



EAR SPEECH: Exploring In-Ear Occlusion Effect on Earphones for Data-efficient Airborne Speech Enhancement

FEIYU HAN, Nanjing University of Information Science and Technology, China and University of Science and Technology of China, China

PANLONG YANG, Nanjing University of Information Science and Technology, China

YOU ZUO, University of Science and Technology of China, China

FEI SHANG, University of Science and Technology of China, China

FENGLEI XU, Suzhou University of Science and Technology, Jiangsu Industrial Intelligent and Low-carbon Technology Engineering Center, China

XIANG-YANG LI, University of Science and Technology of China, China

Earphones have become a popular voice input and interaction device. However, airborne speech is susceptible to ambient noise, making it necessary to improve the quality and intelligibility of speech on earphones in noisy conditions. As the dual-microphone structure (*i.e.*, outer and in-ear microphones) has been widely adopted in earphones (especially ANC earphones), we design EAR SPEECH which exploits in-ear acoustic sensory as the complementary modality to enable airborne speech enhancement. The key idea of EAR SPEECH is that in-ear speech is less sensitive to ambient noise and exhibits a correlation with airborne speech. However, due to the occlusion effect, in-ear speech has limited bandwidth, making it challenging to directly correlate with full-band airborne speech. Therefore, we exploit the occlusion effect to carry out theoretical modeling and quantitative analysis of this cross-channel correlation and study how to leverage such cross-channel correlation for speech enhancement. Specifically, we design a series of methodologies including data augmentation, deep learning-based fusion, and noise mixture scheme, to improve the generalization, effectiveness, and robustness of EAR SPEECH, respectively. Lastly, we conduct real-world experiments to evaluate the performance of our system. Specifically, EAR SPEECH achieves an average improvement ratio of 27.23% and 13.92% in terms of PESQ and STOI, respectively, and significantly improves SI-SDR by 8.91 dB. Benefiting from data augmentation, EAR SPEECH can achieve comparable performance with a small-scale dataset that is 40 times less than the original dataset. In addition, we validate the generalization of different users, speech content, and language types, respectively, as well as robustness in the real world via comprehensive experiments. The audio demo of EAR SPEECH is available on <https://github.com/EarSpeech/earspeech.github.io/>.

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

Additional Key Words and Phrases: Speech Enhancement, Earphone-based Sensing and Computing, In-ear Acoustic Sensing, Occlusion Effect.

Authors' addresses: Feiyu Han, Nanjing University of Information Science and Technology, Nanjing, China, 210044 and University of Science and Technology of China, Hefei, China, 230026, fyhan@mail.ustc.edu.cn; Panlong Yang, Nanjing University of Information Science and Technology, Nanjing, China, 210044, plyang@nuist.edu.cn; You Zuo, University of Science and Technology of China, Hefei, China, 230026, leftright@mail.ustc.edu.cn; Fei Shang, University of Science and Technology of China, Hefei, China, 230026, shf_1998@outlook.com; Fenglei Xu, Suzhou University of Science and Technology, Jiangsu Industrial Intelligent and Low-carbon Technology Engineering Center, Suzhou, China, 215009, xufl@mail.usts.edu.cn; Xiang-Yang Li, University of Science and Technology of China, Hefei, China, 230026, xiangyangli@ustc.edu.cn.

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.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

2474-9567/2024/9-ART104

<https://doi.org/10.1145/3678594>

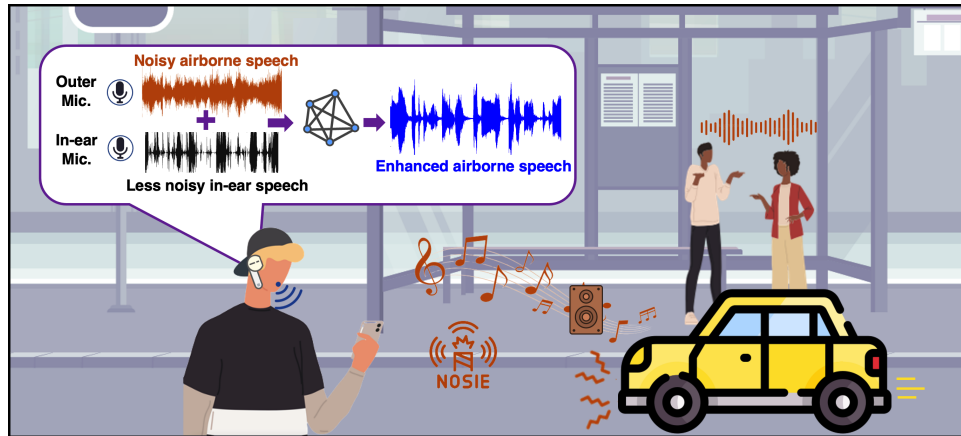


Fig. 1. EAR SPEECH is an earphone-based speech enhancement solution that utilizes outer and in-ear microphones on a single earphone to capture dual-channel speech signals and exploits the correlation between different acoustic channels for noise removal.

ACM Reference Format:

Feiyu Han, Panlong Yang, You Zuo, Fei Shang, Fenglei Xu, and Xiang-Yang Li. 2024. EAR SPEECH: Exploring In-Ear Occlusion Effect on Earphones for Data-efficient Airborne Speech Enhancement. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 8, 3, Article 104 (September 2024), 30 pages. <https://doi.org/10.1145/3678594>

1 INTRODUCTION

The increasing popularity of smart devices such as smartphones and tablets creates a high demand for earphones to provide interaction and entertainment for users. According to a market report [36], the market size of earphones and headphones is expected to reach 163 billion by 2030, showing a significant growth trend. However, a key challenge arises when the earphones are used as a voice input interface between users and smart devices (e.g., making a call with earphones). In particular, the quality of recorded speech drops dramatically in noisy environments, negatively impacting the user experience. Thus, in recent years, academic researchers have shown substantial interest in speech enhancement (SE) on earphones which aims to separate the target speech from external interference and noise.

Compared with the conventional denoising methods [1, 12, 24], audio-based deep learning technologies [3, 15, 45, 46] that rely on a large training speech corpus have shown a greater performance improvement in the speech enhancement task. However, the performance of such technologies is significantly affected by the size and quality of the training dataset, causing poor generalization in practice. Recently, multi-modality technologies performed speech enhancement through the correlation between audio modality and another modality, like video-audio [32], mmWave-audio [30, 41, 63], and ultrasound-audio [11, 55, 65], superior to audio-based enhancement technologies. Nevertheless, such methods require the integration of additional hardware (e.g., wireless antennas or cameras), which cannot be adapted to earphones with small sizes and restraint resources. Some technologies have started to explore speech enhancement for earphones. However, they mainly rely on special sensors like physics-contact bone conduction transducer [60] and high-sampling accelerometer [23], decreasing the adaptability to most earphones. Chatterjee *et al.* [8] leverage the cooperation of binaural earphones for spatial filtering, which cannot adapt to single-earphone usage scenarios (e.g., working with one earphone and charging the other).

To solve the above limitations, we design EARSPEECH that leverages outer and inner (*i.e.*, in-ear) microphones on the same earphone for speech enhancement. Nowadays, with in-ear microphones being widely embedded into earphones (especially ANC earphones), the in-ear acoustic modality boosts the capacity of sensing on earphones [13, 21, 22, 31], which could provide an additional sensory way to enable speech enhancement. As shown in Fig. 1, for the dual-microphone structure on a single earphone, the outer microphone can capture sound vibrations traveling over the air, while the in-ear microphone can capture sound vibrations propagating along the body and ear canal. To avoid ambiguity, we refer to sound vibrations captured by the outer microphone as *airborne speech* and sound vibrations captured by the in-ear microphone as *in-ear speech*. The goal of EARSPEECH is to exploit the correlation between different acoustic channels to improve the quality and intelligibility of airborne speech even in low SNR conditions. The key insight of EARSPEECH is that the airborne speech, which is sensitive to ambient noise, exhibits a unique and complex correlation with the in-ear speech, which is less susceptible to ambient noise. Furthermore, during developing EARSPEECH, we need to address the following key technical challenges:

- (i) *The impact of correlation between dual-channel signals with heterogeneous structures on the final speech enhancement task remains unclear.* Different from the common air-conducted channel, the in-ear channel consists of two pathways. During propagating along the in-ear channel, sound vibrations are first attenuated by bone conduction and then enhanced by the occlusion effect in the ear canal. The differences in propagating channels make airborne and in-ear speech have different acoustic characteristics, including spectral structure, noise sensitivity, and intelligibility. The correlation between heterogeneous dual-channel signals has not been well studied, including theoretical analysis and experimental validation, which prevents us from understanding the effect of such cross-channel correlation on speech enhancement.
- (ii) *A sufficient dataset of paired airborne speech and in-ear speech with labels is still lacking.* Existing public large-scale speech corpora are built on airborne speech (*i.e.*, speech over the air) like LibriSpeech [43], TIMIT [17], and Common Voice [2], but there are no corresponding public in-ear speech corpora. The quality and quantity of the dual-speech dataset are important in data analysis and model building and are closely related to the effectiveness and generalization of speech enhancement. An intuitive way is to manually collect airborne speech samples and corresponding in-ear speech samples in the lab environment to compose a large-scale dual-channel speech corpus. However, it requires extensive collection and labeling efforts, which is not feasible for large-scale dataset construction.
- (iii) *The heterogeneity in speech signals caused by diversities (e.g., different channels and speakers) makes it difficult to extract effective and generalized features to represent the correlation of dual-channel speech.* Considering the differences in spectral structures of the dual-channel speech, EARSPEECH needs to capture representative features that contribute to the speech enhancement task from each acoustic channel and effectively fuse them based on the correlation. In addition, the pronunciation habits and the structure of the body (such as vocal organs, skull, and ear canal) vary from person to person, which affects the consistency of the correlation between dual-channel speech. Furthermore, the noise sensitivity of the in-ear channel is also different from that of the airborne channel, affecting the robustness of our system in real-world environments. Thus, addressing the impact of these factors to improve the system's generalization and robustness continues to pose a challenge.

To address the above challenges, we utilize an electro-acoustic (EA) model to explain the impact of the occlusion effect on the in-ear channel and analyze the characteristics of in-ear speech. Then, we integrate a sound propagation model to theoretically analyze the cross-channel correlation based on the occlusion effect. Based on such cross-channel correlation, we design a data augmentation method to enrich the dual-channel speech corpus with an existing large-scale airborne speech corpus. Specifically, we design a spectral mapping function, named transfer function, to represent the correlation between different acoustic channels. Then, due

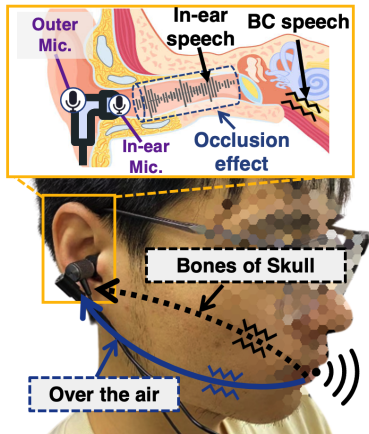


Fig. 2. Illustration of sound propagation in airborne and in-ear channels.

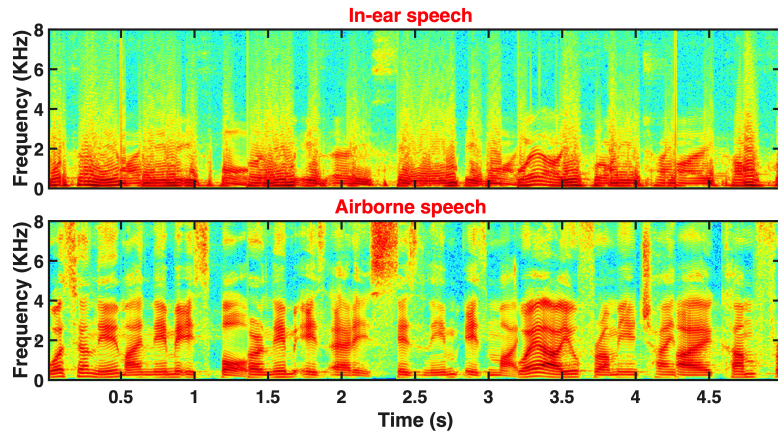


Fig. 3. Spectrograms of 5-second airborne speech signals (lower) and corresponding in-ear speech signals (upper).

to the diversity of individuals, we design a GMM-based transfer function estimation method to improve the generalization of data augmentation. Lastly, we design a **Dual-Channel Speech Enhancement (DC-SE)** model based on a two-branch deep learning network to effectively extract features of dual-channel speech and fuse them for speech enhancement. In addition, to improve the robustness of our system in real-world environments, we design a novel noise mixture scheme that simulates the impact of ambient noise on the in-ear channel, to build the training dataset.

