.4 kHz [88]), impacting the ultrasonic frequency band of *EarSE*

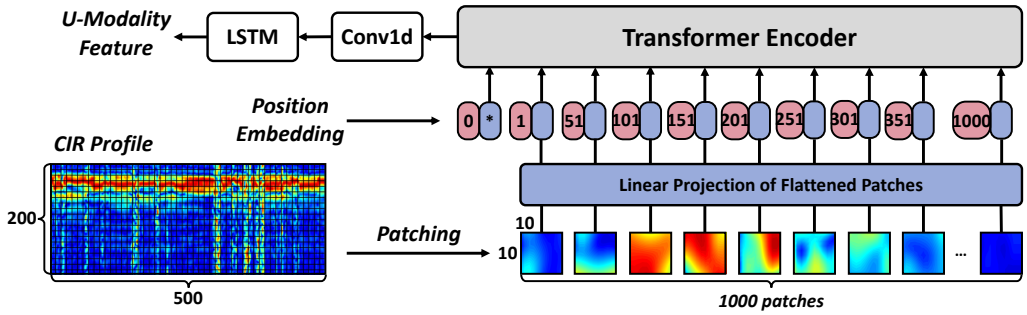


Fig. 8. Using the ViT to adaptively extract informative areas from the CIR profiles generated in Section 6.2.1.

(15 kHz–20 kHz). This phenomenon is manifested as temporary disablement in acoustic sensing, resulting in a disordered profile. Although a transient airflow brings a disordered profile, it can still help *EarSE* to separate the user's speech in complex environments by locating the fricatives of the user's speech, which owes to the precise alignment between the speech and the ultrasonic in a single channel. To describe the correlation between a user's speech and generated CIR profiles, we leverage the aforementioned airflows as landmarks to align the two items and further annotate the speech content (*i.e.*, vocabulary) on the corresponding CIR profiles.

7 DNN MODEL DESIGN

In this section, we elaborate on the DNN design of *EarSE*, which contains five modules: Encoder, Decoder, ViT, Multi-modal Fusion, and Masking Net. To accelerate the training process and minimize the training preparation overhead, we leverage transfer learning techniques and implement a dynamic mixing framework.

7.1 Encoder and Decoder

The encoder module and decoder module have a single 1D (or 1D transposed) convolutional layer. Specifically, the encoder takes a period of noisy speech $x \in \mathbb{R}^{B\times 1\times T}$ as the module input, where B is the batch size and T is the length of time series. Thereafter, the 1D convolutional layer expends x into N channels. After a Rectified Linear Unit (ReLU) activation function layer, the output of the encoder module is an STFT-like representation $h \in \mathbb{R}^{B\times N\times T'}$. This transformation is represented as follows:

$$F_s = \text{ReLU}(\text{conv1d}(x)).$$

After generating a mask m for the STFT-like representation h by Masking Net module (elaborated in Section 7.4), the mask m is point-wise multiplied (denoted as $\odot$) with h and the result is fed into the Decoder module equipped with a 1D transposed convolutional layer. The transformation is defined as follows:

$$x_e = \text{conv1d-transpose}\left(m \odot F_s\right).$$

By using the mask to selectively suppress or enhance the specific area in h, the decoder outputs a same length of clean user speech x_e as x.

7.2 Vision Transformer

The Vision Transformer (ViT) [15] is a computer vision model based on the Transformer architecture, which applies the Transformer to images, bypassing traditional convolutional layers. Its advantages include better generalization, stronger representational power, scalability, and end-to-end training capabilities, making it highly effective for various computer vision tasks. The implementation of this module follows the GitHub project [90].

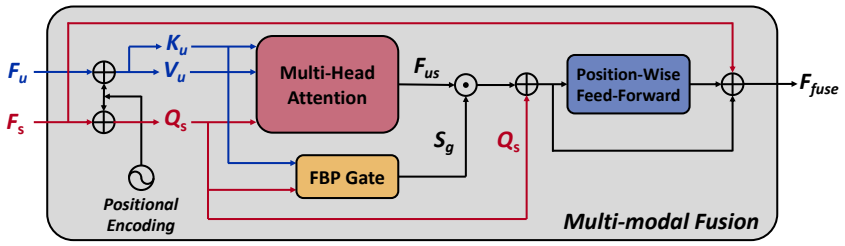


Fig. 9. Enhance speech modality by ultrasonic modality using Multi-modal Fusion module.

As Figure 8 shows, we segment the CIR profile $P \in \mathbb{R}^{H \times W}$ (H and W are 200 and 500 in this paper) obtained in Section 6.2 into 1,000 10×10 patches. The shape 10×10 reflects the patch's resolutions in space distance and time, which are 7 cm and 0.1 s, respectively. In addition, we observe that the informative area spans about four rows of patches, corresponding to a range of approximately 28 cm that matches the typical size of a human face. The patches are flattened into 1D vectors and embedded through a linear layer. Positional encodings are then added to maintain spatial information, and the embedded patches are combined into a sequence. This sequence is passed through multiple Transformer layers, where self-attention mechanisms capture long-range dependencies among the elements. Subsequently, we employ a 1D convolutional layer and Long Short-Term Memory (LSTM) layer to expand the output feature map along the channel and temporal dimensions, respectively. This ensures the final feature map F_u of the ultrasonic modality is dimensionally consistent with the feature map F_s of the speech modality.

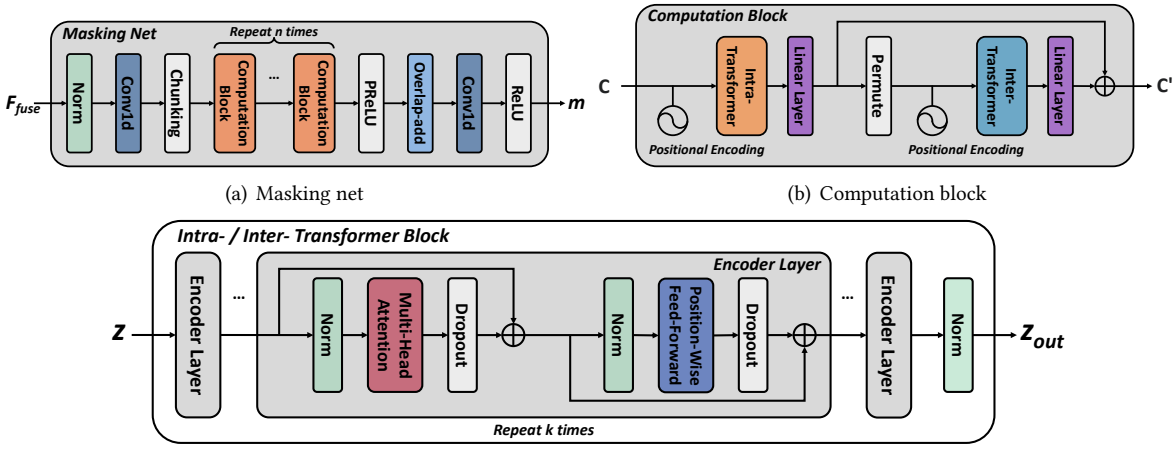
7.3 Multi-Modal Fusion

The speech modality feature F_s and ultrasonic modality feature F_u are fused by the gating-based Multi-modal Fusion module to get the fusion feature F_{fuse}. Figure 9 shows the fusion module has three main components, namely a Multi-head Attention (MHA), a Factorized Bilinear Pooling (FBP) gate, and a Position-wise Feed-Forward network (FFN). Since we use ultrasonic modality as an auxiliary modality to help the selection and suppression in speech modality, we use the Q_s vector from speech modality, and the K_u, V_u vectors from ultrasonic modality as inputs. By employing the Q_s vector from the speech modality, we can probe the ultrasonic modality for information relevant to the speech modality. Conversely, using the K_u and V_u vectors from the ultrasonic modality provides access to the ultrasonic modality's information that is pertinent to the queries from the speech modality. The transformation can be derived as:

$$\begin{aligned} F_{us} &= \text{MHA}\left(Q_s, K_u, V_u\right) \\ &= \text{softmax}\left(\frac{Q_s K_u^\top}{\sqrt{d_{ks}}}\right) V_u \\ &= \text{softmax}\left(\frac{F_s W_{Q_s} W_{K_u}^\top F_u^\top}{\sqrt{d_{ks}}}\right) F_u W_{V_u}, \end{aligned}$$

where the W_{Q_s}, W_{K_u}, and W_{V_u} denotes the weight matrices for mapping separate modality features F_s and F_u to Q_s, K_u, V_u vectors. $\frac{1}{\sqrt{d_{ks}}}$ is a scaling factor that ensures that the dot product results do not become excessively large. It prevents the vanishing gradient issue when calculating attention weights with the softmax function, particularly in cases where the dimensionality is high. The output of the MHA is the interacted features F_{us}.

Simultaneously, the Q_s and K_u are fed into the FBP gate mechanism to generate a temporal gated signal S_g for adaptively controlling the interaction between speech modality and ultrasonic modality. The FBP gate is a feature fusion technique often used in multi-modal tasks. It aims to fuse features from different modalities while



(a) Masking net

(b) Computation block

(c) Transformer block

Fig. 10. The structure of the Masking Net used in *EarSE* is depicted as follows: (a) represents the structure of the Masking Net module; (b) shows the key computation block within the Masking Net module; and (c) provides the details of a transformer encoder layer, which serves as the core of the computational block.

reducing computational complexity. It helps models learn cross-modal relationships, improving performance. The key idea of the FBP gate is to factorize input features and perform bilinear pooling on the decomposed features, capturing interactions while avoiding high complexity.

The Q_s and K_u vectors are transformed by the weight matrices $W_{Q_{s'}}$ and $W_{K_{u'}}$, mapping them to the same hidden dimension d_g. The mapped results, Q'_s and K'_u, then undergo an element-wise multiplication. A sum pooling operation is applied to the multiplied results. After an L2 normalization and a linear projection, the output S_g of the FBP gate is used to perform an element-wise multiplication with F_{us}. Here, we add a shortcut connection from Q_s to F_{us}. Then, the computation results will be fed into the FFN. The output of the FFN, combined with a shortcut vector from F_s, will be output as the two-modal fused features, F_{fuse}. The role of FFN is to introduce non-linearity and local perception to the model. The FFN consists of two linear layers with a non-linear activation function (ReLU) layer inserted between them. Finally, the transformation of the proposed Multi-modal Fusion module can be represented as follows:

$$F_{fuse} = \text{FFN}(F_{us} \odot \text{Sign}(S_g) + Q_s) + F_s,$$

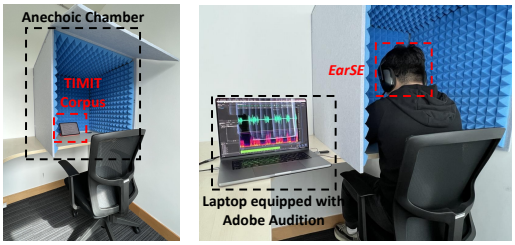
where Sign is a sign function to produce 0 or 1 for binary gating. The two-modal fused features F_{fuse} are used to generate a mask for the STFT-like representation F_s.