Compared with other SE technologies on earphones, EARSPEECH has the following advantages. (1) *Wide availability.* EARSPEECH only leverages two onboard microphones on a single earphone without hardware modification, which allows our system to be applied to most daily scenarios including single-earphone usage. (2) *Data efficiency.* Benefiting from the data augmentation based on the occlusion effect, the DC-SE model can achieve excellent performance through fine-tuning with small-scale real speech samples (about 1/40 of the original dataset collected in the real world), greatly reducing collection costs and achieving high data efficiency. (3) *High Generalization and Robustness.* EARSPEECH has been validated to achieve excellent performance among different users, speech contents, and language types. In addition, with the different impact factors, EARSPEECH can significantly improve the quality and intelligibility of noisy speech in daily scenarios.

We highlight the contributions of our work as follows:

- We design an earphone-based airborne speech enhancement, named EARSPEECH, which takes advantage of the dual-microphone structure on a single earphone with less hardware modification. EARSPEECH exploits the correlation between in-ear speech and airborne speech to improve the quality and intelligibility of airborne speech.
- From theoretical analysis and experimental validation, we study the correlation between dual-channel speech signals based on the occlusion effect. Furthermore, based on the analyzed occlusion effect-based correlation, we design a GMM-based data augmentation method and deep learning-based speech enhancement model for effective, robust, data-efficient, and generalized speech enhancement.
- We conduct comprehensive experiments to evaluate our system in terms of effectiveness (improvement of speech quality and intelligibility), generalization (user groups, speech contents, and language types), robustness in the real world, and data efficiency.

2 BACKGROUND AND PRELIMINARY

2.1 Background: Speech Acquisition on Earphones

Nowadays, most earphones, especially noise-canceling earphones, are equipped with inward-facing microphones (refer to in-ear microphones) and outward-facing microphones (refer to outer microphones). As shown in the orange box of Fig. 2, the outer microphone is mounted on the outside of the earphone, and the in-ear microphone is mounted on the inside of the earphone. Fig. 2 illustrates sound propagation in airborne and in-ear channels. When a participant wears earphones to speak (e.g., making a call and interacting with a voice assistant), sound vibrations propagate over the air and are captured by the outer microphone. In addition, sound vibrations also propagate to the ear canal via the skull and can be captured by the in-ear microphone. In our work, we refer to the sound vibrations captured by the outer microphone as airborne speech and sound vibrations captured by the in-ear microphone as in-ear speech.

The in-ear channel consists of two components, *i.e.*, the skull and the ear canal. Firstly, sound vibrations generated from the sound source propagate along the skull. When these bone-conducted sound vibrations (also known as BC speech) reach the eardrum, they will enter the ear canal and continue to spread. If the ear canal opening is obstructed, these bone-conducted sound vibrations are trapped within the ear canal and bounce back and forth in the obstructed canal [14]. Such obstruction increases the acoustic impedance of the ear canal [6, 54], causing the low-frequency components (mainly below 1 KHz) of sound vibrations to be amplified. This phenomenon is known as the *occlusion effect* [6, 54].

We need to claim that in-ear speech is different from bone-conducted (BC) speech. Although the channels of in-ear speech and bone-conducted speech partially overlap, in-ear speech still experiences the occlusion effect, resulting in a distinct spectral structure that sets it apart from bone-conducted speech. In addition, the acquisition of BC speech requires a special bone conduction transducer that needs to be attached to the head or face. However, in-ear speech is sound vibrations in the ear canal and can be detected by a ubiquitous microphone.

2.2 Characteristics of Dual-channel Speech

We compare the in-ear channel with the airborne channel in terms of spectral structure, speech intelligibility, and ambient noise resistance, to help readers understand their characteristics.

2.2.1 Spectral Structure Difference. Fig. 3 illustrates the spectrogram comparison of in-ear speech (upper) and airborne speech (lower), showing significantly different spectral structures. Intuitively, airborne speech exhibits harmonic and formant structures in a wider frequency range (about 0-6 KHz). However, the most spectral power and harmonic structures of in-ear speech are concentrated below 2 KHz, while almost no obvious harmonic and formant structures can be found in the high-frequency components. To give a quantitative analysis, we calculate the cumulative distribution of spectral power (*i.e.*, $Power_{cdf}$), as shown in Fig. 4. Specifically, for in-ear speech, about 96% of overall spectral power is distributed in the frequency range below 1 kHz, while for airborne speech, only about 60% of overall spectral power is distributed in the frequency range below 1 kHz. During speaking-induced vibrations propagating along the head skull, the higher frequency components will suffer from more severe attenuation [53]. That is why high-frequency components of in-ear speech are different from airborne speech. In addition, the low-frequency components of in-ear speech are enhanced by the occlusion effect [6, 54], as introduced in Sec. 2.1. That is why most spectral power of in-ear speech is concentrated below 1 KHz.

2.2.2 Speech Intelligibility. Speech intelligibility refers to how comprehensible speech is in communication, which is related to high-frequency acoustic features of consonants and vowels [48], such as vowel space area, mean amount of formant movement, harmonic structure, and spectral centroid. Due to the occlusion effect, most of the spectral power of in-ear speech is mainly distributed below 1 KHz. The distortion in high-frequency

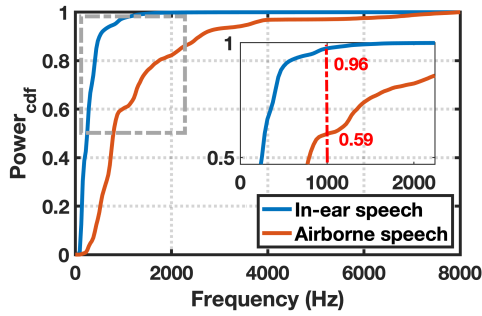


Fig. 4. Cumulative distribution of spectral power.

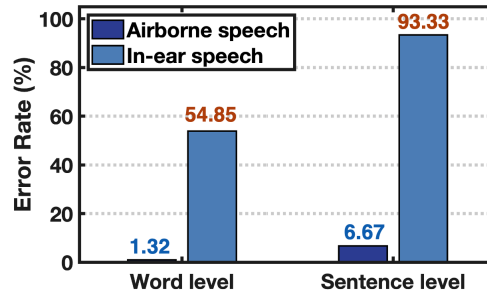


Fig. 5. Speech intelligibility study.

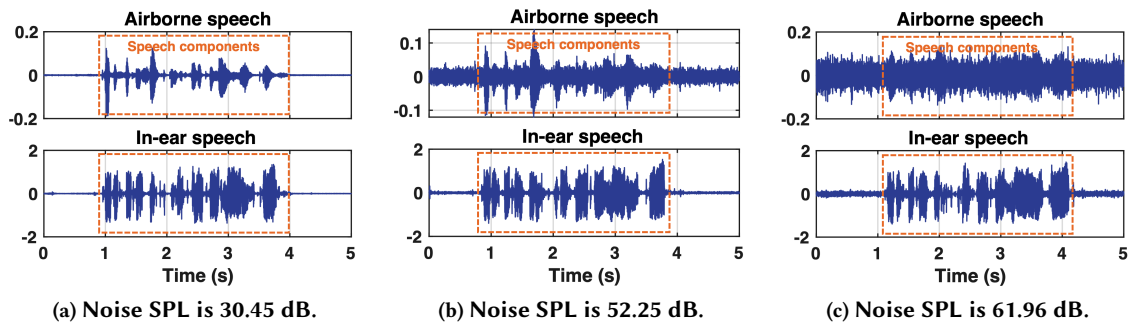


Fig. 6. Noise resistance study of in-ear speech and airborne speech.

components will affect the intelligibility of in-ear speech. Next, we use speech-to-text recognition accuracy as a measure metric to compare the intelligibility between airborne speech and in-ear speech. Specifically, we require five participants to naturally speak 26 letters (*i.e.*, A-Z) and 5 sentences at a normal sound pressure level (SPL) in a quiet room. Then we use *Notta* [38], a popular online transcription platform, to convert airborne speech and in-ear speech into text separately. As shown in Fig. 5, the WER (word error rate) and SER (sentence error rate) of in-ear speech are respectively 54.85% and 93.33%, while the WER and SER of airborne speech are only 1.32% and 6.67, respectively. It indicates that in-ear speech has poor intelligibility and is difficult to meet the requirements of daily speech communication.

2.2.3 Ambient Noise Impact. In this section, we study the impact of ambient noise on two speech channels. The participant is required to wear our prototype and speak several sentences in a quiet meeting room with little fan noise (30.45 dB on average), a student office with conversation and music noise (52.25 dB on average), and an outside street with vehicle noise (61.96 dB on average). Fig. 6 illustrates airborne and in-ear speech collected in different environments. As the noise sound pressure level increases, speech over the air (shown in the orange dotted box) is gradually buried by ambient noise. Especially, in the outside street with a noise SPL of 61.96 dB, we can barely see the airborne speech waveform, while the in-ear speech waveform is still obvious. These results indicate that ambient noise has a subtle impact on in-ear speech. The strong noise resistance of ambient noise comes from two aspects. First, ambient noise widely exists in the airborne channel and will not affect in-ear speech propagating along the in-body channel. Second, benefiting from the special in-ear structure design [35, 49], ear canals fit well with earphones, relatively isolating ear canals from external environments and preventing the entry of ambient noise.

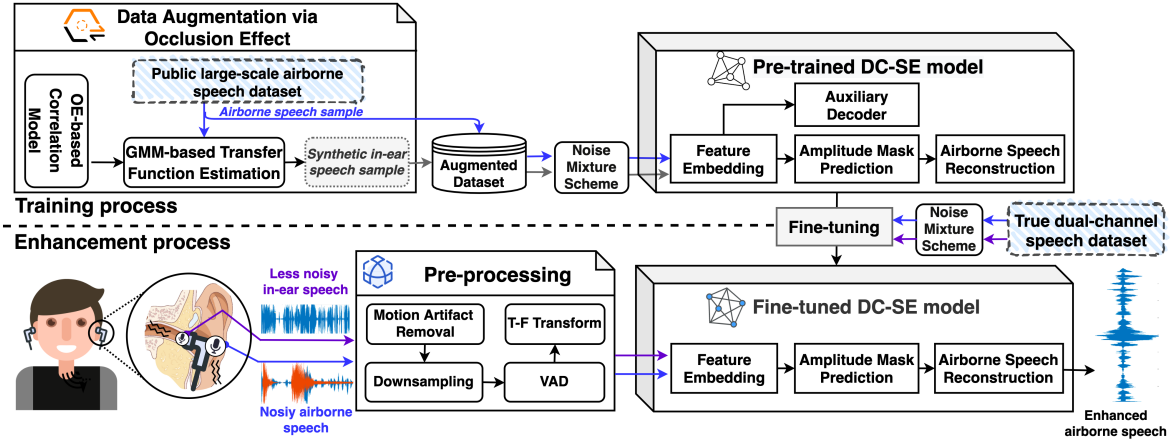


Fig. 7. The framework overview of EARSPEECH.

2.2.4 **Key Insight.** Through the above analysis, there are three observations as follows:

- Airborne and in-ear speech show distinct spectral structures due to differences in propagating channels. The low-frequency components of in-ear speech are amplified by the occlusion effect, while high-frequency components suffer from severe attenuation.
- Compared with the wide bandwidth (about 6-8 KHz) of airborne speech, the narrow bandwidth (about 1-2 KHz) of in-ear speech has relatively lower intelligibility due to the loss of high-frequency information.
- Benefiting from the in-body channel and microphone’s position in earphones, in-ear speech exhibits higher resistance to ambient noise than airborne speech.

Compared with airborne speech, in-ear speech has a stronger capability of noise resistance but weaker intelligibility. Importantly, most commercial devices only support voice input from outer microphones instead of in-ear microphones. Hence, EARSPEECH aims to leverage the in-ear speech as an auxiliary approach to improve the quality of airborne speech in low SNR conditions.

3 SYSTEM OVERVIEW

Fig.7 illustrates the overview framework of EARSPEECH, consisting of three key modules: signal pre-processing, data augmentation, and dual-channel speech enhancement model (also referred to as DC-SE model). In particular, we highlight two processes (*i.e.*, training process and enhancement process) in our system. In the training process, to improve the generalization and performance of the model, EARSPEECH first pre-trains the DC-SE model using a large-scale synthetic dual-channel dataset generated from data augmentation. Then, EARSPEECH utilizes a small-scale dual-channel speech dataset collected in real environments to fine-tune the pre-trained DC-SE model. In the enhancement process, a pair of noisy airborne speech and corresponding in-ear speech are pre-processed and then fed into the fine-tuned DC-SE model for airborne speech enhancement. Next, we introduce the main function of each module separately.

(i) **GMM-based Augmentation for In-ear Speech Corpus Enrichment (Section §4).** As for the problem of lacking paired training data, we design a GMM-based data augmentation method to generate/synthesize in-ear speech samples from existing public airborne speech corpus based on the cross-channel correlation. Building a large-scale synthetic in-ear speech corpus and corresponding airborne speech corpus can help our deep-learning model effectively improve the learning capability of complex relationships between in-ear and airborne speech.

(ii) **Pipeline of Pre-processing (Section §5.1).** Since low-frequency components of in-ear speech are easily influenced by motion artifacts [16, 22], EARSPEECH first removes unnecessary noise to obtain clean in-ear speech. Then, EARSPEECH downsamples speech signals of each channel to improve the computational efficiency and performs voice activity detection to remove silent frames. Lastly, dual-channel speech signals are transformed from the time domain (T domain) to the time-frequency domain (T-F domain) where the target speech and noise show a more distinct distribution difference.

(iii) **Dual-channel Speech Enhancement Model (Section §5.2).** After signal pre-processing, dual-channel amplitude spectrograms are fed into the DC-SE model for speech enhancement. The DC-SE model consists of a *Feature Embedding* sub-network, an *Amplitude Mask Prediction* sub-network, an *Auxiliary Decoder* sub-network, and an *Airborne Speech Reconstruction* sub-network. The *Feature Embedding* sub-network aims to transform two input amplitude spectrograms into the same feature space and extracts high-level feature maps to represent the cross-channel correlation. After that, EARSPEECH concatenates outputs of the *Feature Embedding* sub-network as a feature map along the channel dimension. The concatenated feature map passes through the *Amplitude Mask Prediction* sub-network to generate an ideal ratio mask (IRM) that represents the ratio between clean and noisy spectrograms. Enhanced airborne amplitude spectrogram is calculated by the element-wise multiplication between IRM and noisy airborne amplitude spectrogram. Lastly, *Airborne Speech Reconstruction* converts T-F domain representations into T domain representations for speech construction. It should be particularly emphasized that *Auxiliary Decoder* helps the DC-SE model to learn the cross-channel correlation only in the training process and doesn't participate in the speech enhancement process.

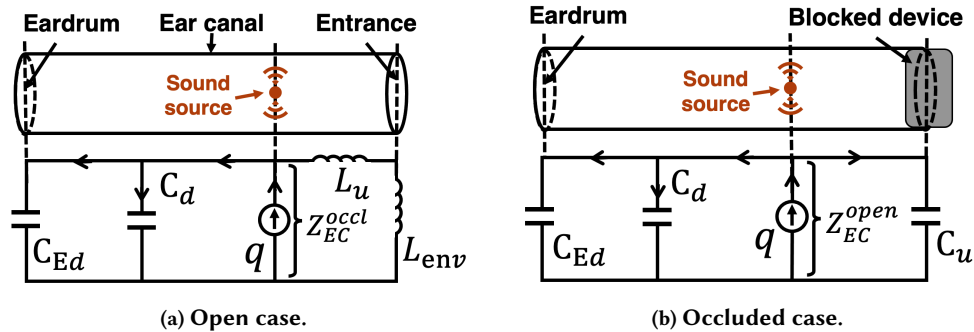


Fig. 8. Leveraging the EA model to analyze the correlation between dual-channel speech. An Electro-acoustic (EA) model of (a) the open ear canal and (b) the occluded ear canal.

4 DATA AUGMENTATION VIA OCCLUSION EFFECT

In this section, we introduce how to exploit the correlation between in-ear and airborne speech for data augmentation to tackle the challenge of lacking paired training data. We first theoretically analyze the cross-channel correlation based on channel differences induced by the occlusion effect (see Sec. 4.1). Then, we propose the transfer function to represent the occlusion effect-based correlation and explore the characteristics of the transfer function by experimental validation (see Sec. 4.2). Based on that, we design an efficient GMM-based augmentation method to synthesize paired in-ear speech samples with a large-scale airborne speech corpus (see Sec. 4.3) and validate the effectiveness of our designed augmentation method (see Sec. 4.4).

4.1 Transfer Function: OE-based Correlation Analysis

Occlusion effect (OE) significantly enhances low-frequency (mainly below 1 KHz) components of in-ear speech [6, 54]. The fundamental mechanism of occlusion effect is the acoustic impedance of the ear canal increases caused by blocking the ear canal entrance. In our work, we leverage an electro-acoustic (EA) model proposed in [6, 7] to help us understand the occlusion effect. Based on the EA model, we theoretically analyze the impact of the occlusion effect on in-ear speech and elaborate on the correlation between in-ear speech and airborne speech. Fig. 8 presents an EA model of the opening and occluded ear canal. In the EA model, the vibration of the ear canal wall induced by speaking is considered a sound source. The ear canal is divided into two spaces: downstream space (*i.e.*, sound source to eardrum) and upstream space (*i.e.*, sound source to ear canal entrance). As for opening ear canal case, the acoustic impedance of ear canal Z_{EC}^{open} can be expressed as:

$$\begin{aligned} Z_{EC}^{open} &= (j\omega C_{Ed} + j\omega C_d + (j\omega L_u + j\omega L_{env})^{-1})^{-1} \\ &= \frac{j\omega(L_u + L_{env})}{1 - \omega^2(C_{Ed} + C_d)(L_u + L_{env})} \end{aligned} \quad (1)$$

where C_{Ed} , C_d , L_u , and L_{env} denote the acoustic compliance of the eardrum, the acoustic compliance of the downstream space, the acoustic mass of the opening upstream space, and the acoustic mass of the external environment, respectively. As for the occluded ear canal case, the acoustic impedance of ear canal Z_{EC}^{occl} can be expressed as:

$$Z_{EC}^{occl} = (j\omega C_{Ed} + j\omega C_d + j\omega C_u)^{-1} \quad (2)$$

where C_u denotes the acoustic compliance of the occluded upstream space. Hence, the occlusion effect F_{OE} can be represented as the ratio of the acoustic pressure of occluded ear canal P_{occl} and the acoustic pressure of opening ear canal P_{open} :

$$\begin{aligned} F_{OE} &= P_{occl}/P_{open} = q * Z_{EC}^{occl} / q * Z_{EC}^{open} \\ &= \frac{\omega^2(C_{Ed} + C_d)(L_u + L_{env}) - 1}{\omega^2(C_{Ed} + C_d + C_u)(L_u + L_{env})} \end{aligned} \quad (3)$$

where q denotes the volume velocity at the ear canal. The aforementioned acoustic compliances and masses are determined by the geometry of the ear canal (*e.g.*, radius and length) and environmental factors [6]. Thus, Eq. 3 is simplified as follows:

$$F_{OE}(f) = \Gamma - \frac{1}{(2\pi f)^2 \chi} \quad (4)$$

where $\omega = 2\pi f$, $\Gamma = \frac{(C_{Ed} + C_d)}{(C_{Ed} + C_d + C_u)}$, and $\chi = (C_{Ed} + C_d + C_u)(L_u + L_{env})$. For the speech $s(f)$ produced by vocalization organs, it will propagate along two channels: air and bone. In our work, we define the *transfer function* as the ratio of in-ear speech $s_{ie}(f)$ to airborne $s_{air}(f)$ speech, which is expressed as follows:

$$F_{if} = \frac{s_{ie}(f)}{s_{air}(f)} = \frac{F_{OE}(f) * H_{bone}(f) * \cancel{s(f)}}{H_{air} * \cancel{s(f)}} \quad (5)$$

where H_{air} and H_{bone} are frequency-dependent attenuation function, respectively. In general, H_{air} is related to the airborne medium and can be considered as a constant. $H_{bone}(f)$ is related to the structure of the skull, which varies from person to person.

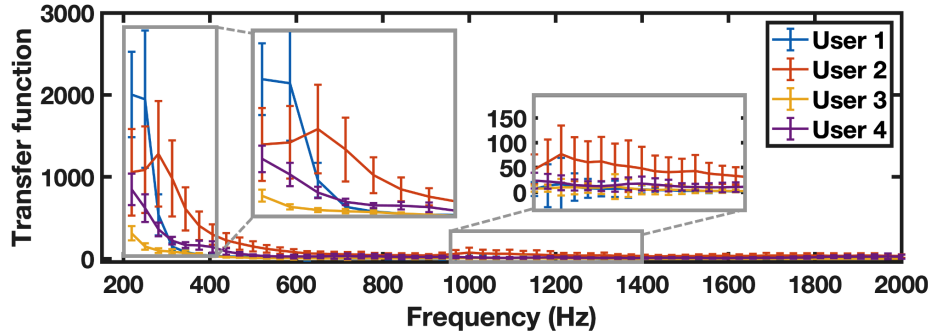


Fig. 9. Transfer function measurements in the frequency range below 2000 Hz.

4.2 Measurement of Transfer Function

Since $F_{OE}(f)$ and $H_{bone}(f)$ are associated with the individual as introduced in Sec.4.1, we use dual-channel speech collected from four participants (users) to explore the diversity of transfer function. Each user is required to wear earphones with dual microphones and speak for 2 minutes. Then we split the 2-minute speech into 24 5-second speech segments. For each speech segment, we calculate the spectrograms of in-ear speech S_{ie} and air-conducted speech S_{air} using the Short-Time Fourier Transform with a hamming window of 25 ms, an overlap length of 20 ms, and a FFT length of 512 points. To eliminate the impact of individual phonemes [19, 20], we accumulate each time point at the frequency bin f and calculate the transfer function $F_{tf}(f)$ as following:

$$F_{tf}(f) = \frac{\sum_t^T S_{ie}(t, f)}{\sum_t^T S_{air}(t, f)} \quad (6)$$

As introduced in the previous analysis, the occlusion effect only enhances the low-frequency components of speech. Thus, we mainly focus on the transfer function in the frequency range below 2000 Hz. The transfer function measurement of four users is shown in Fig. 9. Combining theoretical analysis and experimental validation, we can conclude the following characteristics of the transfer function. (i) The transfer function varies from person to person due to the unique body-physics structure of the individual. (ii) Despite the diversity of individuals' transfer functions, they still follow the same underlying variation pattern. As shown in Fig. 9, the transfer function shows an exponential decay pattern in the frequency range below 400 Hz. As the frequency increases, especially above 600 Hz, the transfer function becomes approximately stable. In addition, we can find that the transfer functions of different users exhibit significant differences in the frequency range between 200-600 Hz and only subtle differences in the frequency range above 800 Hz.

4.3 GMM-based Transfer Function Estimation

Intuitively, if we estimate the transfer function F_{tf} , we can synthesize the in-ear speech S_{ie} by calculating the product of F_{tf} and S_{air} . In other words, we can derive S_{air} from the public airborne speech dataset and generate the large-scale in-ear speech dataset with the help of the transfer function. However, as illustrated in Fig. 9, the diversity of transfer functions makes it difficult to directly estimate the transfer function by hard mapping.

In our work, we consider airborne speech as the *source domain* and in-ear speech as the *target domain*. To improve the generalization, a Gaussian mixture model is established to present the joint distribution of source and target domains, which can accurately represent the correlation between different domains via probability distribution. We denote the source and target feature vectors at frame t by X_t and Y_t , respectively. The joint

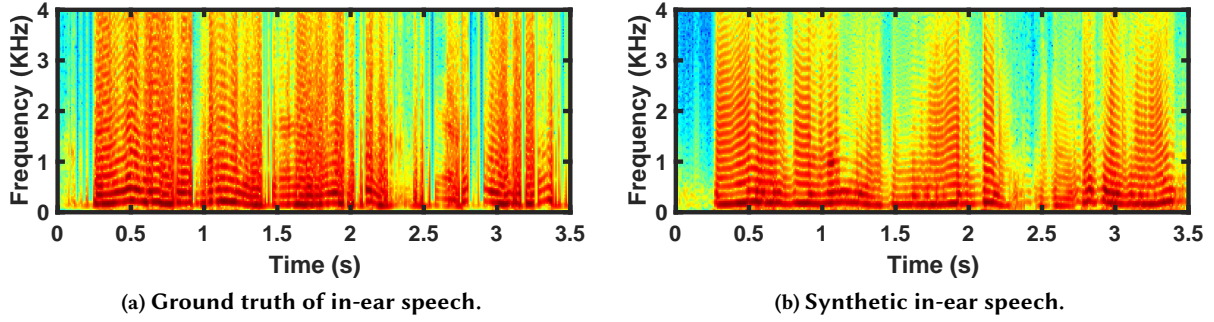


Fig. 10. Spectrogram comparison between (a) ground truth of in-ear speech and (b) synthetic in-ear speech.

distribution function is a combination of M normal distributions, which is expressed as follows:

$$P(X_t, Y_t | \lambda) = \sum_{m=1}^M \omega_m \mathcal{N}(X_t, Y_t; \Psi_m, \Sigma_m) \quad (7)$$

where λ denotes the parameters set of the joint distribution function. Ψ_m and Σ_m denote the mean vector and covariance matrix of the m -th component and are expressed as follows:

$$\Psi_m = \begin{bmatrix} \Psi_m^x \\ \Psi_m^y \end{bmatrix}, \Sigma_m = \begin{bmatrix} \Sigma_m^{xx} & \Sigma_m^{xy} \\ \Sigma_m^{yx} & \Sigma_m^{yy} \end{bmatrix} \quad (8)$$

We use the EM algorithm to train the joint distribution function and use DTW to align source and target feature vectors [57]. After training, we can calculate the predicted target feature vectors \hat{Y} by maximizing the following conditional probability function:

$$\hat{Y}_t = \underset{m}{\operatorname{argmax}} P(Y_t | X_t, \lambda) = \sum_{m=1}^M P(m | X_t, \lambda) P(Y_t | X_t, m, \lambda) \quad (9)$$

Thus, we can find that the key to the method is how to select representative features for the correlation modeling. As demonstrated in Sec. 4.1, there is a correlation between spectral structures of in-ear speech and airborne speech. Thus, we extract spectral features (*i.e.*, fundamental frequency and Mel spectrogram) and corresponding delta features from source and target domains for GMM modeling. To eliminate the over-smoothing effect, we leverage the global variance of the frame sequence as a supplement feature [57].