7.4 Masking Net

Figure 10(a) shows the structure of the Masking Net module, which takes the fused features F_{fuse} as input and outputs a mask m for the user's speech STFT-like representation F_s. Specifically, $F_{fuse} \in \mathbb{R}^{B \times H_{fuse} \times T'}$ is normalized with a normalization layer [5]. We then employ a 1D convolutional layer (with kernel size 1) to adjust the channel number from H_{fuse} to hidden dimensional d_m and capture local features and patterns. Next, the processed $F' \in \mathbb{R}^{B \times d_m \times T'}$ is chunked into N_c overlapped chunks with C as the length of each chunk for simplifying computation and capturing local features. The chunks $\mathbf{C} \in \mathbb{R}^{B \times C \times N_c}$ are then fed into the key component, two repeated computation blocks, to capture the intra-chunk features and inter-chunk features for generating the mask m. Figure 10(b) shows the details of the computation block. *EarSE* leverages two transformer encoders [83]

(a) Logitech G733 (b) ATH-G1WL (c) XM4 + Antlion

Fig. 11. Three devices. The Antlion is a modular microphone which can be magnetically adhered to the left side or the right side of a headphone (*e.g.*, Sony WH-1000XM4).



(a) Anechoic chamber (b) Setting

Fig. 12. Data collecting setting. The anechoic chamber can reduce echos and reverberations, which contributes to obtain the clean speech as the ground truth.

in each computation block for extracting the short-term dependencies (Intra-Transformer) and the long-term dependencies (Inter-Transformer). As Figure 10(c) shows, each transformer encoder contains eight encoder layers. The transformation can be derived as:

$$z' = \text{Dropout}\left(\text{MHA}\left(\text{Norm}\left(z\right)\right)\right) + z,$$

$$z_{out} = \text{Dropout}\left(\text{FFN}\left(\text{Norm}\left(z'\right)\right)\right) + z',$$

where z is the input of each encoder layer. Then, z_{out} is sent into a Parametric Rectified Linear Unit (PReLU) layer to increase the nonlinear capability. After conducting an overlap-add operation [52], the computed chunks $\mathbf{C}'$ are converted back to $F'' \in \mathbb{R}^{B \times d_m \times T'}$. Finally, a 1D convolutional layer followed by a ReLU layer is used to generate a mask $m \in \mathbb{R}^{B \times N \times T'}$ for F_s.

7.5 Transfer Learning and Dynamic Mixing.

Since the proposed DNN model is a powerful transformer-based neural network that requires sufficient training data to feed it, this introduces high deployment overhead and labor-intensive data collection. To tackle this issue, we leverage a public dataset [27] to create a WSJ0-2mix dataset, which is created by randomly mixing the speech of two users for training and evaluating speech separation models. Thereafter, we use the proposed deep neural network (*i.e.*, Encoder, Masking Net, and Decoder module), excluding the ViT and Multi-modal Fusion module, to perform a speech separation task on the WSJ0-2mix dataset. We set the output channel number of the last "Conv1d" to 2 in the Masking Net module, generating two masks for speech separation. Using transfer learning techniques to learn the prior knowledge as a pre-trained model for *EarSE*, the training time will be significantly reduced. Moreover, we implement a dynamic mixing framework for model training instead of creating mixed training data in advance. Compared with the long initial bootstrapping period introduced by creating the entire training dataset in advance, dynamic mixing enables a quick start-up of model training.

8 EVALUATION

In this section, we first describe the data collection process and experiment settings, followed by the comparison with seven baselines. Subsequently, we conduct comprehensive evaluations to assess the impact of different factors, followed by an ablation study. Next, we verify the practicality of *EarSE* in real-world applications and evaluate the impact of auxiliary spacers from three aspects (*i.e.*, privacy leakage, sound quality, and noise isolation). Finally, we conduct a user study to gather user feedback.

8.1 Data Construction and Experiment Settings

Table 2. The *EarSE*-BooMic dataset is consist of four sub-datasets, with a total of 21 participants after excluding overlap. Each sub-dataset serves a specific evaluation purpose: the 1st evaluates user diversity; the 2nd evaluates differences across devices; the 3rd evaluates the side of the modular mic; and the 4th evaluates re-wearing. (❍–Not Included, ●–Included).

Sub-dataset	Participant Number	Device			Side		Recording Time (h)
		Antlion+XM4	G733	G1WL	L	R	
1	18	●	❍	❍	●	❍	9
2	2	●	●	●	●	❍	3
3	4	●	❍	❍	●	●	4
4	3	❍	●	❍	●	❍	3

8.1.1 Data Collection. We invited 21 fluent English speakers (9 female, 12 male, with ages ranging from 18 to 27) in our study and collected a 19-hour dataset[2], called the *EarSE* Boom-Microphone-based (*EarSE*-BooMic) dataset. As illustrated in Figure 12, each participant was instructed to speak a minimum of 300 sentences (approximately 30 minutes) from the TIMIT speech corpus [28], using one of three boom-microphone-equipped headphones (Figure 11): Antlion Mod Microphone [3] mounted on Sony WH-1000XM4 [79], Logitech G733 [50], or Audio-Technica ATH-G1WL [4]. These recordings took place in an anechoic chamber to ensure optimal audio quality. The details of the *EarSE*-BooMic dataset are summarized as Table 2. During the data collection, we utilized a laptop (MacBook Pro with M1 chip) with the Adobe Audition 2023 software to collect these data. The chirp sequence is pre-modulated and is played at the same time. The boom/mod microphone is positioned on the left side of the headphone, and the ultrasonic audio is played by the right speaker of the headphone. The chirp signal leaked from the right ear pad, together with the user's spoken voice, was captured by the boom microphone simultaneously. We segmented the raw audio into 5-second clips to ensure a fair comparison [14, 85]. Overall, we collected 13,680 five-second clean speech segments.

8.1.2 Data Synthesis. After constructing the clean speech data, we introduce the detailed manipulations [14, 85] to generate datasets used in this research. For the convenience of the fairness of the later comparisons, we adhere to the methodology used in UltraSE and UltraSpeech. In order to generate the *EarSE*-BooMic dataset, we first randomly split the recorded clean speech data into training, validation, and testing datasets in the ratio of 70%, 20%, and 10%, respectively. To synthesize samples in the dataset, we linearly combine four types of audios, including background noise S_{noise} from the WHAM! dataset [96], interfering speech S_{is} from the WSJ0 dataset [27], the perturbed ultrasonic audio U_i, and recorded clean speech S_i (ground truth). This process is repeated 20 times for each clip of clean speech, which guarantees that each clean speech is mixed with a sufficient amount of interference settings. Overall, the initial SiSDR, STOI, and PESQ of *EarSE*-BooMic dataset are 4.86, 0.73, and 1.87. And the training dataset comprising 273k five-second segments of noisy speech (amounting to over 380 hours).

8.1.3 Experiment Setting. We implement the DNN model using the PyTorch framework [66] and train the neural network on a desktop equipped with an NVIDIA RTX 3090 GPU, featuring 24 GB of memory, and an AMD Ryzen Threadripper PRO 3955WX 16-Core processor. We adopt SiSNR as the objective function and utilize the Adam optimizer with an initial learning rate of $1.5e{-}04$ and the maximum number of training epochs is set to 200. When the number of training epochs exceeds 85, the learning rate will be reduced by multiplying it by 0.5 every time the loss value has not improved for two consecutive epochs until the learning rate reaches the minimum value ($1e{-}06$). The batch size is set to 1.

[2]Ethical approval has been obtained (No. H002969).

We use four common metrics to evaluate speech quality and intelligibility, and introduce the Subjective Mean Opinion Score (MOS) as an additional evaluation metric for evaluating subjective perception:

- SiSNR [53] within $(-\infty, \infty)$: Scale-invariant Signal-to-Noise Ratio. SNR (Signal-to-Noise Ratio) is a measure of the ratio between the signal power and the noise power in the acoustic signal. Scale-invariant Signal-to-Noise Ratio (SiSNR) is a variant of SNR that compensates for the difference in the amplitude scale of the original and reconstructed signals. Unlike SNR, SiSNR is insensitive to the scaling factor of the signals, which makes it a more robust measurement of speech quality.
- SiSDR [46] within $(-\infty, \infty)$: Scale-invariant Signal-to-Distortion Ratio. Roux *et al.* argue that the Signal-to-Distortion Ratio (SDR [89]) has been improperly used especially in the case of single-channel separation, resulting in misleading results [46]. They proposed scale-invariant SDR (SI-SDR), a slightly modified version of SDR that can overcome the shortages of SDR in various cases.
- STOI [87] within $[0, 1]$: Short-Time Objective Intelligibility, which is a measurement of speech quality that evaluates the intelligibility of speech signals in the presence of noise or distortion.
- PESQ [72] within $[1, 5]$: PESQ-Low Quality Option (PESQ-LQO), which is a variant of the Perceptual Evaluation of Speech Quality (PESQ [73]) algorithm that is optimized for low-quality speech signals.
- MOS within $[1, 5]$: Subjective Mean Opinion Score (MOS), which is a measure of speech quality obtained through human listeners' subjective testing. The score for MOS ranges from 1 (poor) to 5 (excellent). By defining the noisy speech as the worst (*i.e.*, 1) and the ground truth as the best (*i.e.*, 5), we invite 30 participants to evaluate the enhanced speech and calculate the average value.

8.2 Overall Performance

In this section, we compare *EarSE* with seven advanced baselines. For a fair comparison, we re-implemented the seven baselines. Then, we train and test these models on the 1st sub-dataset (Table 2) of *EarSE*-BooMic collected by Antlion+XM4. The seven baselines are listed as follows:

- SepFormer [83]: SepFormer is an RNN-free, attention-based neural network for speech separation and enhancement. The SepFormer employs a multi-scale approach that uses transformers to learn short- and long-term dependencies. It inherits the parallelization advantages of Transformers and achieves advanced performance in single-modal methods.
- AvaTr V2 [34]: AvaTr is an "avatar" and attention-based speaker extraction neural network, which can be applied in speech separation and enhancement. It assumes that a very short reference speech (speaker ID) is available from each target speaker [33]. The avatar summarizes the characteristics of the speaker, contributing to selective attention. We use the more advanced AvaTr V2 structure as our baseline.
- PHASEN [102]: PHASEN is a phase-and-harmonics-aware Deep Neural Network (DNN) for single-channel speech enhancement. It separates the predictions of amplitude and phase using two parallel streams.
- Conv-TasNet [54]: Conv-TasNet is a fully convolutional time-domain network designed for speech separation. It can model long-term dependencies.
- VoiceFilter [91]: VoiceFilter is a speaker-conditioned voice separation neural network that uses a trainable speaker recognition network to generate speaker-discriminative embeddings for filtering in target voice.
- UltraSpeech [14]: UltraSpeech uses the same hardware settings as UltraSE but improves performance on background noise reduction. This is achieved by introducing a complex neural network and a complex interaction module.
- UltraSE [85]: UltraSE is a single-channel, multi-modal speech enhancement solution based on COTS smartphones. It leverages the Doppler shift of reflected ultrasonic waves to capture the user's facial articulatory gestures as auxiliary information. It also incorporates a powerful two-stream DNN that can fuse the ultrasonic modality and speech modality features by concatenating them.

Table 3. Performance comparison with seven baselines.

Methods	SiSNR	SiSDR	STOI	PESQ	MOS
EarSE	**19.48**	**19.47**	**0.95**	**3.32**	**4.43**
UltraSE	15.42	15.74	0.85	3.01	4.03
UltraSpeech	13.52	13.52	0.81	2.87	3.47
SepFormer	17.14	17.13	0.89	3.02	4.10
AvaTr V2	17.00	17.00	0.94	2.83	3.87
PHASEN	13.57	13.63	0.82	2.94	3.67
Conv-TasNet	12.46	12.45	0.77	2.58	2.87
VoiceFilter	11.65	11.65	0.86	2.23	2.97
Noisy speech	6.07	4.86	0.73	1.87	1.00

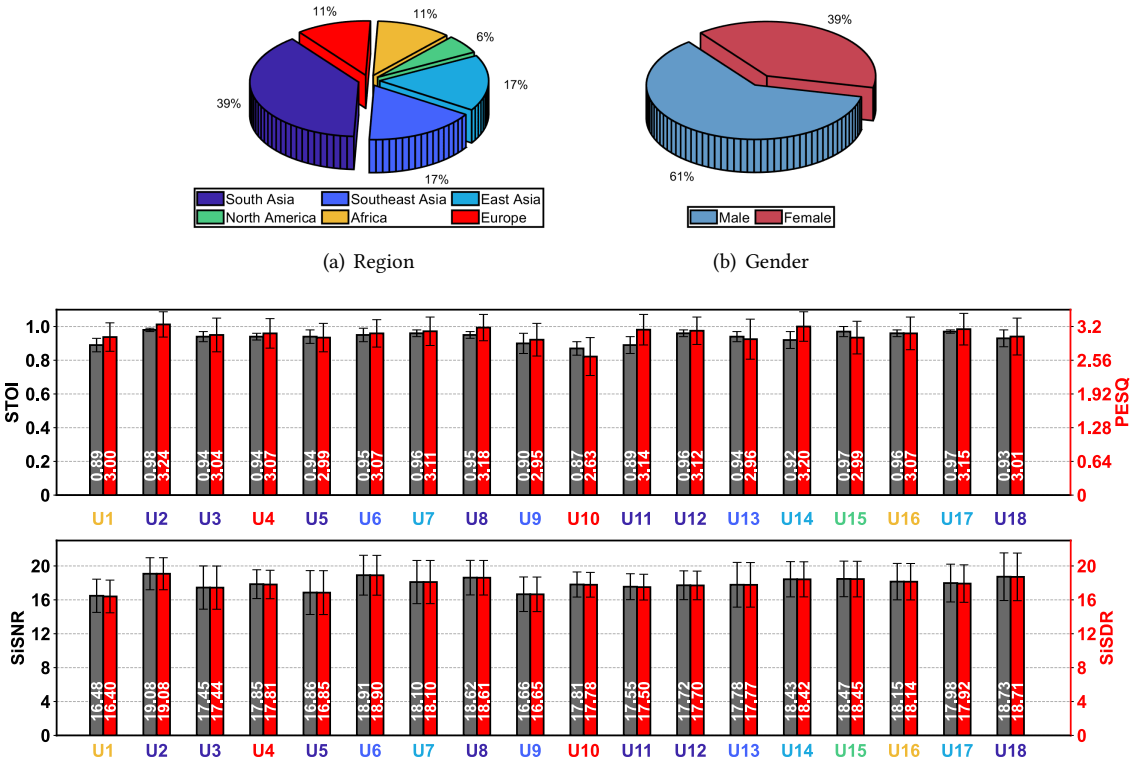
Table 3 shows the performance comparison with seven baselines. Overall, *EarSE* achieves 19.48 dB SiSNR, 19.47 dB SiSDR, 0.95 STOI, 3.32 PESQ, and 4.43 MOS. In comparison, *EarSE* outperforms the baselines by 2.34 dB–7.83 dB in SiSNR, 2.34 dB–7.82 dB in SiSDR, 1.1–23.3% in STOI, and 9.9–48.9% in PESQ. In a nutshell, *EarSE* outperforms these baselines due to its novel hardware setting, powerful transformer-based neural network, and the emulation of human selective listening capability. Compared with hand-held methods (UltraSpeech [14] and UltraSE [85]), the head-mounted property of *EarSE* provides a stable sensing field that can mitigate performance degradation caused by hand tremors and human activities. In addition, *EarSE* employs a fully attention-based neural network. The ViT module adaptively extracts informative areas from the ultrasonic modality features to adapt to different head shapes and sizes of users. The Multi-modal Fusion module designed based on the multi-head attention mechanism integrates auxiliary modality information to effectively enhance the user's speech even in extremely noisy environments. By combining the auxiliary modality and selectively attending to the user's speech, *EarSE* significantly outperforms the five single-modal solutions and performs robust speech enhancement even in noisy and complex acoustic environments.

8.3 Performance Impacted by Different Factors

We evaluate the performance of *EarSE* under the impact of different factors, we re-train models for each specific experiment setting. The general settings for each experiment are as described in Section 8.1.3. The results are integrated into Figure 13, 14, 15 and Tabel 4.

8.3.1 Impact of User Diversity. To evaluate the performance of *EarSE* on different users, we use the 1st sub-dataset (Table 2) collected from 18 participants from different regions in both genders, as shown in Figure 13(a) and 13(b). We conduct training and testing on each participant individually. Figure 13(c) present the results, indicating that *EarSE* can improve the speech quality of these users with relatively consistent performance. We observed that the User 1's African dreadlocks obstruct the gap created by auxiliary spacers, leading to a lower SiSNR and SiSDR. Additionally, User 10's soft speech, produced by facial articulatory gestures with shorter displacement, results in less effective auxiliary information and lower STOI and PESQ scores. Despite these challenges, *EarSE* significantly improves SiSNR, SiSDR, STOI, and PESQ, showing its ability to accommodate various accents and cater to diverse users.

8.3.2 Impact of Headphones Hardware. We evaluate *EarSE* on three devices, including two COTS devices (Logitech G733 [50] and Audio-Technica ATH-G1WL [4]) and a self-assembly system using a COTS headphone Sony WH-1000XM4 [79] and an Antlion Mod Microphone [3]. We use the 2nd sub-dataset (Table 2) to evaluate the performance of different devices. Figure 14 shows that *EarSE* achieves at least 0.82 STOI, 2.99 PESQ, 16.74 dB



(a) Region (b) Gender

(c) The performance of *EarSE* on individual users

Fig. 13. User diversity. (a) and (b) show the demographics of users; (c) displays the STOI, PESQ, SiSNR, and SiSDR performance for each user. The color of the user's text corresponds to the region in (a) (*e.g.*, User 1 is from Africa).

SiSNR, and 16.74 dB SiSDR on all of the devices. The Antlion+XM4 achieves the best performance (18.98 dB SiSNR, 3.28 PESQ, and 0.92 STOI), which can be attributed to the Sony WH-1000XM4's high-quality speakers with a frequency response range of 4 Hz–40,000 Hz, leading to minor signal distortions when playing chirps.

8.3.3 Impact of Position of Modular Mic. In this experiment, we investigate the impact of positioning a mod microphone on different sides of the user's head. Typically, modular microphones are placed on the left side. We emit ultrasonic in left ear and ask the four participants to switch the detachable modular mic to the right side and repeat the experiment using Antlion+XM4 prototype. The collected data is the 3rd sub-dataset (Table 2). Figure 14 shows that *EarSE* is robust to different position of the boom/mod mic. Although there is a slight difference in performance between the left and right sides, the system achieves at least 18.57 dB in SiSNR, 18.56 dB in SiSDR, 0.89 in STOI, and 3.14 in PESQ for both configurations. This suggests that *EarSE* can effectively adapt to user preferences and physical constraints in head-mounted headphone scenarios.

Additionally, we also evaluate the impact of the position of the modular microphone. We first train a model for *EarSE* using data collected from an initial position **O**. Then we adjust the malleable arm of the modular microphone along two orthogonal axes by a displacement of approximately 1.5 cm. The new positions are shown

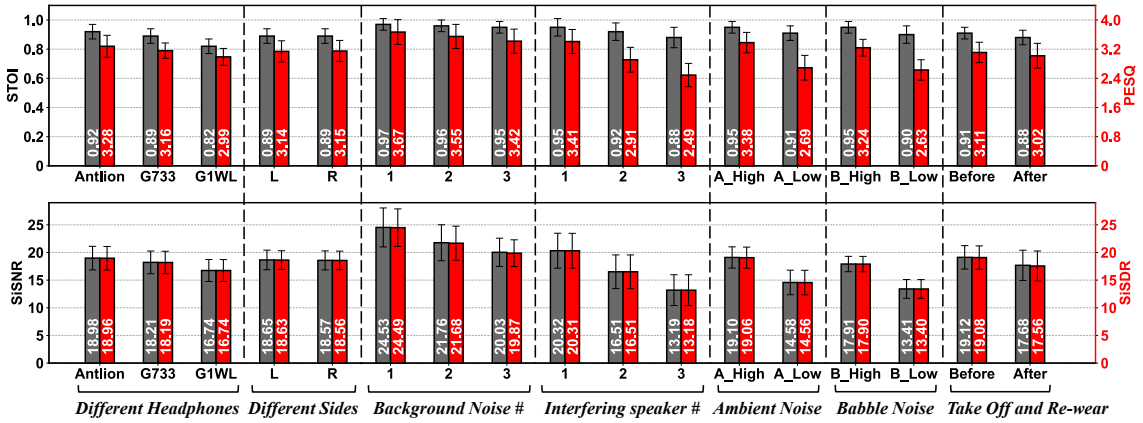
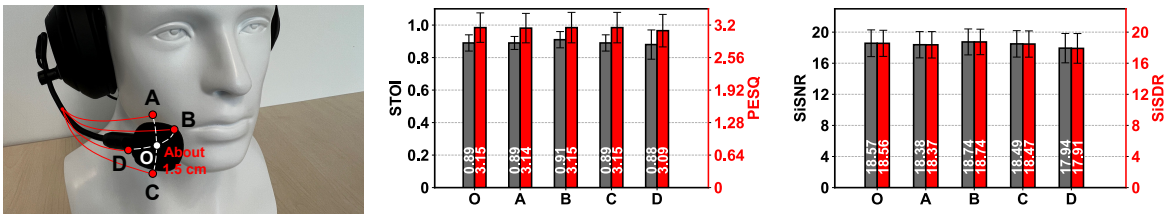


Fig. 14. Micro-benchmarks evaluation results.