4.4 In-ear Speech Synthesis

Next, we synthesize in-ear speech from existing airborne speech datasets based on the estimated transfer function. After training, a GMM-based transfer function estimation is denoted by \hat{F}_{tf} . Given an airborne speech spectrogram S^{air} , the synthetic in-ear speech spectrogram can be expressed as:

$$\hat{S}_{ie} = \hat{F}_{tf}(S^{air}) \quad (10)$$

Then, \hat{S}_{ie} is converted to the time domain via inverse short-time Fourier transform (iSTFT). We use synthetic in-ear speech and ground-truth in-ear speech collected in the real world to validate the effectiveness of the designed data augmentation method preliminarily. First, we use paired airborne speech and in-ear speech of 6 participants to train the data augmentation model. Then, paired airborne speech and in-ear speech of the other 6 participants compose the validation dataset. We input each airborne speech sample in the validation dataset to

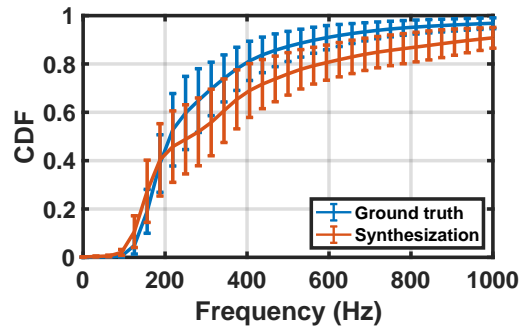


Fig. 11. Cumulative distribution of spectral power below 1000 Hz.

generate a synthetic in-ear speech sample. Lastly, we use all synthetic in-ear speech samples and corresponding ground-truth in-ear speech samples to conduct a preliminary validation of our data augmentation method.

Fig. 10 demonstrates an example comparison between ground-truth in-ear speech and synthetic in-ear speech. We can clearly observe that the key spectral structures (e.g., formant) in synthetic speech are similar to those in ground-truth speech. We want to obtain synthetic in-ear speech with a similar distribution to ground-truth speech, which can help the speech enhancement task analyze/learn the distribution difference between clean speech and noise. In our work, we calculate the cumulative distribution of spectral power to represent the distribution difference in spectral structure. As shown in Fig. 11, we can observe that variation patterns of two cumulative distributions are approximately similar, indicating the effectiveness of our designed method for data augmentation. Furthermore, in Sec. 6.4, we will evaluate the contribution of data augmentation to the final speech enhancement task in detail.

5 AIRBORNE SPEECH ENHANCEMENT UTILIZING DUAL-CHANNEL SPEECH

Intuitively, the speech enhancement task exploits the distribution difference between the desired speaker's speech and noise. Compared with the time domain, the distribution difference in the T-F domain is more discriminable. In addition, as a special complementary modality, in-ear speech is also distributed by motion and heartbeat artifacts. Therefore, it is necessary to pre-process collected raw dual-channel speech before enhancing target speech.

5.1 Data Pre-processing

Prior works [16, 22, 62] have reported that the frequency responses of motion-induced and heartbeat-induced in-ear sounds are mainly distributed below 100 Hz. Thus, we first use a high-pass filter with 100 Hz to remove motion and heartbeat artifacts. Since the frequency components above 8 KHz have little contribution to speech quality and intelligibility, we resample the dual-channel speech from 44.1 KHz to 16 KHz to improve the computational efficiency. Furthermore, we adopt a well-known voice activity detection (VAD) method [18] that extracts short-term energy and spectral spread for voice frame recognition and silent frame removal. Lastly, we transform dual-channel speech from the T domain to the T-F domain via short-time Fourier transform (STFT). The hamming window length, window step length, and FFT length in STFT are set to 400, 100, and 512, respectively.

5.2 DC-SE Model Design

After signal pre-processing, we can obtain the amplitude spectrogram and the phase spectrogram, respectively. Compared with the phase spectrogram, the amplitude spectrogram of speech has a more obvious structure pattern which can help distinguish noise. In our work, we design a deep-learning speech enhancement model to

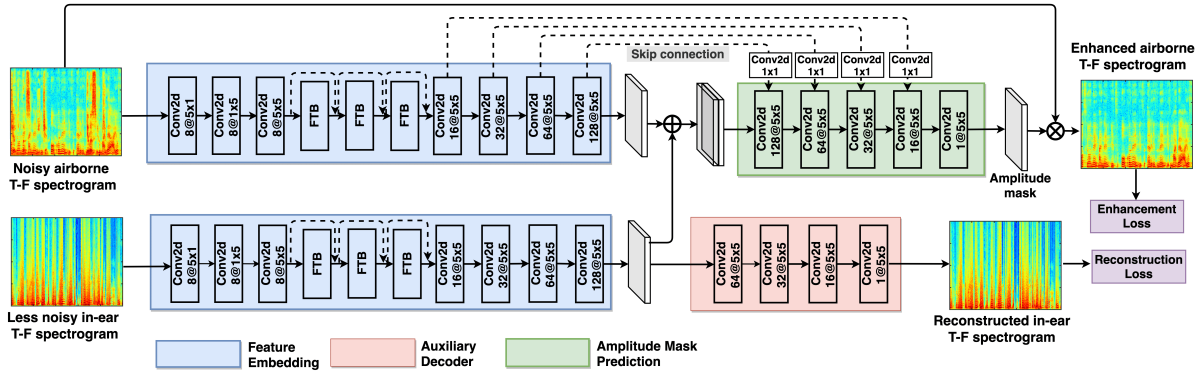


Fig. 12. Detailed structure of the dual-channel speech enhancement network. It is noted that the auxiliary decoder plays an important role in forcing the DC-SE model to learn multi-channel correlation information during the training process.

learn correlation information between amplitude spectrograms of dual-channel speech signals for separating target airborne speech from noise. The detailed network structure is shown in Fig. 12. The total parameters of the model are about 3.8 MB¹.

5.2.1 Feature Embedding. Although the sensory data of both channels are of the same modality, each channel of speech has its own unique temporal-spectral structure. It is difficult to directly extract features to represent the correlation between these dual-channel spectrograms. Therefore, we design a dual-branch *Feature Embedding* network that transforms dual channels of speech into the same feature space and extracts high-level representations to represent the cross-channel correlation.

We denote amplitude spectrograms of airborne speech and in-ear speech as $S_{air}^N \in \mathbb{R}^{T \times F \times C_{air}}$ and $S_{ie}^N \in \mathbb{R}^{T \times F \times C_{ie}}$, respectively. Here, F and T denote the frequency and time bins, respectively. $C_{air} = 1$ and $C_{ie} = 1$ are the number of channels of airborne amplitude spectrogram and in-ear amplitude spectrogram, respectively. As shown in the blue box of Fig. 12, the airborne speech feature embedding sub-network (upper row) and the in-ear speech feature embedding sub-network (lower row) have the same structure that can effectively capture corresponding correlation features. At the beginning of each sub-network, EARSPEECH utilizes three 2-D convolution layers with kernel sizes of 1×5 , 5×1 , and 5×5 to extract time-related correlation, frequency-related correlation, and global time-frequency correlation from input amplitude spectrogram. Each 2-D convolution layer is followed by a batch normalization layer and a ReLU function. In addition, we also borrow the FTB block from Phasen [64] to capture the detailed harmonic structure in the speech spectrogram. Three FTB blocks are connected by shortcut connections [29] to accelerate convergence. With the inspiration of Redundant Convolutional Encoder-Decoder (R-CED) network [45], EARSPEECH encodes output feature maps of the last FTB block into high dimension using four repetitions of a 2-D convolution layer with the kernel size of 5×5 , a batch normalization layer, and a ReLU activation layer. No pooling layer is present during the process of feature encoding, making a fully convolutional network more efficient. The FTB block does not change the input spectrogram’s size by the reshaping operation. Zero padding is applied in all convolution layers to keep the size of the output feature map consistent with that of the input feature map. Thus, the outputs of two feature embedding sub-networks are denoted by $U_{air} \in \mathbb{R}^{T \times F \times C_{air}}$ and $U_{ie} \in \mathbb{R}^{T \times F \times C_{ie}}$, respectively, where $C_{air} = 128$ and $C_{ie} = 128$. Lastly, EARSPEECH concatenates the outputs of

¹The source code of DC-SE model is available on <https://github.com/EarSpeech/earspeech.github.io/tree/main/model>

two feature embedding sub-networks as a feature map $U_{air}^{ie} = \text{cat}(U_{air}, U_{ie})$ along the channel dimension, where $U_{air}^{ie} \in \mathbb{R}^{T \times F \times (C_{ie} + C_{air})}$.

5.2.2 Amplitude Mask Prediction. The amplitude ideal ratio mask represents the ratio of speech power to noise power at each unit in the amplitude spectrogram. EARSPEECH exploits the concatenated feature map U_{air}^{ie} to predict the amplitude mask $M_{air} \in \mathbb{R}^{T \times F \times 1}$. As introduced in Sec. 5.2.1, EARSPEECH encodes feature maps into the high dimension by leveraging repetitions of convolution layers. Thus, in *Amplitude Mask Prediction* network, the concatenated high-dimensional feature map is compressed into the low-dimension amplitude mask. As shown in the green box of Fig. 12, *Amplitude Mask Prediction* can be considered as a decoder that has a symmetric structure with *Feature Embedding* network. In addition, we connect the last four convolution layers of *Feature Embedding* network and the first four convolution layers (i.e., decoder) of *Amplitude Mask Prediction* network by element-wise skip connections with 1×1 2-D convolution operations. Lastly, we use a 2-D convolution layer with a Sigmoid activation layer to map the output feature map of the decoder into the amplitude ideal ratio mask M_{air} .

5.2.3 Airborne Speech Reconstruction. The enhanced amplitude spectrogram is the element-wise multiplication of the input noisy airborne amplitude spectrogram and the predicted amplitude mask, which can be expressed as $S_{air}^{EN} = S_{air}^N \odot M_{air}$. To reconstruct airborne speech with high quality and intelligibility, we first combine the enhanced airborne amplitude spectrogram and the noisy airborne phase spectrogram $P_{air}^N \in \mathbb{R}^{T \times F \times 1}$, as $Y_{EN} = S_{air}^{EN} e^{(-jP_{air}^N)}$. Although prior works [42, 64] have reported that the phase spectrogram can help improve the quality of reconstructed speech, accurate phase spectrogram estimation brings additional computational overhead. Thus, by trading off the computational overhead and speech quality, we directly leverage the noisy airborne phase spectrogram for speech reconstruction. Lastly, EARSPEECH leverages the inverse short-time Fourier transform (iSTFT) to convert the enhanced speech from the T-F domain to the T domain.

5.2.4 Auxiliary Decoder. Since the enhanced airborne speech spectrogram is highly correlated with the noisy airborne speech spectrogram, the DC-SE model easily ignores the information of the in-ear branch. Thus, we additionally design an *Auxiliary Decoder* network to force the DC-SE model to learn multi-channel correlation during the training process. The DC-SE model transforms from a single-task (speech enhancement) learning case to a multi-task learning (speech enhancement and in-ear speech reconstruction) case. It is noted that *Auxiliary Decoder* network only participates in the training process and not in the enhancement process. The detailed structure of *Auxiliary Decoder* network is shown in the red box of Fig. 12, consisting of repetitions of convolution layers. The goal of the *Auxiliary Decoder* network is to reconstruct the in-ear speech spectrogram $S_{ie}^{rec} \in \mathbb{R}^{T \times F \times 1}$ using the high-dimensional feature map $U_{ie} \in \mathbb{R}^{T \times F \times C_{ie}}$ encoded by the in-ear feature embedding sub-network.

5.3 Training Methodology

5.3.1 Noise Mixture Scheme. We adopt a scheme that mixes various types of noise with clean speech to synthesize noisy speech. As introduced in Sec. 2.2.3, ambient noise has a subtle impact on the in-ear sound channel. Thus, one of the most intuitive ways is to artificially mix various types of noise with clean airborne speech and ignore the subtle impact of noise on in-ear speech, like previous solutions [3, 23, 46, 55, 64, 65]. However, this noise mixture way may decrease the robustness of our system in the real world, since ambient noise indeed affects the in-ear channel. Thus, in our work, we design an effective noise mixture scheme that simultaneously adds noise to airborne and in-ear channels. We observe that ambient noise needs to pass through earphones before entering the ear canal. In other words, if we could measure the attenuation of ambient noise by earphones, we can simulate the impact of ambient noise on the in-ear channel. Based on that, we perform a field test to calculate the SPL of ambient noise inside the ear canal (denoted by P_{ie}^n) and that of ambient noise outside the ear canal (denoted by P_{air}^n). Specifically, we require three participants wearing our prototype to keep silent and collect noise including music noise, conversation noise, ambient noise, and vehicle noise. In total, we collect

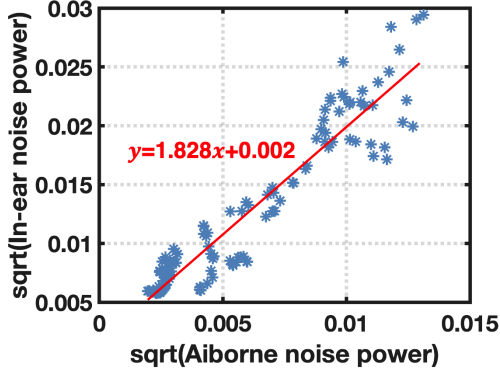


Fig. 13. The mapping function between $\sqrt{P_{air}^n}$ and $\sqrt{P_{ie}^n}$ shows a linear relationship which can be described by a linear fitting function (red line).