(a) Different Positions (b) STOI and PESQ performance (c) SiSNR and SiSDR performance

Fig. 15. Impact of different positions of Modular Microphone.

as **A**, **B**, **C**, **D** in Figure 15(a), we collect 3 min testing data from each position under the same setting with the 3rd sub-dataset. Then, we use the well-trained model obtained from position **O** to evaluate the performance of the other four positions. The results are shown in Figures 15(b) and 15(c). We observe that the STOI for all four positions is consistently around 0.89, indicating that the intelligibility of enhanced speech is hardly affected by short-distance microphone displacements. The PESQ, SiSNR, and SiSDR all exhibit different performance, depending on positions. The model performs best at the position **B**. This is attributed to the fact that a closer distance between the microphone and the user's mouth can improve the quality of the processed audio, resulting in a higher SiSNR.The performance is the worst at the position **D**, which is the farthest from the user's mouth. This is caused by perturbations of the same size occupying a smaller proportion in a larger sensing field, causing a decrease in the resolution of the informative area. The positions **A** and **C** are roughly on par, with position **C** showing a very slight performance advantage because the acoustic sensing field constructed by the speaker and position **C** can cover the user's articulatory area more extensively. The results demonstrate that *EarSE* extracts the user's speech from numerous sound sources, relying on acoustic features with high synchronicity and correlation while being insensitive to the position and resolution of the informative area.

8.3.4 Impact of Noise Levels. We also investigate the performance of *EarSE* at various noise levels using the 1st sub-dataset (Table 2), the results are integrated in Figure 14. In this experiment, we increase the number of simultaneous background noise sources and interfering speakers from one to three, respectively. The result shows that *EarSE*, when only facing background noise, achieves at least 20 dB in SiSNR and SiSDR. Even with three

Table 4. The performance of different train-test splitting methods.

Methods	SiSNR	SiSDR	STOI	PESQ
EarSE (seen)	19.48	19.47	0.95	3.32
EarSE (unseen time period)	19.23	19.23	0.95	3.32
EarSE (unseen users)	17.82	17.83	0.92	3.21

simultaneous interfering speakers, *EarSE* still achieves a 13.19 dB in SiSNR. To evaluate *EarSE*'s performance under two types of noise (ambient noise [96] and babble noise [68]) with varying SNR, we divide the range from −3 dB to 10 dB into two levels—high SNR ([3.5 dB, 10 dB]) and low SNR ([−3 dB, 3.5 dB]). Subsequently, we generate a value k randomly drawn from a normal distribution within each range, which serves as the SNR for each synthetic sample. By adjusting the noise intensity, synthetic audio achieves an SNR of k by linear superposition. Then, we evaluate the performance of *EarSE* under the two types of noise conditions (high SNR and low SNR) for ambient and babble noise. The results show that *EarSE* performs slightly better (about 1 dB in SiSNR) in ambient noise and significantly better with high-SNR audio. The babble noise [68] used in this section is extremely challenging for speech enhancement due to the complex acoustic composition of the mixture of ten sound sources of varying intensity. In this extreme scenario, *EarSE* delivers speech enhancement that consistently exceeds 13.40 dB SiSNR.

8.3.5 Impact of Taking off and Re-wearing. Slight differences in the relative position between the head and device may occur if the user takes off and re-wears the device. We use the 4^{th} sub-dataset (Table 2) to evaluate the impact of such user behaviors. During data collection, participants were instructed to continuously record for 30 min for training, validation, and testing a model, the results are represented as "Before" in Figure 14. Then, We then collected three 10 min speech data sections, with participants instructed to remove the device for a break and re-wear it between each section. We evaluated the trained model's performance on the three sections directly without any training, and the average results from the three segments is represented as "After" in Figure 14. The results show that re-wearing the device may cause a 1.5 dB performance degradation, as the minute differences in the wearing position introduce a domain shift problem.

8.3.6 Impact of Unseen Time Period and Unseen Users on Model Performance. To evaluate the effectiveness of *EarSE*, we need to guarantee the training and testing datasets have a similar distribution and avoid the impact of domain shift problem [84] on the model's generalization. Therefore, we split the training and testing data in a random manner to evaluate the aforementioned sections. It helps to prevent the overconcentration of certain segments in training and testing sets within specific time periods or with particular characteristics (*e.g.*, the mild hoarseness caused by prolonged speaking is typically concentrated in the latter half of each participant's data), thus enhancing the generalization performance of the model. However, the model trained with the training data split according to time sequence may be more approximate to performance in the real world. Therefore, we change the data splitting method and assess the impact on performance. Furthermore, we evaluated the performance of *EarSE* on unseen users by performing leave-one-user-out cross-validation.

As Table 4 shows, splitting the training test data in a random or time-order manner has no impact on the model performance. The model trained in a time-order manner exhibits a 0.25 dB decline in SiSNR and SiSDR. As *EarSE* operates by sensing facial articulatory gestures, and given the significant variance in 3D facial structures across individuals, the performance of unseen users is inferior to that of "registered" users. However, by fine-tuning a base model, we can significantly reduce the overhead of training a personalized model. Specifically, we leave one participant out as an unseen user and train a base model using the remaining participants. We found that by increasing the amount of data used for fine-tuning, *EarSE* can achieve a speech enhancement performance

Table 5. Ablation study of *EarSE*.

Methods	SiSNR	SiSDR	STOI	PESQ
EarSE	**19.48**	**19.47**	**0.95**	**3.32**
EarSE (w/o Multi-modal Fusion)	17.26	17.24	0.89	3.17
EarSE (w/o auxiliary modality)	14.69	14.69	0.84	3.05

of 18.61 dB SiSNR on a new user with only 3 min of user-dependent data. On average, it requires just 5 min of user-dependent data to achieve comparable performance with training from scratch.

8.4 Ablation Study

To evaluate the effectiveness of key components of *EarSE*, we design two baseline versions of *EarSE* and repeat the same experiment presented in Section 8.2. In the first version, we do not use the ultrasonic modality as the auxiliary modality (w/o auxiliary modality). This model only consists of the Encoder, Masking Net, and Decoder modules in *EarSE*. This version evaluates the efficacy of the transformer-based DNN and the significance of the ultrasonic modality. In the second version, we exclude the Multi-modal Fusion module (w/o Multi-modal Fusion), meaning that we concatenate the features from the speech and ultrasonic modalities without any fusion or gating control. This version evaluates the effectiveness of the multi-modal fusion method proposed in Section 7.3. Table 5 demonstrates that, although the version *EarSE* (w/o auxiliary modality) improves performance in speech enhancement to some extent, the results are unsatisfactory compared with *EarSE*. This is because the model suffers from the label permutation problem in the absence of auxiliary information. In comparison, version *EarSE* outperforms *EarSE* (w/o Multi-modal Fusion) by 2.26 dB in SiSNR, 0.06 in STOI, and 0.15 in PESQ, respectively, which demonstrates the efficacy of the proposed Multi-modal Fusion module. Both the ultrasonic modality and the proposed multi-modal fusion method significantly contribute to the speech enhancement performance of *EarSE* by addressing the label permutation problem.

8.5 Real-World Application

In this section, we evaluate the performance of *EarSE* in real-world applications. As the user speech is polluted before being recorded by the microphone in real-world situations, obtaining clean speech (ground truth) for evaluating the metrics used in Section 3 is challenging. Therefore, we assess *EarSE*'s real-world performance using word error rate (WER), which represents the percentage of minimum word-level edit operations (insertions, deletions, or substitutions) required to transform the recognized text (hypothesis) into the reference text (ground truth). Specifically, we asked ten users to speak ten sentences from the TIMIT speech corpus and five sentences not included in the TIMIT corpus while using *EarSE* and engaging in various human activities across different scenarios. By comparing the WER obtained from the noisy speech (w/o SE solutions), enhanced speech (w/ SE solution), and text ground truth, we can evaluate *EarSE*'s effectiveness in enhancing the quality and intelligibility of users' speech in noisy environments. We also compare *EarSE*'s performance in speech recognition with two industry solutions, AirPods Pro [35] and Galaxy Buds2 Pro [75], to demonstrate the practicality of *EarSE*.

8.5.1 Automatic Speech Recognition System Used for Calculating WER. This study uses the Speech Recognition library [104], an up-to-date speech recognition engine, as the Automatic Speech Recognition (ASR) system. The ASR system utilizes the Google Cloud Speech API to convert spoken language into text. We choose this engine as the core technology of our ASR system because it is a stable, accurate, and continuously updated speech recognition engine with over 7.3 k stars on GitHub. The ASR system was used to convert both recorded noisy speech and enhanced speech into text. The predicted text is then compared with the ground-truth text to calculate

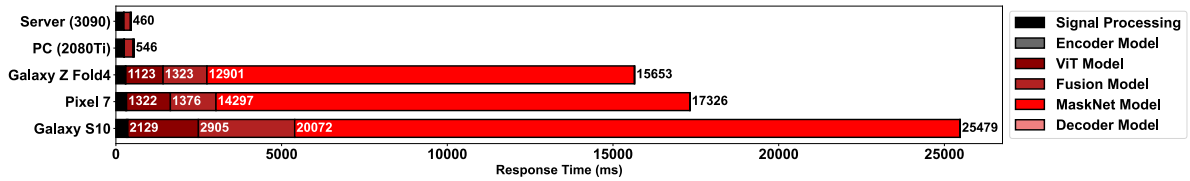


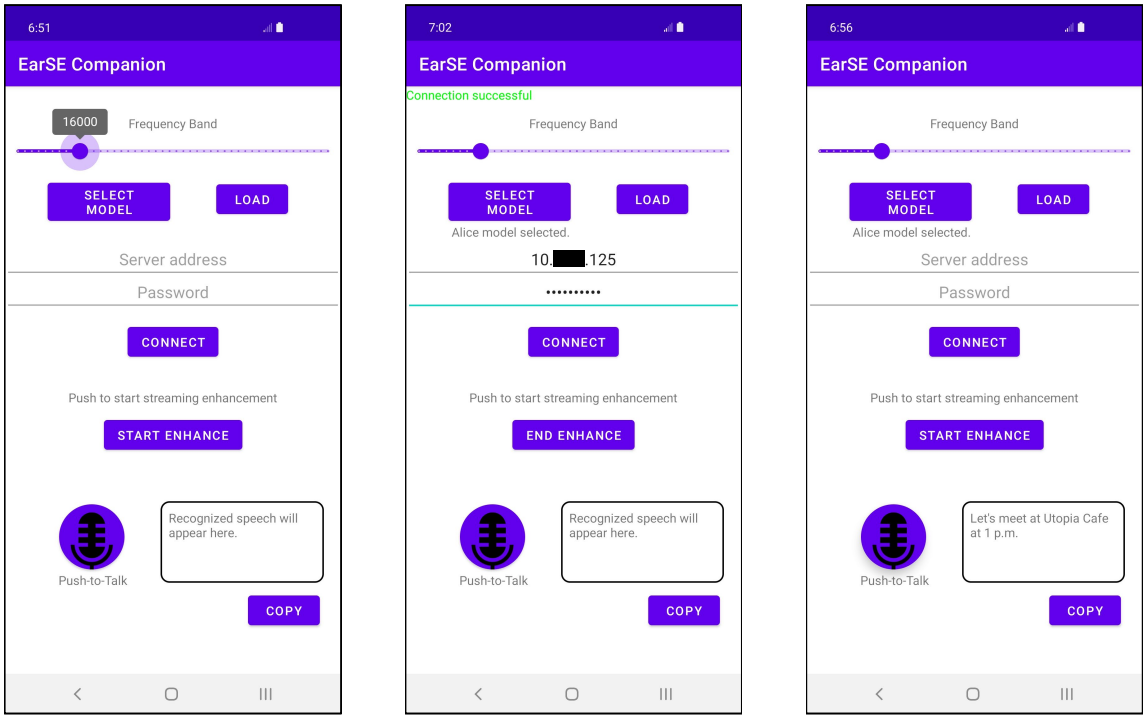
Fig. 16. The computational delay of *EarSE* on mobile devices.

the WER as the evaluation metric using the following formula: WER = $\frac{S+D+I}{N}$, where S represents the number of substitutions, D stands for the number of deletions, I indicates the number of insertions, and N denotes the total number of words in the reference text. Note that if an audio segment is too noisy to preserve intelligibility, the engine may recognize it as "could not understand". In such cases, we define the WER of the audio as 100%. By using the ASR system, the WER can reflect practical performance in real-world speech recognition and indirectly indicate the effectiveness of *EarSE* in speech enhancement, even in the absence of perfect reference speech.

8.5.2 Practicality Analysis and Sample Applications. We evaluate the computational delay of *EarSE* on five devices: three mobile devices (Samsung Galaxy S10, Google Pixel 7, Samsung Galaxy Z Fold4), a PC equipped with an RTX 2080Ti GPU, and a cloud server equipped with an RTX 3090 GPU. We measured the computational delay across six distinct components (*i.e.*, signal processing, Encoder model, ViT model, Fusion model, MaskNet model, and Decoder model), executing each device ten times and using the mean value as the measurement outcome. The results are visualized in Figure 16.

We observe that the execution time for the encoder and decoder models is negligible, even on the least powerful device (Samsung Galaxy S10). While the signal processing part is about 300 ms, indicating that signal processing does not cause a significant bottleneck to run *EarSE* on mobile devices. However, the computational delay of the MaskNet model consumes a significant portion of the total latency (accounting for 79–83% across three smartphones). In comparison, the introduction of CIR profiles and ViT module as an auxiliary modality for selective feature extraction and the multi-modal fusion does not introduce an excessive delay (approximately 2,500 ms on mobile devices). The size of the *EarSE* app on Android phones is 509 MB, with a maximum memory usage of 498 MB during computation. When deployed on a laptop, the maximum memory used during computation is 5,348 MB. Furthermore, we find that as devices undergo continuous upgrades, their computational capabilities improve incrementally, resulting in a significant reduction in calculation-induced delay. On the Z Fold4 launched in 2022, it is feasible to process a 5-second voice clip in 15.7 s without relying on additional computational resources (such as cloud servers). Furthermore, for mobile devices that have certain computing power, such as a laptop with a local RTX 2080Ti GPU, the processing time can be reduced to 546 ms. Moreover, with a cloud server equipped with an RTX 3090 GPU, the processing time can be further reduced to 460 ms. Also note that both UltraSE and UltraSpeech process 5-second segments as input due to the non-causal structure of the Bi-LSTM [85]. For a fair comparison, *EarSE* also adopts 5-second segments as input. However, since *EarSE* is fully based on transformers, which can be configured as either causal (*e.g.*, text generation) or non-causal (*e.g.*, text translation) structures, the algorithmic delay, also known as the initial bootstraping period, can be shorter than 5 s. Naturally, using shorter segments can significantly reduce the computational delay shown in Figure 16. By using the Overlap-Add (OLA) method [82], a method for efficiently concatenating audio frames, this latency value fully meets the requirements for real-time speech enhancement.

Based on the practicality analysis above, *EarSE* can be used for real-time speech enhancement on laptops and smartphones when network and cloud servers are available. However, when cloud servers are unavailable, using *EarSE* on mobile devices faces slowdown problems. In this case, we can still combine *EarSE* with push-to-talk functionality to facilitate speech recognition or voice messages. We build an Android application for *EarSE* named "*EarSE* Companion", as shown in Figure 17. Once paired with *EarSE* hardware, it enables various voice



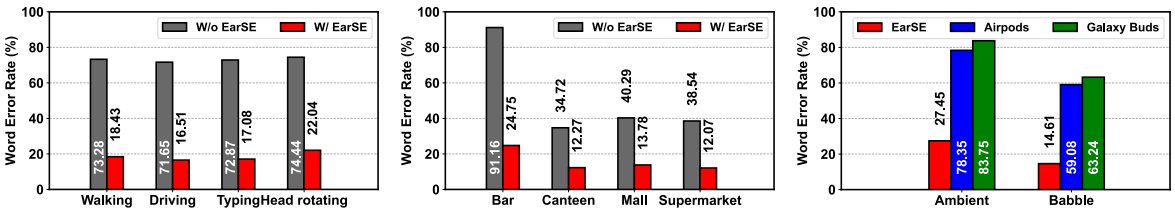
(a) Teenager mode (b) Streaming enhancement (c) Push-to-talk and voice input method

Fig. 17. Sample applications.

enhancement features. Considering teenagers' auditory sensitivity, we have designed a "Teenager mode" for the *EarSE* Companion app. As shown in Figure 17(a), the sliding bar controls the starting frequency of the FMCW signal, facilitating the collection of training data in the teenager mode and the activation of speech enhancement. Figure 17(b) shows how *EarSE* enhances streaming audio in conjunction with a cloud server. Upon entering the correct cloud server address and password, the smartphone offloads the captured noisy speech (including FMCW) to the cloud server for computation in real-time. The enhanced speech, post-processing, is returned to the smartphone via a protocol, serving as the input for the in-use microphone process. Figure 17(c) shows that in the absence of network connectivity, speech enhancement, and recognition can be achieved only using the smartphone's computing resource, offering users clean speech input in noisy environments. This feature remains available and provides an even better experience when the cloud server is accessible.

8.5.3 Impact of Human Activities. To assess the robustness of *EarSE* when users are performing various daily human activities, we asked users to engage in four activities (*i.e.*, walking, driving, typing, and head rotating) while recording their speech in a room under competition from one interfering speaker and loud music. The recorded audio which contains 15 sentences for each participant are enhanced by the model trained in Section 8.2. Subsequently, WER was calculated using the ASR system described in Section 8.5.1 as experimental results.

Figure 18(a) presents the results, which demonstrate that *EarSE* effectively addresses speech enhancement challenges across a variety of real-life scenarios involving different human activities. During walking, *EarSE* exhibits minimal performance degradation, maintaining a high WER reduction. This highlights the system's robustness against the relative displacement issues encountered in smartphone SSE solutions. While driving and



(a) Different activities (b) Different scenarios (c) Comparison with industry solutions

Fig. 18. Real-world usage WER.

typing, *EarSE* enables users to communicate effortlessly without the need to hold a smartphone, providing a clear speech signal and liberating users' hands compared to handheld smartphone SSE solutions. This advantage makes *EarSE* more user-friendly and convenient for multitasking scenarios. In situations involving head rotations, such as face-to-face or hybrid online-offline conferences with multiple speakers, the head-mounted form factor of the headphones ensures the adaptiveness of *EarSE* in head movements.

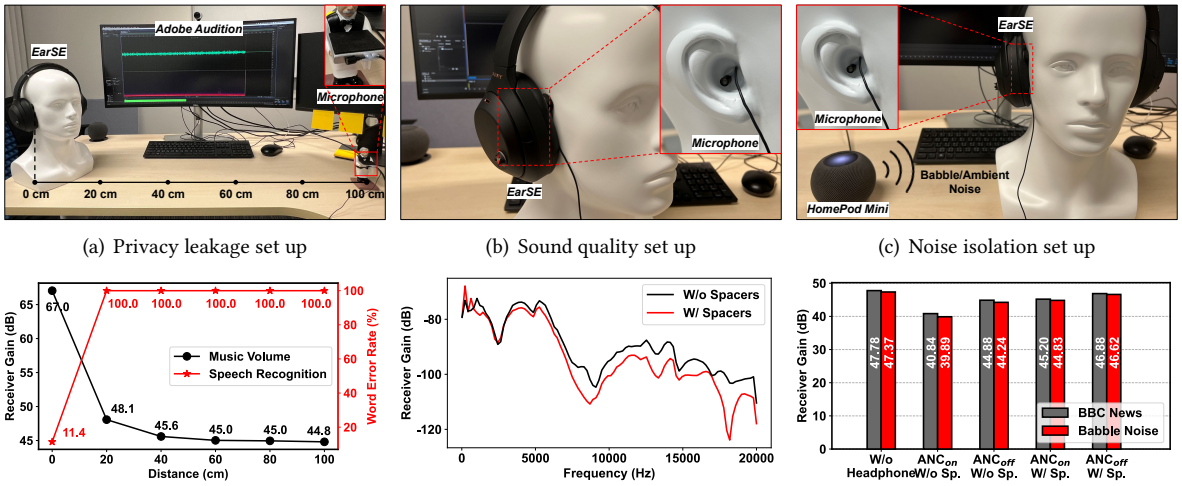
8.5.4 Impact of Environmental Scenarios. To investigate the effectiveness of *EarSE* in various environmental scenarios, we evaluate its performance in four common noisy scenarios (*i.e.*, bar, canteen, mall, supermarket) where users may need to communicate. In bars with loud audio, *EarSE* effectively suppresses background music and delivers clear speech signals. In situations with multiple speakers, such as canteens, shopping malls, and supermarket, the system isolates the user's speech from surrounding conversations, reducing the WER of user speech recognition significantly by 22.45–66.41% as shown in Figure 18(b).