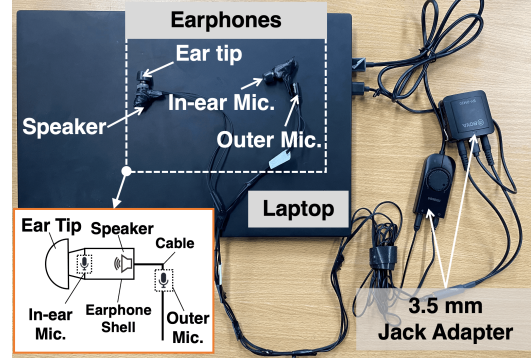


Fig. 14. Proof-of-concept earphone prototype and recording device (a Lenovo ThinkPad laptop).

125 5-second noise recordings in the air channel and corresponding noise recordings in the in-ear channel. We surprisingly find that $\sqrt{P_{air}^n}$ and $\sqrt{P_{ie}^n}$ are linearly related, i.e., $\sqrt{P_{ie}^n} = k * \sqrt{P_{air}^n} + b$, where $k = 1.828$ and $b = 0.002$, as shown in Fig. 13.

Given ambient noise signals n_{air} and clean airborne speech signals s_{air}^c , we generate noisy airborne speech signals s_{air}^n with the SNR of Ω by:

$$s_{air}^n = s_{air}^c + \beta * n_{air}, \quad \text{where} \quad \beta = \sqrt{\frac{\sum (s_{air}^c)^2}{10^{\frac{\Omega}{10}} \sum (n_{air})^2}} \quad (11)$$

where β denotes the scaling factor of noise signals. When we determine the β , the noise power in the in-ear channel can be calculated by:

$$P_{ie}^n = (k * \sqrt{\sum (\beta * n_{air})^2} + b)^2 \quad (12)$$

Thus, the synthetic noisy in-ear speech s_{ie}^n can be expressed as follows:

$$s_{ie}^n = s_{ie}^c + \beta_1 * n_{air} \quad \text{where} \quad \beta_1 = \sqrt{\frac{P_{ie}^n}{\sum (n_{air})^2}} \quad (13)$$

where s_{ie}^c and β_1 are clean in-ear speech signals and the scaling factor of ambient noise signals.

5.3.2 Loss Function Design. The *Auxiliary Decoder* network and the *Amplitude Mask Prediction* network perform different tasks and share the in-ear feature embedding subnetwork. Hence, we define two loss functions to improve learning efficiency and effectiveness. The first loss function is the enhancement loss, i.e., $\mathcal{L}_{en}(S_{air}^{EN}, S_{air}^C) = |S_{air}^{EN} - S_{air}^C|^2$, that measures the element-wise mean squared error (MSE) between the predicted airborne amplitude spectrogram and the clean airborne amplitude spectrogram. The minimization of \mathcal{L}_{en} helps the DC-SE model to learn the feature correlation between dual-channel speech and the distribution difference between the target speech and noise. Similarly, another loss function is the reconstruction loss, i.e., $\mathcal{L}_{rec}(S_{ie}^{rec}, S_{ie}^C) = |S_{ie}^{rec} - S_{ie}^C|^2$ that represents the element-wise mean squared error between the reconstructed in-ear amplitude spectrogram and the clean in-ear amplitude spectrogram. The minimization of \mathcal{L}_{rec} ensures that the information of the in-ear channel is not ignored. In practice, the reconstruction loss \mathcal{L}_{rec} is about 100-120 times larger than the enhancement loss

\mathcal{L}_{en} in the first few training epochs. The imbalance between the loss functions of the two tasks may make one task overfitting and another task underfitting [25, 58]. Thus, we formulate the final joint loss function by applying a logarithmic operation on each loss function:

$$\mathcal{L}_{joint} = \log_{10}\mathcal{L}_{en} + \log_{10}\mathcal{L}_{rec} \quad (14)$$

6 EVALUATION ANALYSIS

6.1 Experimental Setup

6.1.1 Proof-of-concept Prototype. Since most earphone manufacturers generally do not provide open API access to the audio stream, we design a proof-of-concept hardware prototype to collect dual-channel audio streams in our work. As shown in Fig. 14, we embed two AS-B6027AL30-RC electret microphones [39] on each earphone. The outer microphone is mounted on the outside of the earphone to collect airborne speech. The in-ear microphone is embedded into the earphone and faces towards the ear canal to collect in-ear speech. The specifications of both microphones include a 44.1 KHz sampling rate, 2200 Ω acoustic impedance, and -30 ± 2 dB acoustic sensitivity. Earphones are connected to a Lenovo ThinkPad laptop with 3.5 mm jack adapters.

6.1.2 Dataset for Evaluation. (i) *Dual-channel Speech Collection in Real World.* We recruit 24 participants, including 8 females and 16 males, with an age range of [22, 28]. Each participant is required to wear the earphone prototype in the way that is most comfortable for them. In addition, each participant's ear canals are completely occluded by the ear tips of earphones. The reading material consists of 15 daily English conversations selected from an online website [34]. Each participant says the reading material naturally at the sound pressure level of a normal conversation and repeats it 5-10 times. Then, we split collected parallel dual-channel (airborne and in-ear) speech samples into multiple 5-second segments for evaluation. In total, we can get 2450 pairs of airborne and in-ear sound segments.

(ii) *Noise Dataset.* We select three types of noise (*i.e.*, environmental noise denoted by n_{env} , speech noise denoted by n_{sph}^n , and music noise denoted by n_{msc}) to mix with clean airborne and in-ear speech samples. The ambient noise all comes from the ESC-50 dataset [47] that contains 2000 environmental audio recordings. We select more than 2900 audio recordings of 40 subjects from the LibriSpeech [43] to be speech noise. Lastly, we randomly select 25 songs with different music categories from MUSAN [52] to be music noise. We believe that such a wide range of noise categories can cover most daily scenarios.

(iii) *Training/Testing Dataset Generation.* We adopt the noise mixture scheme introduced in Sec. 5.3.1 for training and testing dataset generation. A training/testing instance can be denoted by $\{s_{air}^c, s_{ie}^c, s_n\}$, where s_{air}^c and s_{ie}^c denote the clean airborne, clean in-ear, and noise samples, respectively. For each pair of $\{s_{air}^c, s_{ie}^c\}$, we utilize 5 speech noise samples, 10 ambient noise samples, and 5 music noise samples to synthesize noisy airborne and in-ear speech samples, respectively. The SNR of synthetic noisy airborne samples is distributed randomly in a range of $[-5, 15]$ dB (4.37 on average) that can cover most daily real-world scenarios.

(iv) *Data Augmentation Using Public Dataset.* We use collected dual-channel speech samples of 6 participants to train the data augmentation model. Then, based on the trained augmentation model, we utilize a public airborne speech corpus LibriSpeech for data augmentation. We select airborne speech samples of 40 subjects in the LibriSpeech to generate in-ear samples. The generated/synthetic in-ear samples and corresponding real airborne samples are mixed with noise to pre-train the DC-SE model.

6.1.3 DC-SE Model Implement and Training/Testing. We implement the DC-SE model with PyTorch [37]. In the training process, the noisy airborne speech and corresponding in-ear speech are fed into the DC-SE model. The clean airborne speech and in-ear speech are the ground truth samples that are used for the loss calculation. Based on the joint loss function designed in Sec. 5.3.2, an Adam optimizer [26] is applied for weights updating

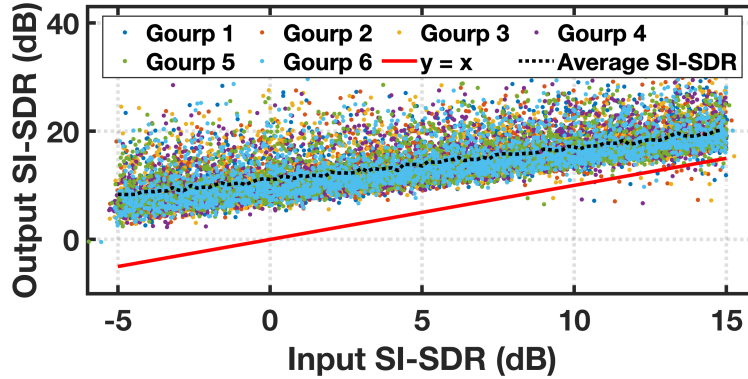


Fig. 15. Input-output SI-SDR of each testing user group (involving 3 participants). The results of each user group are calculated by the DC-SE model that is trained based on the remaining 5 user groups (involving 15 participants), *i.e.* leave-one-group-out cross-validation.

during training. The training process is completed on a server that is equipped with 40 Intel(R) Xeon(R) Silver 4210R@2.4GHz CPUs and 4 NVIDIA GeForce RTX 3090 graphics cards.

We first use a pre-training dataset generated from the LibriSpeech corpus to pre-train the DC-SE model with a maximum epoch of 30 and a batch size of 16. Then, we use the dual-channel speech dataset that is collected in the real world, to fine-tune the pre-trained DC-SE model. We split the remaining 18 participants into 6 groups. We utilize the leave-one-group-out cross-validation (*i.e.*, utilizing 5 groups for training and leaving the remaining one group for testing at a time) for the performance evaluation of EARSPEECH.

6.1.4 Evaluation Metrics. In our work, we adopt three speech quality metrics [45, 64] as follows:

- *PESQ*. Perceptual Evaluation of Speech Quality is a speech quality metric that ranges from -0.5 to 4.5, with higher PESQ indicating better voice quality.
- *STOI*. Short-time Objective Intelligibility measure is a speech intelligibility metric that ranges from 0 to 1, with higher STOI indicating better intelligibility.
- *SI-SDR*. Signal-to-distortion Ratio measures the distortion between the enhanced speech and clean speech, with higher SI-SDR indicating that the enhanced speech is close to the clean speech.

6.2 Overall Performance

As introduced in Sec. 6.1.3, we evaluate the overall performance of EARSPEECH via the leave-one-group-out cross-validation approach that can better estimate the effectiveness and generalization of our system on users who are not involved in the training process. Fig. 15 shows the input-output SI-SDR comparison of each testing user group. The red line represents the output SI-SDR of noisy speech is equal to the input SI-SDR, *i.e.*, no improvement in speech quality and intelligibility. We can clearly observe that input-output SI-SDR distributions of different testing user groups are similar, which indicates EARSPEECH can yield a high generalization performance among different users. In addition, we also find that our system can significantly improve the SI-SDR of noisy speech, especially in poor SI-SDR conditions. For example, for noisy speech with SI-SDR ranging from -5 dB to 0 dB, EARSPEECH can achieve 12.11 dB SI-SDR improvement on average.

We also select *Phasen* which has been considered a state-of-the-art airborne speech enhancement solution [11, 32, 55, 65], as our baseline to evaluate the overall performance of EARSPEECH. Tab. 1 illustrates the performance comparison between EARSPEECH and *Phasen* in different noise conditions, *i.e.*, environmental noise (EN), music

Table 1. Overall performance in different noise types with cross-validation. EN, MN, and SN represent environmental noise, music noise, and speech noise, respectively.