8.5.5 Comparison with Industry Solutions. We compare *EarSE*'s performance in speech recognition WER with AirPods Pro and Galaxy Buds2 Pro in two noise settings. In the two settings, we use a HomePod mini [36] to play BBC news (ambient) and babble noise as interference at volume 50% directly in front of the user. Five participants are invited to speak 15 sentences (10 of them included in TIMIT and 5 of them not included) wearing different devices. Figure 18(c) shows that both AirPods Pro and Galaxy Buds2 Pro are vulnerable to the noise that contains human speech, due to the lack of multi-modal support in both earbuds. Furthermore, because of the proximity of the microphones used for beamforming and noise originating directly in front does not bring TDoA to the binaural microphones, they perform poorly in filtering out noise that contains human voices, resulting in clear noise being recorded in the audio. For babble noise, as its interference with ASR results is not as intractable as human speech, the WER for all three solutions is lower than ambient noise at the same volume. The WER for the speech enhanced by *EarSE* is only 14.61%, while AirPods Pro and Galaxy Buds2 Pro are similar, both around 61%. This is because neither of these two industry solutions effectively handles the noise coming directly from the front of the user.

8.6 The Impact of Auxiliary Spacers.

In this section, we evaluate the impact of the auxiliary spacers from three aspects: privacy leakage, sound quality, and noise isolation. The experiment settings and results are shown in Figure 19 (the audio playback of all devices is set to 50% volume).

8.6.1 Impact on Privacy Leakage. As shown in Figure 19(a), we fit the prototype of *EarSE* on a human head model and place a microphone close to *EarSE* to collect the leaked sound of *EarSE* at varying distances ranging from 20 cm to 100 cm. At each position, we evaluate the decibel value of the volume of leaked music and the WER of the leaked human speech recognized by the ASR system. Figure 19(d) shows that the volume of leaked music attenuates rapidly with increasing distance and reaches stable at 40 cm position. For speech recognition, the ASR



(a) Privacy leakage set up (b) Sound quality set up (c) Noise isolation set up

(d) Privacy leakage result (e) Sound quality result (f) Noise isolation result

Fig. 19. The impact of auxiliary spacers.

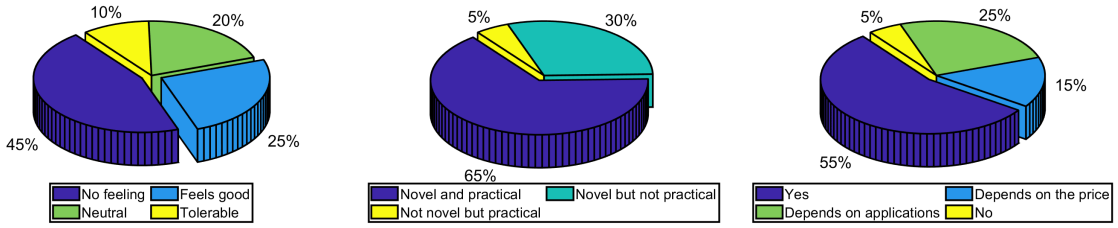
system cannot recognize the leaked speech with an extremely low SNR (defined as 100% WER). Both results demonstrate that the privacy leakage issue with *EarSE* is not severe.

8.6.2 Impact on Sound Quality. We modulate a 10 s chirp signal ranging from 20 Hz to 20,000 Hz to test receiver gain in different frequency bins. We use hot melt adhesive to fix a microphone to the external auditory meatus of a head model and use *EarSE* to play the modulated chirp signal. By analyzing the recorded audio, we observed that the receiver gains of the two settings have identical shapes in Figure 19(e), with only a slight decrease in gain at high frequencies (over 5,000 Hz) when the spacers were attached. This difference became more noticeable at even higher frequencies (over 8,000 Hz). Since human voice and music typically occupy frequencies below 8,000 Hz [9], the nearly overlapped receiver gains in this frequency range suggest that using spacers in *EarSE* would have a negligible impact on intelligibility. However, the high-frequency differences indicate that the rich high-order harmonics contained in some high-fidelity (Hi-Fi) audio may be affected by the spacers when heard by human ears.

8.6.3 Impact on Noise Isolation and Active Noise Cancellation (ANC). We use a HomePod Mini to play babble or ambient noise (BBC news) around *EarSE* (about 40 cm). Figure 19(f) shows that using auxiliary spacers leads to 73% and 60% performance degradation on noise isolation and ANC performance, respectively. However, the reason behind the degradation is that ANC perceives the external signal and emits it internally. The gain of the transmit compensation signal is calculated and constant. Using spacers alters the intensity of intrusive noise, whereas the compensation emitted by the ANC remains unchanged, resulting in an imbalance and leading to a decline in ANC effectiveness. However, this problem can be solved by adjusting the gain of the internal emission of ANC. We discuss the limitation and potential solutions in Section 9.3.

8.7 *EarSE* Form Factor User Study

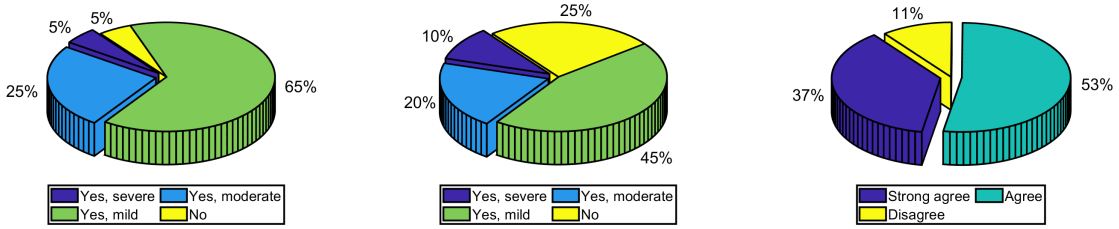
As elaborated in Section 5.1, the key idea behind *EarSE* is to create a sensing field formed by a boom/modular microphone and the opposite-side speaker to sense the user's articulatory gestures for enhancing their speech. To amplify the ultrasonic signal leaking from the ear pad, we place two auxiliary spacers on the ear pad, creating a gap for the ultrasonic waves to escape. These auxiliary spacers are made of a soft, lightweight material and are



(a) When using *EarSE*, has the auxiliary spacers caused you any discomfort?

(b) Do you think the design principle of *EarSE* is novel and practical?

(c) Would you be willing to pay for a detachable modular microphone for intelligent applications?



(d) Is there a degradation in audio quality during music playback, and if applicable, to what extent?

(e) Are the auxiliary spacers might cause privacy leakage issues, and if applicable, to what extent?

(f) If you give positive answers in Q4 or Q5, do you think the designs shown in Figure 21 can solve the problem?

Fig. 20. *EarSE*'s user experience and user study.

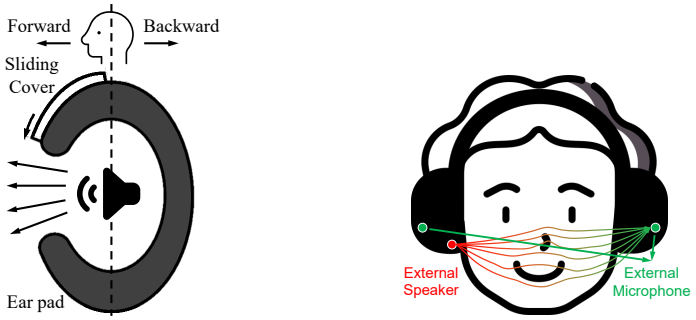
designed to integrate seamlessly with the ear pad, minimizing any potential discomfort for the user. We also gathered feedback of three questions regarding the comfort and usability of *EarSE* from 20 volunteers (10 of them are participate in data collection while others are unseen users). The three questions are:

- **Q1:** When using *EarSE*, has the auxiliary spacers caused you any discomfort?
- **Q2:** Do you think the design principle of *EarSE* is novel and practical?
- **Q3:** Would you be willing to pay for a detachable modular microphone for intelligent applications?

As the pie chart 20(a), 20(b), and 20(c) shows, 70% users regard *EarSE* to be comfortable to wear and 95% users regard *EarSE* to be novel. Furthermore, 55% of users make it clear that they are willing to spend extra money to buy a modular microphone to realize intelligent applications, while 40% of users are hesitant about the price or applications. Overall, the participant feedback indicated that the shape and material of the auxiliary spacers did not have a negative impact on their comfort or experience while wearing the headphones. The innovative concept of *EarSE—enabling the ability to sense the user's facial articulatory gestures by using the form factor of COTS headphones equipped with a boom microphone*—has promising market potential.

To evaluate the impact of auxiliary spacers on the listening experience and privacy consideration, we randomly invite 20 volunteers to listen music using *EarSE* and complete the following three questions:

- **Q4:** Is there a degradation in audio quality during music playback, and if applicable, to what extent?
- **Q5:** Are the auxiliary spacers might cause privacy leakage issues, and if applicable, to what extent?
- **Q6:** If you give positive answers in Q4 or Q5, do you think the designs (proposed in Section 9.3 as potential solutions) shown in Figure 21 can solve the problem?



(a) Design 1 (b) Design 2

Fig. 21. Commercial designs.

As the pie chart 20(d), 20(e), and 20(f) show, more than 70% of volunteers think the auxiliary spacers have negligible impact on the listening experience. While for audiophiles with a keen sensitivity to sound quality and pursuit of high fidelity (30%), there is a noticeable decline in audio quality. For privacy leakage, only 25% of volunteers consider that the gap created by auxiliary spacers will not lead to privacy leakage, and 75% of the volunteers believe that there are varying degrees of privacy breach issues. 89% of the volunteers think that the designs shown in Figure 21 can solve the problems to different extents.

9 LIMITATIONS, POTENTIAL SOLUTION, AND FUTURE WORK

In this section, we discuss the limitations of *EarSE* and possible solutions that are worth the effort in the future.

9.1 Face-touching behaviors

EarSE establishes an acoustic sensing field across the user's face to capture their articulatory gestures, which is vulnerable to the user's face-touching behaviors. Although we noted that face touching was not often the case during speaking in our observations, mitigating its effect remains a worthwhile future direction. A potential solution is to detect face-touching behaviors before conducting speech enhancement and suspend speech enhancement to avoid distortions in the output audios once these behaviors are detected.

9.2 Deployment on Resource-Constrained Mobile Devices

In Section 8.5.2, we observe that the algorithmic delay can be up to 15.7 s on resource-limited devices (*e.g.*, smartphones), apparently affecting the real-time experience. To improve the user experience on these resource-limited devices, we propose several possible solutions: (1) to reduce computational complexity further, one can consider using model compression techniques such as model pruning [43, 56] or knowledge distillation [48, 77] to reduce the number of parameters; (2) offload enhancement processing to a powerful cloud server; (3) deploy *EarSE* on mobile devices in a push-to-talk manner, such that it can be applied to instant messaging applications. It reduces the requirement for immediacy and yields high-quality user speech on resource-constrained mobile devices, requiring only a short processing period.

9.3 The Form Factor of *EarSE*

Several challenges exist in transitioning from the prototype of *EarSE* proposed in this paper to a commercially valuable product, including the degradation of ANC functionality and noise isolation caused by auxiliary spacers, the risk of privacy leakage from the headphone speaker, and the rarity of boom mics.

To further address the limitations caused by the use of auxiliary spacers, the prototype in this paper can be improved to the design shown in Figure 21(a), featuring a specially designed ear pad with a directional gap and a sliding cover. The inner side is made of the same sound-insulating material as the ear pad and is kept closed when not used. A user study conducted for this design yields positive feedback and can somewhat mitigate these issues. Moreover, as Figure 21(b) shows, an advanced commercial solution involves installing a low-frequency ultrasonic speaker (or an add-on module) in an earcup, which emits towards the opposite side, and using the external microphone (for ANC) in the other earcup as a receiver to construct an acoustic sensing field. Note that external microphones take low-frequency ambient sound as cues for ANC, which will not be affected by the emitted ultrasound. This design can bypasses the issues related to privacy leakage, ANC performance degradation, and reliance on boom/modular microphones while achieving robust speech enhancement.

10 CONCLUSION

In this paper, we propose *EarSE*, the first robust, hands-free, multi-modal speech enhancement solution based on COTS headphones equipped with a boom microphone and a fully attention-based DNN. By leveraging the form factor of the headphones and well-designed FMCW, we establish a stable acoustic sensing field across the user's face to capture the subtle facial articulatory gestures as auxiliary information for speech enhancement. A comprehensive evaluation (including three devices, 21 participants from 11 countries, and 19 hours of data recording) demonstrates that *EarSE* outperforms seven baselines in five evaluation metrics. A follow-up study evaluates the stability of *EarSE* under the impact of various factors. Then, we verify the practicality of *EarSE* by conducting an in-depth analysis of the computational delay on mobile devices and a cloud server. We also design sample applications to show how *EarSE* will be applied in the real world and test the performance of *EarSE* in the real world. The experimental results show that *EarSE* maintains stability and practicality and outperforms the seven baselines by 38.0% in SiSNR, 12.4% in STOI, and 20.5% in PESQ on average.

ACKNOWLEDGMENTS

The work was also supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 21201420 and CityU 11201422). The work was also partially supported by CityU APRC grant 9610471, CityU MFPRC grant 9680333, CityU SIRG grant 7020057, CityU SRG-Fd grant 7005666 and 7005984.

REFERENCES

[1] 2023. Krisp. https://krisp.ai/.
[2] M Abd El-Fattah, Moawad Ibrahim Dessouky, Salah Diab, and Fathi Abd El-Samie. 2008. Speech enhancement using an adaptive wiener filtering approach. *Progress In Electromagnetics Research M* 4 (2008), 167–184.
[3] Antlion Audio. 2023. *Antilion Mod Mic.* Retrieved April 6, 2023 from https://antlionaudio.com/collections/microphones/products/modmic-usb
[4] Audio-Technica. 2023. *ATH-G1WL.* Retrieved April 6, 2023 from https://www.audio-technica.com/en-us/ath-g1wl
[5] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).
[6] BioDigital. 2022. *Face: Superficial Fascia of the Face and Facial Nerve.* Retrieved December 17, 2022 from https://human.biodigital.com/view?id=production/maleAdult/face_superficial_fascia_of_the_face_and_facial_nerve_quiz&lang=en
[7] Nam Bui, Nhat Pham, Jessica Jacqueline Barnitz, Zhanan Zou, Phuc Nguyen, Hoang Truong, Taeho Kim, Nicholas Farrow, Anh Nguyen, Jianliang Xiao, et al. 2019. ebp: A wearable system for frequent and comfortable blood pressure monitoring from user's ear. In *Proceedings of the 25th annual international conference on mobile computing and networking (MobiCom)*. 1–17.
[8] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2023. hEARt: Motion-resilient Heart Rate Monitoring with In-ear Microphones. In *Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE, 200–209.
[9] Chao Cai, Rong Zheng, and Jun Luo. 2022. Ubiquitous acoustic sensing on commodity iot devices: A survey. *IEEE Communications Surveys & Tutorials* 24, 1 (2022), 432–454.

[10] Gaoshuai Cao, Kuang Yuan, Jie Xiong, Panlong Yang, Yubo Yan, Hao Zhou, and Xiang-Yang Li. 2020. Earphonetrack: involving earphones into the ecosystem of acoustic motion tracking. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys)*. 95–108.

[11] Ishan Chatterjee, Maruchi Kim, Vivek Jayaram, Shyamnath Gollakota, Ira Kemelmacher, Shwetak Patel, and Steven M Seitz. 2022. ClearBuds: wireless binaural earbuds for learning-based speech enhancement. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 384–396.

[12] E Colin Cherry. 1953. Some experiments on the recognition of speech, with one and with two ears. *The Journal of the acoustical society of America* 25, 5 (1953), 975–979.

[13] Marc Delcroix, Keisuke Kinoshita, Tomohiro Nakatani, Shoko Araki, Atsunori Ogawa, Takaaki Hori, Shinji Watanabe, Masakiyo Fujimoto, Takuya Yoshioka, Takanobu Oba, et al. 2013. Speech recognition in living rooms: Integrated speech enhancement and recognition system based on spatial, spectral and temporal modeling of sounds. *Computer Speech & Language* 27, 3 (2013), 851–873.

[14] Han Ding, Yizhan Wang, Hao Li, Cui Zhao, Ge Wang, Wei Xi, and Jizhong Zhao. 2022. UltraSpeech: Speech Enhancement by Interaction between Ultrasound and Speech. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 3 (2022), 1–25.

[15] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[16] Yariv Ephraim and David Malah. 1984. Speech enhancement using a minimum-mean square error short-time spectral amplitude estimator. *IEEE Transactions on acoustics, speech, and signal processing* 32, 6 (1984), 1109–1121.

[17] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. 2018. Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 1–11.

[18] Hakan Erdogan, John R Hershey, Shinji Watanabe, and Jonathan Le Roux. 2015. Phase-sensitive and recognition-boosted speech separation using deep recurrent neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 708–712.

[19] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Proceedings of the 2nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, Vol. 96. 226–231.

[20] SOZO Exchange. 2022. *Learn English with sozo exchange.* Retrieved December 17, 2022 from http://sozoexchange.com

[21] Xiaoran Fan, Longfei Shangguan, Siddharth Rupavatharam, Yanyong Zhang, Jie Xiong, Yunfei Ma, and Richard Howard. 2021. HeadFi: bringing intelligence to all headphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 147–159.

[22] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: gait-based user identification with in-ear microphones. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 337–349.

[23] Centers for Disease Control and Prevention. 2023. Public Health and Scientific Information. https://www.cdc.gov/nceh/hearing_loss/public_health_scientific_info.html Accessed: 2023-07-31.

[24] Susanne Fuchs, Martine Toda, and Marzena Żygis. 2010. *Turbulent sounds: An interdisciplinary guide*. Vol. 21. Walter de Gruyter.

[25] Yang Gao, Yincheng Jin, Jagmohan Chauhan, Seokmin Choi, Jiyang Li, and Zhanpeng Jin. 2021. Voice in ear: Spoofing-resistant and passphrase-independent body sound authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5, 1 (2021), 1–25.

[26] Yang Gao, Wei Wang, Vir V Phoha, Wei Sun, and Zhanpeng Jin. 2019. EarEcho: Using ear canal echo for wearable authentication. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 3, 3 (2019), 1–24.

[27] John Garofolo, David Graff, Doug Paul, and David Pallett. 1993. CSR-I (WSJ0) Complete LDC93S6A. *Web Download. Philadelphia: Linguistic Data Consortium* 83 (1993).

[28] John S Garofolo et al. 1988. DARPA TIMIT acoustic-phonetic speech database. *National Institute of Standards and Technology (NIST)* 15 (1988), 29–50.

[29] Bryan Gick, Ian Wilson, and Donald Derrick. 2013. *Articulatory phonetics*. John Wiley & Sons.

[30] Giuliana Grimaldi and Mario Manto. 2008. Tremor: from pathogenesis to treatment. *Synthesis lectures on biomedical engineering* 3, 1 (2008), 1–212.

[31] Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. 2023. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 14–27.

[32] Guoning Hu and DeLiang Wang. 2004. Monaural speech segregation based on pitch tracking and amplitude modulation. *IEEE Transactions on neural networks* 15, 5 (2004), 1135–1150.

[33] Shell Xu Hu. 2021. Avatar-based speech separation by Upload AI LLC. https://github.com/cajal/AvaTr/tree/v2.