Method	PESQ				STOI				SI-SDR (dB)			
	EN	MN	SN	Avg	EN	MN	SN	Avg	EN	MN	SN	Avg
Noisy speech	2.65	2.25	2.29	2.46	0.84	0.75	0.74	0.79	5.08	5.05	5.09	5.07
Phasen	3.24	3.00	2.91	3.05	0.86	0.85	0.80	0.84	11.05	9.93	9.36	10.11
EARSPeech	3.25	3.06	2.97	3.13	0.91	0.89	0.88	0.90	15.16	12.89	12.73	13.98
-w/o FTB	3.08	2.78	2.70	2.91	0.88	0.85	0.84	0.87	13.80	11.66	10.96	12.55
-w/o SK	3.11	2.87	2.79	2.97	0.89	0.87	0.86	0.88	13.87	11.80	11.31	12.70
-w/o AD	3.16	2.93	2.78	3.01	0.90	0.88	0.86	0.88	14.09	12.16	11.52	12.96
-w/o IC	2.32	2.23	1.98	2.21	0.76	0.76	0.68	0.74	5.93	5.88	3.15	5.24

noise (MN), and speech noise (SN). We observe that EARSPeech can significantly improve the quality and intelligibility of noisy speech with an average PESQ improvement of 0.67 (improvement ratio of 27.23%), an average STOI improvement of 0.11 (improvement ratio of 13.92%), and an average SI-SDR improvement of 8.91 dB. In addition, compared with music noise and speech noise, EARSPeech can more effectively separate the environmental noise components from target speech. This is because the distribution pattern of environmental noise is more distinguishable from the distribution pattern of speech. Furthermore, EARSPeech significantly outperforms *Phasen* in average PESQ, average STOI, and average SI-SDR by 0.08, 0.06, and 3.87 dB, respectively. *Phasen* adopts a BiLSTM and three fully connected layers to learn context information and predict the amplitude mask, while EARSPeech takes advantage of convolution layers and skip connection instead. Therefore, the total parameters of EARSPeech (about 3.8 MB) are 2 times smaller than *Phasen* (about 7.8 MB) and still yield a performance improvement, showing a high applicability of EARSPeech on most commercial earphones.

6.3 Ablation Study

We conduct the ablation study to validate the contributions of key components in our speech enhancement model (*i.e.*, DC-SE model). The results of the ablation study are shown in Table 1.

(1) "w/o FTB" represents the "feature embedding" model without the FTB block that captures the detailed harmonic structure of speech. Since harmonic information can help the DC-SE model recover clean speech with minimal distortion, the quality and intelligibility metrics are decreased due to the loss of the FTB block.

(2) "w/o SK" represents the model without the skip connection block. We can find that the performance of EARSPeech is decreased without the assistance of the skip connection block. That is because the skip connection block can connect decoder and encoder blocks, avoiding the loss of context information.

(3) "w/o AD" represents the model without the "auxiliary decoder" block. The lack of the "auxiliary decoder" block makes the DC-SE model unable to fully exploit the correlation between airborne and in-ear channels, leading to a slight decrease in metrics.

(4) "w/o IC" represents the model without the in-ear channel branch (*i.e.*, without "in-ear feature embedding" and "auxiliary decoder" networks). The quality and intelligibility metrics exhibit a significant drop. Compared with noisy speech, the DC-SE model w/o IC only improves the average SI-SDR from 5.07 dB to 5.24 dB. In addition, the DC-SE model w/o IC even has a negative effect on the improvement of PESQ and STOI. This is because the lack of assistance from the in-ear channel branch leads to overfitting of the model during the training process, resulting in poor performance on unknown users in the testing dataset.

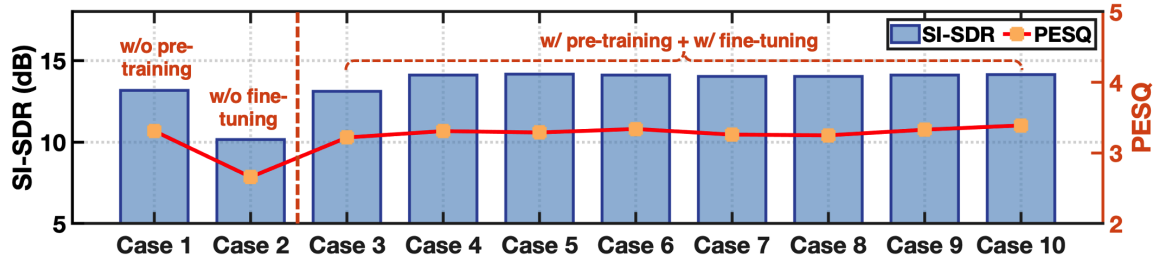


Fig. 16. Data efficiency validation. {Case 1}: only utilizing the real dual-channel speech dataset for training. {Case 2}: only utilizing the augmented dataset for training. {Case 3-9}: utilizing the augmented dataset for pre-training and then utilizing 50, 100, 150, 200, 400, 600, and 800 pairs of samples randomly selected from the real dataset for fine-tuning, respectively. {Case 10}: utilizing the augmented dataset for pre-training and utilizing all samples in the real dataset for fine-tuning.

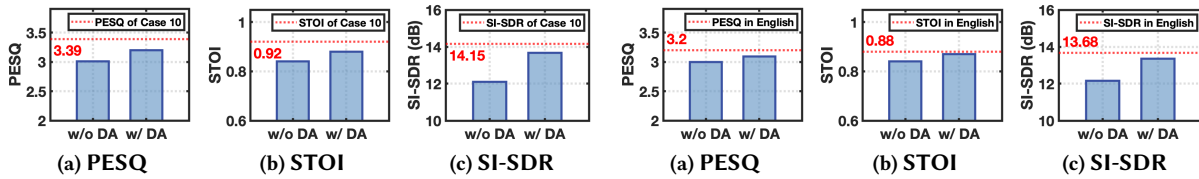


Fig. 17. Generalization validation on new sentences.

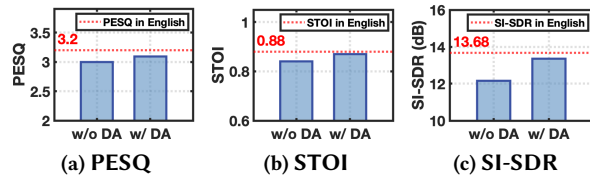


Fig. 18. Generalization validation on Mandarin.

6.4 Data Augmentation Effectiveness and Generalization Capability

The GMM-based data augmentation method designed in Sec. 4 can expand the scale of the training dataset, resulting in the improvement of generalization and sensing capacity. Although we have conducted a preliminary experiment to study the performance of data augmentation in Sec. 4.4, we quantify the contribution of the data augmentation method to the final speech enhancement task in this section from the following aspects.

Efficiency of Training Data. We first select sample pairs (1984 pairs in total) of 5 user groups (involving 15 participants) collected in real-world environments as the real dataset. The augmented dataset is generated as introduced in Sec. 6.1.2. Then, we train the DC-SE model as the following approaches:

- Only utilizing the real dataset for model training (Case 1) and only utilizing the augmented dataset for model training (Case 2). We also refer to Case 1 as the model without pre-training and Case 2 as the model without fine-tuning.
- First, utilizing the augmented dataset for model pre-training and then utilizing 50, 100, 150, 200, 400, 600, and 800 pairs of speech samples randomly selected from the real dataset for model fine-tuning, respectively. (Case 3-9)
- Utilizing the augmented dataset for model pre-training and utilizing all sample pairs of the real dataset for fine-tuning. (Case 10)

After each training process is completed, we use the remaining user group (involving 3 participants) to evaluate the performance of our system. The average PESQ and SI-SDR are shown in Fig. 16. The average PESQ and SI-SDR of the model that has not been fine-tuned on the real dataset (*i.e.*, Case 2) are lower than the model that is not pre-trained on the augmented dataset (*i.e.*, Case 1). This is because there is still a certain distribution difference between the augmented dataset and the real dataset. However, when we use small-scale speech samples in the

Reference	In your high school, most of the teachers there are helpful and friendly.
Noisy speech	In miao miao high school, most of the teachers they are here for are friends .
Enhanced speech	In your high school, the post of the teachers they are helpful and friendly.

Fig. 19. An example of transcription texts. Texts that do not match the reference are marked red.

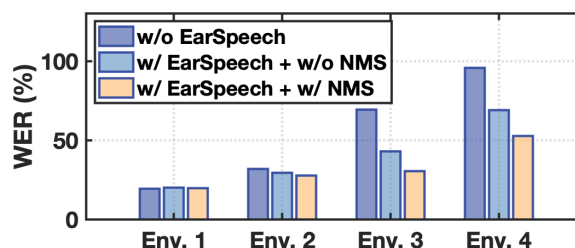


Fig. 20. WER in four real-world environments. NMS: noise mixture scheme in Sec. 5.3.1

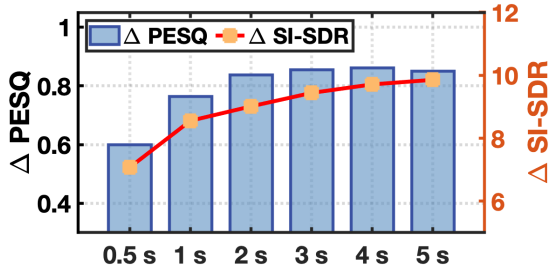
real dataset to fine-tune the pre-trained model (*i.e.*, Case 3-10), we find that the performance of the system will gradually increase and stabilize. For example, in Case 3, we only utilize 50 pairs of real two-channel speech samples to fine-tune the model pre-trained on the augmented dataset, and the fine-tuned model can achieve an average PESQ of 3.22 and an average SI-SDR of 13.11, which is comparable to the performance of the model (with an average PESQ of 3.31 and an average SI-SDR of 13.18) trained using 1984 pairs of real samples (*i.e.*, Case 1). However, the number of real training data used in Case 3 is about 40 times less than that of Case 1. In particular, when we use 100 pairs of samples to fine-tune the model (*i.e.*, Case 4), EARSPEECH achieves an average PESQ of 3.31 and an average SI-SDR of 14.12, which has outperformed Case 1. It indicates that with the benefit of our designed data augmentation method, EARSPEECH can achieve satisfactory enhancement performance by utilizing only 1/40 to 1/20 of the original data size, greatly saving collection and labeling costs.

Generalization of New Sentence. To validate the generalization of new sentences (not present in the training dataset), we still leverage the original training dataset (involving 15 participants) to train the DS-SE model with the pre-training model involved and without the pre-training model involved respectively. Then, we require three participants in the testing dataset to wear our prototype and read several new sentences. The dual-channel speech samples corresponding to new sentences are utilized to validate the performance. The speech quality and intelligibility metrics are shown in Fig. 17. We use the average PESQ, STOI, and SI-SDR of the model trained following Case 10 as references which are shown as the red dotted lines. As we can see, the performance of our system on new sentences is close to that of original sentences. It indicates a high generalization performance on new sentences. In addition, we can observe that with the benefit of data augmentation, PESQ, STOI, and SI-SDR of EARSPEECH on new sentences are significantly improved.

Generalization of Mandarin Language. In addition, we also collect dual-channel speech samples of three participants in Mandarin to study the generalization performance of different language types. The results are shown in Fig. 18. In addition, we use the average PESQ, STOI, and SI-SDR of new English sentences as references which are shown as the red dotted lines. Compared with the English corpus, EARSPEECH still achieves a comparable performance on the Mandarin corpus. In addition, the data augmentation can still benefit the improvement of average PESQ, average STOI, and average SI-SDR.

6.5 Real-world Study and Noise Mix Scheme Validation

We conduct experiments in real-world environments to validate the effectiveness of the designed noise mixture scheme. Specifically, we additionally collect dual-channel speech samples of 3 participants (not involved in training) in four environments, *i.e.*, a student office with fan noise (Env. 1, about 60.98 dB on average), a cafe with music noise (Env. 2, about 65.02 dB on average), an outside street with environmental noise (Env. 3, about 74.02 dB on average), and an indoor mall with conversation noise (Env. 4, about 73.08 dB on average). Since we are unable to obtain the clean speech signals as a reference, we adopt WER (word error rate) to evaluate the performance of







	(ET 1)	(ET 2)	(ET 3)	(ET 4)
				
Material	Memory foam	Silicone	Silicone	Silicone
Geometry	Single flange	Single flange	Double flange	Wingtips

Fig. 21. Impact of the audio length on EARSPEECH. Fig. 22. List of materials and geometries of ear tips (earbuds).

speech enhancement. We use Notta [38], a commercial online Transcribe Audio to Text platform, to process noisy speech signals and enhanced speech signals, respectively. Fig. 19 illustrates an example of transcription texts. We can find that the noisy speech enhanced by EARSPEECH is with satisfactory quality and intelligibility. Fig. 20 shows the WER in different environments. "w/o EARSPEECH" represents that the noisy speech is not enhanced by EARSPEECH. "w/ EARSPEECH+ w/o NMS" and "w/ EARSPEECH+ w/o NMS" represent the EARSPEECH is trained with and without the noise mixture scheme as introduced in Sec. 5.3.1. In Env. 1 and Env. 2, there is a subtle difference in WER of with and without EARSPEECH. That is because Notta is built on a large language model that is robust to slight external disturbance. However, when the noise becomes serious, the improvement of EARSPEECH is obvious. For example, EARSPEECH can reduce the WER from 70% to 30% in Env. 3. In addition, the WER of EARSPEECH increases significantly in Env. 4 where competing speakers are present. However, compared with "w/o EARSPEECH", our system still brings the improvement ratio of 45.3%. In addition, we also find that the performance of EARSPEECH with NMS is better than that without NMS, especially in very noisy environments. The noise mixture scheme can simulate the impact of noise on the in-ear channel. When the environment has a higher SNR, external noise has a subtle impact on the in-ear channel, as introduced in Sec. 5.3.1. Thus, our system performs equally well with and without the help of NMS. When the environment has a poor SNR, the impact of external noise is not ignored, causing a decrease in the performance of EARSPEECH without NMS. Therefore, these results also indicate the effectiveness of the designed noise mixture scheme.