[34] Shell Xu Hu, Md. Rifat Arefin, Viet-Nhat Nguyen, Alish Dipani, Xaq Pitkow, and Andreas Savas Tolias. 2021. AvaTr: One-Shot Speaker Extraction with Transformers. In *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*.
[35] Apple Inc. 2023. *AirPods Pro*. https://www.apple.com/airpods-pro/
[36] Apple Inc. 2023. *HomePod mini*. https://www.apple.com/homepod-mini/
[37] Nan Jiang, Terence Sim, and Jun Han. 2022. EarWalk: towards walking posture identification using earables. In *Proceedings of the 23rd Annual International Workshop on Mobile Computing Systems and Applications (HotMobile)*. 35–40.
[38] Yincheng Jin, Yang Gao, Xiaotao Guo, Jun Wen, Zhengxiong Li, and Zhanpeng Jin. 2022. EarHealth: an earphone-based acoustic otoscope for detection of multiple ear diseases in daily life. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 397–408.
[39] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: " Hearing" Your Silent Speech Commands In Ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 2 (2022), 1–28.
[40] Vimal Kakaraparthi, Qijia Shao, Charles J Carver, Tien Pham, Nam Bui, Phuc Nguyen, Xia Zhou, and Tam Vu. 2021. FaceSense: sensing face touch with an ear-worn system. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5, 3 (2021), 1–27.
[41] Sunil Kamath, Philipos Loizou, et al. 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise.. In *ICASSP*, Vol. 4. Citeseer, 44164–44164.
[42] Toshiki Kikuchi and Yuko Ozasa. 2018. Watch, listen once, and sync: Audio-visual synchronization with multi-modal regression CNN. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3036–3040.
[43] Woosuk Kwon, Sehoon Kim, Michael W Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. 2022. A fast post-training pruning framework for transformers. *Proceedings of the Neural Information Processing Systems (NeurIPS)* 35 (2022), 24101–24116.
[44] Peter Ladefoged and Keith Johnson. 2014. *A course in phonetics*. Cengage learning.
[45] Ying-Hui Lai and Wei-Zhong Zheng. 2019. Multi-objective learning based speech enhancement method to increase speech quality and intelligibility for hearing aid device users. *Biomedical Signal Processing and Control* 48 (2019), 35–45.
[46] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. 2019. SDR–half-baked or well done?. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 626–630.
[47] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. Eario: A low-power acoustic sensing earable for continuously tracking detailed facial movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 2 (2022), 1–24.
[48] Chen Liang, Haoming Jiang, Zheng Li, Xianfeng Tang, Bing Yin, and Tuo Zhao. 2023. HomoDistil: Homotopic Task-Agnostic Distillation of Pre-trained Transformers. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
[49] Qianru Liao, Yongzhi Huang, Yandao Huang, Yuheng Zhong, Huitong Jin, and Kaishun Wu. 2022. MagEar: eavesdropping via audio recovery using magnetic side channel. In *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 371–383.
[50] Logitech. 2023. *G733*. Retrieved April 6, 2023 from https://www.logitechg.com/en-us/products/gaming-audio/g733-rgb-wireless-headset.981-000863.html
[51] Philipos C Loizou. 2005. Speech enhancement based on perceptually motivated Bayesian estimators of the magnitude spectrum. *IEEE Transactions on Speech and Audio Processing* 13, 5 (2005), 857–869.
[52] Yi Luo, Zhuo Chen, and Takuya Yoshioka. 2020. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 46–50.
[53] Yi Luo and Nima Mesgarani. 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 696–700.
[54] Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 27, 8 (2019), 1256–1266.
[55] Dong Ma, Andrea Ferlini, and Cecilia Mascolo. 2021. OESense: employing occlusion effect for in-ear human sensing. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 175–187.
[56] Jiachen Mao, Huanrui Yang, Ang Li, Hai Li, and Yiran Chen. 2021. Tprune: Efficient transformer pruning for mobile devices. *ACM Transactions on Cyber-Physical Systems (TCPS)* 5, 3 (2021), 1–22.
[57] Brian B Monson, Eric J Hunter, Andrew J Lotto, and Brad H Story. 2014. The perceptual significance of high-frequency energy in the human voice. *Frontiers in psychology* 5 (2014), 587.
[58] Arun Narayanan and DeLiang Wang. 2013. Ideal ratio mask estimation using deep neural networks for robust speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7092–7096.
[59] Zhaoheng Ni and Michael I Mandel. 2020. Mask-dependent phase estimation for monaural speaker separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7269–7273.

[60] Harry Nyquist. 1928. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers* 47, 2 (1928), 617–644.

[61] National Institute on Deafness and Other Communication Disorders (NIDCD). 2023. Noise-Induced Hearing Loss. https://www.nidcd.nih.gov/health/noise-induced-hearing-loss Accessed: May 15, 2023.

[62] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, Min Wu, and KJ Ray Liu. 2023. Radio SES: mmWave-Based Audioradio Speech Enhancement and Separation System. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 31 (2023), 1333–1347.

[63] K Paliwal and Anjan Basu. 1987. A speech enhancement method based on Kalman filtering. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 12. IEEE, 177–180.

[64] Ashutosh Pandey and DeLiang Wang. 2019. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6875–6879.

[65] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. 2017. SEGAN: Speech Enhancement Generative Adversarial Network. *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)* (2017), 3642–3646.

[66] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems (NeurIPS)* 32 (2019).

[67] Nhat Pham, Tuan Dinh, Taeho Kim, Zohreh Raghebi, Nam Bui, Hoang Truong, Tuan Nguyen, Farnoush Banaei-Kashani, Ann Halbower, Thang N Dinh, et al. 2021. Detection of Microsleep Events with a Behind-the-ear Wearable System. *IEEE Transactions on Mobile Computing (TMC)* (2021).

[68] S. Pigeon. 2023. Babble Noise Background Noise Generator. https://mynoise.net/NoiseMachines/babbleNoiseGenerator.php. Accessed: 2023-06-29.

[69] Jay Prakash, Zhijian Yang, Yu-Lin Wei, Haitham Hassanieh, and Romit Roy Choudhury. 2020. EarSense: earphones as a teeth activity sensor. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 1–13.

[70] D. Purves, G.J. Augustine, D. Fitzpatrick, et al. 2001. *The Audible Spectrum*. Sinauer Associates.

[71] Akam Rahimi, Triantafyllos Afouras, and Andrew Zisserman. 2022. Reading To Listen at the Cocktail Party: Multi-Modal Speech Separation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10493–10502.

[72] ITUT Rec. 2003. P. 862.1: Mapping function for transforming P. 862 raw result scores to MOS-LQO. *International Telecommunication Union, Geneva* 24 (2003).

[73] ITU-T Recommendation. 2001. Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs. *Rec. ITU-T P. 862* (2001).

[74] Dario Rethage, Jordi Pons, and Xavier Serra. 2018. A wavenet for speech denoising. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5069–5073.

[75] Samsung. 2023. *Galaxy Buds2 Pro*. https://www.samsung.com/global/galaxy/galaxy-buds2/

[76] Philipp Schilk, Niccolò Polvani, Andrea Ronco, Milos Cernak, and Michele Magno. 2023. In-Ear-Voice: Towards Milli-Watt Audio Enhancement With Bone-Conduction Microphones for In-Ear Sensing Platforms. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation (IOTDI)*. 1–12.

[77] Florian Schmid, Khaled Koutini, and Gerhard Widmer. 2023. Efficient large-scale audio tagging via transformer-to-cnn knowledge distillation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[78] Claude E Shannon. 1949. Communication in the presence of noise. *Proceedings of the IRE* 37, 1 (1949), 10–21.

[79] Sony. 2023. *WH-1000XM4*. Retrieved April 6, 2023 from https://electronics.sony.com/audio/headphones/headband/p/wh1000xm4-b

[80] Tanmay Srivastava, Prerna Khanna, Shijia Pan, Phuc Nguyen, and Shubham Jain. 2022. MuteIt: Jaw Motion Based Unvoiced Command Recognition Using Earable. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 3 (2022), 1–26.

[81] Kenneth N Stevens. 2000. *Acoustic phonetics*. Vol. 30. MIT press.

[82] Thomas G Stockham Jr. 1966. High-speed convolution and correlation. In *Proceedings of the April 26-28, 1966, Spring joint computer conference*. 229–233.

[83] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. 2021. Attention is all you need in speech separation. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 21–25.

[84] Baochen Sun, Jiashi Feng, and Kate Saenko. 2016. Return of frustratingly easy domain adaptation. In *Proceedings of the AAAI conference on artificial intelligence (AAAI)*, Vol. 30.

[85] Ke Sun and Xinyu Zhang. 2021. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 160–173.

[86] Ke Sun, Ting Zhao, Wei Wang, and Lei Xie. 2018. Vskin: Sensing touch gestures on surfaces of mobile devices using acoustic signals. In *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 591–605.

[87] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

IEEE, 4214–4217.
[88] Ingo R Titze and Daniel W Martin. 1998. Principles of voice production.
[89] Emmanuel Vincent, Rémi Gribonval, and Cédric Févotte. 2006. Performance measurement in blind audio source separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 14, 4 (2006), 1462–1469.
[90] Phil Wang. 2020. vit-pytorch: An implementation of Vision Transformers in PyTorch. https://github.com/lucidrains/vit-pytorch.
[91] Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John Hershey, Rif A Saurous, Ron J Weiss, Ye Jia, and Ignacio Lopez Moreno. 2019. VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking. *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)* (2019).
[92] Tianben Wang, Daqing Zhang, Yuanqing Zheng, Tao Gu, Xingshe Zhou, and Bernadette Dorizzi. 2018. C-FMCW based contactless respiration detection using acoustic signal. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 1, 4 (2018), 1–20.
[93] Yuntao Wang, Jiexin Ding, Ishan Chatterjee, Farshid Salemi Parizi, Yuzhou Zhuang, Yukang Yan, Shwetak Patel, and Yuanchun Shi. 2022. FaceOri: Tracking Head Position and Orientation Using Ultrasonic Ranging on Earphones. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. 1–12.
[94] Zi Wang, Yili Ren, Yingying Chen, and Jie Yang. 2022. Toothsonic: Earable authentication via acoustic toothprint. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 2 (2022), 1–24.
[95] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An Ear Canal Deformation Based Continuous User Authentication Using In-Ear Wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5, 1 (2021), 1–27.
[96] Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. 2019. WHAM!: Extending Speech Separation to Noisy Environments. *Proceedings of the Conference of the International Speech Communication Association (INTERSPEECH)*.
[97] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. 2015. Complex ratio masking for monaural speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 24, 3 (2015), 483–492.
[98] Yi Wu, Vimal Kakaraparthi, Zhuohang Li, Tien Pham, Jian Liu, and Phuc Nguyen. 2021. BioFace-3D: continuous 3d facial reconstruction through lightweight single-ear biosensors. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 350–363.
[99] Yadong Xie, Fan Li, Yue Wu, Huijie Chen, Zhiyuan Zhao, and Yu Wang. 2022. TeethPass: Dental Occlusion-based User Authentication via In-ear Acoustic Sensing. In *Proceedings of the 2022-IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 1789–1798.
[100] Chao Xu, Junwei Zhu, Jiangning Zhang, Yue Han, Wenqing Chu, Ying Tai, Chengjie Wang, Zhifeng Xie, and Yong Liu. 2023. High-fidelity Generalized Emotional Talking Face Generation with Multi-modal Emotion Space Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6609–6619.
[101] Xuhai Xu, Haitian Shi, Xin Yi, WenJia Liu, Yukang Yan, Yuanchun Shi, Alex Mariakakis, Jennifer Mankoff, and Anind K Dey. 2020. Earbuddy: Enabling on-face interaction via wireless earbuds. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*. 1–14.
[102] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. 2020. Phasen: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. 9458–9465.
[103] Guochen Yu, Andong Li, Chengshi Zheng, Yinuo Guo, Yutian Wang, and Hui Wang. 2022. Dual-branch attention-in-attention transformer for single-channel speech enhancement. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7847–7851.
[104] Anthony Zhang. 2023. Speech Recognition (Version 3.10.0) [Software]. https://github.com/Uberi/speech_recognition.
[105] Qian Zhang, Dong Wang, Run Zhao, Yinggang Yu, and Junjie Shen. 2021. Sensing to hear: Speech enhancement for mobile devices using acoustic signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 5, 3 (2021), 1–30.
[106] Marvin C Ziskin. 1993. Fundamental physics of ultrasound and its propagation in tissue. *Radiographics* 13, 3 (1993), 705–709.