6.6 Robustness Study

In this section, we study various factors that may affect the robustness of EARSPEECH. The base model is firstly pre-trained via the augmented dataset. Then we fine-tune the pre-trained model with dual-channel samples of 15 participants and evaluate the performance of EARSPEECH using three participants under different conditions. All dual-channel samples are collected with the 16 KHz sampling rate.

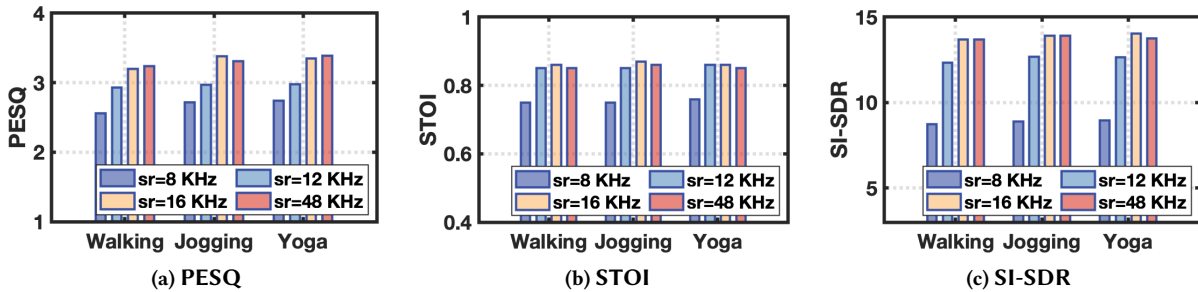
6.6.1 Impact of audio length. EARSPEECH can handle audio inputs of different lengths. With the default audio input length of 5 seconds, EARSPEECH achieves outperforming enhancement performance as introduced in Sec. 6.2. In this subsection, we also explore the impact of audio length on the enhancement performance. As shown in Fig. 21, we can observe that Δ PESQ and Δ SI-SDR increase as the audio length increases. When the length exceeds 2 s, our system performance tends to be stable. That is because longer audio clips can provide richer temporal context information (including variations in speech characteristics, background noise statistics, and contextual dependencies) that can potentially improve speech enhancement performance. However, even with an audio length of 1 second, the system still achieves an average PESQ improvement of 0.76 and an average SI-SDR improvement of 8.55 dB, which can meet the requirements of most voice input scenarios.

Table 2. The enhancement performance of EARSPEECH with different types (materials and geometries) of ear tips (ET 1, ET 2, ET 3, and ET 4).

	ET 1	ET 2	ET 3	ET 4	std
PESQ	3.28	3.26	3.25	3.26	0.0126
STOI	0.91	0.90	0.91	0.91	0.0050
SI-SDR (dB)	14.12	14.07	14.11	14.15	0.0330

Table 3. Impact of audio channel on EARSPEECH. "R": right channel. "L": left channel. "RL": fused right and left channels.

	R-R	L-R	RL-R	std
PESQ	3.21	3.16	3.17	0.0265
STOI	0.90	0.89	0.90	0.0058
SI-SDR (dB)	14.94	14.52	14.67	0.2128

**Fig. 23. Impact of different sampling rates and body movements.**

6.6.2 Impact of earphone form-fitting geometry. The form-fitting geometry of earphones may influence the occlusion effect and how much external noise is captured. Since current earphone manufacturers cannot provide audio APIs for customers, we are unable to deploy EARSPEECH on different types of commercial earphones. As an alternative solution, we select different types of ear tips (also known as earbuds) to simulate the impact of earphone form-fitting geometry. The materials (*i.e.*, memory foam and silicone) and geometries (*i.e.*, single flange, double flange, and wingtips) of selected ear tips are shown in Fig. 22. *ET 1* and *ET 2* are both single-flange ear tips but are made of different materials. *ET 2*, *ET 3*, and *ET 4* are all made of silicone material but have different geometries. Metrics of different ear tip types are shown in Tab. 2. The standard deviations of PESQ, STOI, and SI-SDR are 0.0126, 0.0050, and 0.0330, respectively, indicating that EARSPEECH generates a high generalization capability among different earphone form-fitting geometries.

6.6.3 Impact of audio channel. The dual-channel speech enhancement model is designed as a structure with two input branches, *i.e.*, less noisy in-ear speech and noisy airborne speech. In the entire evaluation, we leverage the right in-ear channel and right airborne channel by default. Due to the asymmetry of the human body, the differences between the right in-ear channel and the left in-ear channel may affect the performance of EARSPEECH. Thus, we input speech samples from different channels into the speech enhancement model to explore the impact of audio channels. Specifically, we combine the right-ear airborne channel and different in-ear channels. "R-R" refers to paired speech samples from right in-ear and right airborne channels. "L-R" refers to paired speech samples from left in-ear and right airborne channels. In addition, we also fuse in-ear speech samples from left in-ear and right in-ear channels by averaging to obtain fused single-channel in-ear speech samples. Then, we input the fused single-channel in-ear speech samples and right-channel airborne speech samples to the speech enhancement model, denoted as "RL-R". The quality metrics are shown in Tab. 3. We can observe that different in-ear channels only have a subtle impact on enhancement performance. The "R-R" still outperforms the other two ways, since our designed enhancement model is trained based on the "R-R" way.

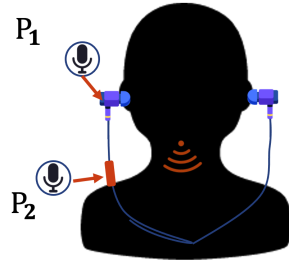


Fig. 24. Position of the outer microphone on earphones.

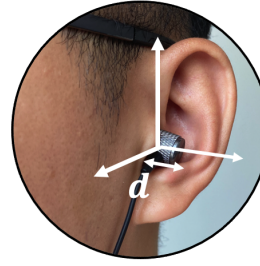


Fig. 25. Depth of earplug in the ear canal.

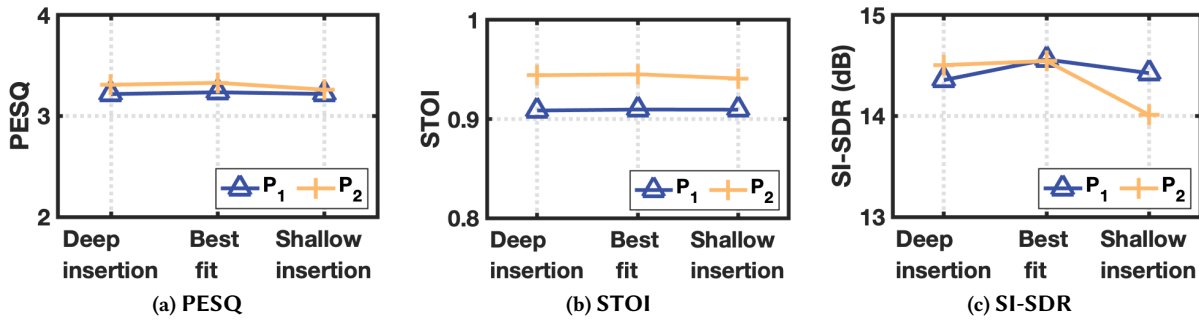


Fig. 26. Impact of relative distance and wearing position.

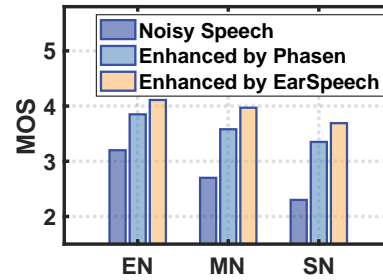
6.6.4 Impact of Body Movement. We select three daily outdoor sports scenarios to study the impact of body movement on the robustness of our system. The speech quality and intelligibility metrics are shown in Fig. 23. At the sampling rate of 16 KHz, we can find that PESQ, STOI, and SI-SDR have only a small fluctuation range among the three scenarios with STD of 0.096, 0.005, and 0.187, respectively. Prior works [16, 22, 62] have studied that in-ear acoustic responses induced by body movements are concentrated below 100 Hz, which can be removed by a low-pass filter in the pre-processing module. Thus, EARSPEECH is resistant to body movements and works well in both static and dynamic scenarios.

6.6.5 Impact of Sampling Rate. The base model is trained using dual-channel speech with 16 KHz, while we use dual-channel speech with 8 KHz, 12 KHz, and 48 KHz to study the impact of sampling rate on the robustness of EARSPEECH. The speech quality and intelligibility metrics with different sampling rates are shown in Fig. 23. We take the walking scenario as an example to analyze. As the sampling rate increases, PESQ, STOI, and SI-SDR all gradually increase. That is because frequency components below 8 KHz mainly contribute to speech quality and intelligibility. When speech samples are downsampled with a lower sampling rate (*i.e.* ≤ 16 KHz), the effective information is aliased [19]. However, microphones on most earphones support a sampling rate of 44.1 KHz or 48 KHz by default, indicating that EARSPEECH is available to most earphones.

6.6.6 Relative Distance between In-ear and Outer Microphones. The relative distance between the outer and in-ear microphones is also a key factor that needs to be considered. Generally, the in-ear microphone is fixed on the inside of the earphone, facing inward. However, the outer microphone is located in two positions as shown in Fig. 24. One position (P1) is on the exterior of the earphone, facing outward. Another position (P2) is on

Table 4. Impact of sound played from on-earphone speakers with a comfortable volume.

	PESQ	STOI	SI-SDR (dB)
w/ played sounds	3.37	0.92	13.89
w/o played sounds	3.49	0.96	14.16

**Fig. 27. Mean opinion score comparison among different types of noise, i.e., environmental noise (EN), music noise (MN), and speech noise (SN).**

the earphone cable. The results are shown in Fig. 26. We can observe that the relative distance between the in-ear and outer microphones has a subtle influence on the enhancement performance.

6.6.7 Wearing Position of Earphones. Wearing habits also vary from person to person. The insertion depth of the earplug (shown in Fig. 25) can influence the occlusion effect [5]. We qualitatively divide the depth of the earplug in the ear canal into depth insertion, best fit, and shallow insertion. It's worth noting that the earplug is ensured to completely block the ear canal, regardless of depth. We can observe from Fig. 26 that although the PESQ, STOI, and SI-SDR of shallow insertion are slightly lower than those of deep insertion and best fit, the variations of these metrics are all within a small range. The above results demonstrate that our system is robust to variations of earplug insertion depth.

6.6.8 Impact of Sound Played from On-earphone Speakers. Users often wear headphones to listen to music or other sounds. In hardware structure design, the on-earphone speaker is close to the in-ear microphone (as shown in Fig. 14), which indicates that the sounds played back over the earphone speakers may interfere with the in-ear speech collection. To explore the impact of sound played from on-earphone speakers, we select the top 5 songs from Billboard Hot 100^{TM2}. When participants wear earphones to speak, these songs are also played at a comfortable volume from on-earphone speakers. Metrics are shown in Tab. 4. We can clearly observe that EARSPEECH is robust to sounds played from on-earphone speakers. Although PESQ, STOI, and SI-SDR exhibit a little drop, EARSPEECH can still improve PESQ, STOI, and SI-SDR by 0.76, 0.13, and 8.54, respectively. That is because the noise mixture scheme designed in Sec. 5.3.1 has simulated the impact of noise on the in-ear channel. In addition, the noise dataset used in model training has included various types of music noise. Nowadays, most earphones can automatically lower the playback volume or pause playback when they detect the user's voice [40]. Thus, EARSPEECH can also be applied to scenarios where sound is played from earphones.

6.7 Subject Measurement

The subjective measurement (i.e., Mean Opinion Score, MOS) from human participants is also an important metric to measure speech quality and intelligibility. In our work, we divide the quality of speech into five levels [59], i.e., *Excellent* - 5, *Good* - 4, *Fair* - 3, *Poor* - 2, and *Bad* - 1. Then, we invited 10 participants aged between 20 - 40 years old to score noisy audio samples, audio samples enhanced by *Phasen*, and audio samples enhanced by EARSPEECH. It is noted that the involved participants did not suffer from hearing loss before. The mean opinion score comparison among different types of noise is shown in Fig. 27. EARSPEECH outperforms *Phasen* by 0.26, 0.39, and 0.34 in terms of environment noise, music noise, and speech noise, respectively. These surprising results

²<https://www.billboard.com/charts/hot-100/>, WEEK OF APRIL 20, 2024.

indicate that our system not only yields the best performance in objective evaluation metrics (*i.e.*, PESQ, STOI, and SI-SDR) but also performs best in the subjective evaluation metric.

Table 5. Run-time latency for a pair of 5 s audio clips.

Platform	Pre-processing	Inference	Total
Laptop GPU	4.79 ms (± 0.72 ms)	38.39 ms (± 0.06 ms)	36.51 ms (± 7.35 ms)
Laptop CPU	7.80 ms (± 0.78 ms)	1.64 s (± 0.16 s)	1.71 s (± 0.13 s)

6.8 Run-time Latency Evaluation

Considering the limited computing capability of earphones, we adopt the client-server mode to evaluate the system latency. The earphone prototype only collects the dual-channel speech and then transmits it to a paired laptop via a wired connection. EARSPEECH is deployed on the paired laptop with two hardware settings. One hardware setting is with CPUs (Intel(R) Xeon(R) Silver 4210R) and another setting is with a GPU (NVIDIA GeForce RTX 3090). In our work, we define the system latency as the time difference between the user ending the voice input and EARSPEECH outputting the enhanced airborne speech [55, 65]. It is noted that we ignore the data uploading latency since the data is transmitted via the wired connection. Thus, the measured latency can also be considered as the run-time latency. Tab. 5 shows the average run-time latency for a pair of 5 s audio clips repeated to be calculated 10 times. EARSPEECH can achieve an average overall latency of 36.51 ms on the GPU platform, even on the CPU platform, our system can still achieve an average overall latency of 1.71 s. According to the definition of real-time [27, 61], our system can meet the real-time requirements, since the processing latency is less than the lengths of raw audio clips.

7 RELATED WORK

7.1 Audio-only Speech Enhancement

Deep learning-based technologies have been widely adopted in the speech enhancement task [56], showing significant improvement in performance compared with conventional denoising methods such as spectral subtraction [24], filtering [1], and subspace decomposition [12]. Previous DL-based denoising technologies mainly utilize distribution differences between clean speech and noise in the time or time-frequency domains. Generally, these technologies artificially add noise at several SNRs to the clean speech to train the speech enhancement model. Based on the training target, these technologies are divided into mapping-based and masking-based. The mapping-based technologies aim to directly map noisy speech into clean speech during the model training process [44–46]. However, masking-based technologies aim to predict an amplitude mask from the input noisy amplitude spectrogram [56, 64, 66]. The amplitude mask represents the ratio of clean target speech to noise. The enhanced amplitude spectrogram is determined by multiplying the predicted amplitude mask with the noisy amplitude spectrogram. Compared with mapping-based technologies, masking-based technologies are more sensitive to SNR variations but are more efficient in training on the limited-scale training dataset [66]. Therefore, in our work, we take the basic idea of amplitude masking for speech enhancement.

7.2 Multi-modality Speech Enhancement

With the advance of multi-modality fusion, many studies have started to leverage non-audio modalities (*e.g.*, visual signal [32], mmWave [30, 41, 63], and ultrasound [11, 55, 65]) that are less sensitive to ambient noise as complementary modalities for speech enhancement. Compared with visual-based methods, wireless-based

methods don't raise too many concerns about privacy and attract more research attention. For example, Liu *et al.* [30] have explored fusing mmWave signals with audio signals for noise-resistant speech recognition. Unlike mmWave signals, ultrasound-based methods are based on available sensors (*i.e.*, microphone and loudspeaker) on most commercial devices, which don't induce additional hardware costs. Sun *et al.* [55], Zhang *et al.* [65] and Ding *et al.* [11] all leverage a loudspeaker of the mobile phone to emit ultrasound signals to capture articulatory gestures and exploit the relationship between ultrasound signals and audio signals for speech enhancement.

7.3 Speech Enhancement Towards Earphones

Earphones have become one of the most popular wearable devices, making them considered a promising sensing platform [4, 10, 13, 21, 22, 31]. As people use earphones more and more frequently for voice interaction (*e.g.*, making a call), improving the quality of input voice on earphones is a problem worthy of long-term exploration. Unlike other smart devices, earphones are small-size and resource-constrained, making it difficult to integrate additional sensors on them. Recently, He *et al.* [23] leverage the microphone and accelerometer on earphones to capture airborne speech and corresponding bone-conducted vibrations. However, it requires the accelerometer to support sampling rates up to 1.6 KHz which is not accessible to most commercial earphones as claimed in [28, 33]. In addition, Chatterjee *et al.* [8] leverage left and right microphones on earphones for speech enhancement, which is not suitable for single-earphone scenarios (*e.g.*, using one earphone for work and charging the other earphone). Recently, prior works started exploring bone conduction (BC) speech enhancement [9, 50, 51, 60, 67]. For example, Wang *et al.* [60] study the relationship of complex spectral mapping between BC speech and airborne speech for speech enhancement. However, the acquisition of BC speech requires a special transducer that is in contact with the skull bones and converts the bone vibrations to speech. Unlike BC speech, in-ear speech refers to sound vibrations affected by the occlusion effect in the ear canal. This type of speech can be easily detected by a ubiquitous microphone which can be seamlessly integrated into most current earphones. Thus, our work focuses on exploiting a cross-channel relationship between in-ear speech and airborne speech for speech enhancement.

8 DISCUSSION

Our work focuses on exploring the occlusion effect-based correlation between airborne and in-ear channels for speech enhancement. Through real-world experiments, we demonstrate the effectiveness, robustness, and generalization of our system. Nevertheless, there are several limitations that need to be addressed.

(1) *OS Interface's Public Access.* We conducted a survey about the API access to operation systems on existing earphones including Apple AirPods Series, Bose QuietComfort Series, Samsung Galaxy Buds Series, and HUAWEI FreeBuds Series. We find that due to considerations such as intellectual property protection and security, most earphone manufacturers generally do not provide us with earphone operation system interfaces, *e.g.*, audio access interfaces, network interfaces, and application programming interfaces. Therefore, we are unable to deploy EARSPEECH on commercial earphones to conduct an evaluation of practical performance. To solve this problem, in our work, we implement a proof-to-concept prototype and adopt a client-server mode to conduct a preliminary evaluation of the performance of our system. In the future, as researchers pay more attention to earphone-based sensing and computing, we believe that earphone manufacturers will open rich operation system interfaces to the public, allowing us to deploy EARSPEECH on the earphone side.

(2) *Lightweight Computing.* Existing commercial earphones vary in terms of their computing capabilities, as they are primarily designed for audio playback and communication purposes rather than intensive computation. Some advanced earphones like Apple AirPods Pro, Sony WF-1000XM4, and Samsung Galaxy Buds Pro, have incorporated certain computing capabilities to support audio processing, noise cancellation, connectivity, etc. However, it is still difficult for them to support complex model computing. At the beginning of the enhanced model design, we did consider the size and computational complexity of the model. Although the total parameters

of the DC-SE model are approximately 3.8 MB, EARSPEECH may be difficult to directly run on earphone sides. We can explore model compression technologies (like pruning, knowledge distillation, and quantization) to accelerate computing while keeping a good enhancement performance.

(3) *Sound Played from On-earphone Speakers*. In Sec. 6.6.8, we explore the impact of sound playback from on-earphone speakers and find that replayed sound with a normal volume only has a subtle impact on EARSPEECH. However, when sound is played at high volume, the in-ear speech will be overwhelmed by noise, leading to degradation of enhancement performance. However, in practice, users only listen to music at comfortable volumes. Additionally, most earphones have supported Speak-to-Chat mode, which can automatically lower the playback volume or pause playback when they detect the user’s voice [40].

9 CONCLUSION

In this work, we design EARSPEECH, an earphone-based speech enhancement technology. EARSPEECH utilizes outer and in-ear microphones on a single earphone to capture sound vibrations propagating along different channels and exploits the occlusion effect-based correlation between them to improve the quality and intelligibility of airborne speech. Based on the occlusion effect, we design a data augmentation method to generate a large-scale synthetic dual-channel speech corpus. Then, a deep learning-based speech enhancement model is designed to effectively fuse dual-channel speech signals with heterogeneous structures to remove noise components from target speech components. Through pre-training with the augmented dataset, EARSPEECH can achieve excellent enhancement performance by using only about 1/40 of the original samples for fine-tuning. Comprehensive experiments have demonstrated the effectiveness, robustness, and generalization ability of EARSPEECH.

ACKNOWLEDGMENTS

Panlong Yang and Fenglei Xu are the corresponding authors. This work is partially supported by the National Natural Science Foundation of China with No. 62072004, and Jiangsu Province Higher Education Connotation Construction and Development-2024 Double First-Class-Talent Start-up Fee 1523142401052. We also appreciate the valuable suggestions and feedback from the anonymous reviewers.

REFERENCES

- [1] Marwa A Abd El-Fattah, Moawad I Dessouky, Alaa M Abbas, Salaheldin M Diab, El-Sayed M El-Rabaie, Waleed Al-Nuaimy, Saleh A Alshebeili, and Fathi E Abd El-samie. 2014. Speech enhancement with an adaptive Wiener filter. *International Journal of Speech Technology* 17 (2014), 53–64.
- [2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. arXiv:1912.06670 [cs.CL]
- [3] Deepak Baby and Sarah Verhulst. 2019. Sergan: Speech enhancement using relativistic generative adversarial networks with gradient penalty. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 106–110.
- [4] Yetong Cao, Chao Cai, Fan Li, Zhe Chen, and Jun Luo. 2023. HeartPrint: Passive Heart Sounds Authentication Exploiting In-Ear Microphones. *Heart* 50, S1 (2023), S2.
- [5] Yetong Cao, Chao Cai, Anbo Yu, Fan Li, and Jun Luo. 2023. EarAce: Empowering Versatile Acoustic Sensing via Earable Active Noise Cancellation Platform. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 2 (2023), 1–23.
- [6] Kévin Carillo, Olivier Doutres, and Franck Sgard. 2020. Theoretical investigation of the low frequency fundamental mechanism of the objective occlusion effect induced by bone-conducted stimulation. *The Journal of the Acoustical Society of America* 147, 5 (2020), 3476–3489.
- [7] Kévin Carillo, Olivier Doutres, and Franck Sgard. 2021. On the removal of the open earcanal high-pass filter effect due to its occlusion: A bone-conduction occlusion effect theory. *Acta Acustica* 5 (2021), 36.
- [8] 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*. 384–396.
- [9] Liangliang Cheng, Yunfeng Dou, Jian Zhou, Huabin Wang, and Liang Tao. 2023. Speaker-Independent Spectral Enhancement for Bone-Conducted Speech. *Algorithms* 16, 3 (2023), 153.

- [10] Romit Roy Choudhury. 2021. Earable computing: A new area to think about. In *Proceedings of the 22nd International Workshop on Mobile Computing Systems and Applications*. 147–153.
- [11] 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* 6, 3 (2022), 1–25.
- [12] Yariv Ephraim and Harry L Van Trees. 1995. A signal subspace approach for speech enhancement. *IEEE Transactions on speech and audio processing* 3, 4 (1995), 251–266.
- [13] Xiaoran Fan, David Pearl, Richard Howard, Longfei Shangguan, and Trausti Thormundsson. 2023. APG: Audioplethysmography for Cardiac Monitoring in Hearables. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.
- [14] 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*. 337–349.
- [15] Szu-Wei Fu, Yu Tsao, Xugang Lu, and Hisashi Kawai. 2017. Raw waveform-based speech enhancement by fully convolutional networks. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 006–012.
- [16] 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* 5, 1 (2021), 1–25.
- [17] John S Garofolo. 1993. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993* (1993).
- [18] Theodoros Giannakopoulos. 2009. A method for silence removal and segmentation of speech signals, implemented in Matlab. *University of Athens, Athens 2* (2009).
- [19] Feiyu Han, Panlong Yang, Haohua Du, and Xiang-Yang Li. 2022. Accuth: Anti-Spoofing Voice Authentication via Accelerometer. In *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*. 637–650.
- [20] Feiyu Han, Panlong Yang, Haohua Du, and Xiang-Yang Li. 2024. Accuth⁺⁺: Accelerometer-Based Anti-Spoofing Voice Authentication on Wrist-Worn Wearables. *IEEE Transactions on Mobile Computing* 23, 5 (2024), 5571–5588.
- [21] Feiyu Han, Panlong Yang, Yuanhao Feng, Weiwei Jiang, Youwei Zhang, and Xiang-Yang Li. 2024. EarSleep: In-ear Acoustic-based Physical and Physiological Activity Recognition for Sleep Stage Detection. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–31.
- [22] Feiyu Han, Panlong Yang, Shaojie Yan, Haohua Du, and Yuanhao Feng. 2023. BreathSign: Transparent and Continuous In-ear Authentication Using Bone-conducted Breathing Biometrics. In *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 1–10.
- [23] 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*. 14–27.
- [24] Sunil Kamath, Philippos Loizou, et al. 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise.. In *ICASSP, Vol. 4*. Citeseer, 44164–44164.
- [25] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7482–7491.
- [26] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).
- [27] Sen M Kuo, Bob H Lee, and Wenshun Tian. 2013. *Real-time digital signal processing: fundamentals, implementations and applications*. John Wiley & Sons.
- [28] Jianwei Liu, Wenfan Song, Leming Shen, Jinsong Han, and Kui Ren. 2022. Secure user verification and continuous authentication via earphone imu. *IEEE Transactions on Mobile Computing* (2022).
- [29] Tianyi Liu, Minshuo Chen, Mo Zhou, Simon S Du, Enlu Zhou, and Tuo Zhao. 2019. Towards understanding the importance of shortcut connections in residual networks. *Advances in neural information processing systems* 32 (2019).
- [30] Tiantian Liu, Ming Gao, Feng Lin, Chao Wang, Zhongjie Ba, Jinsong Han, Wenyao Xu, and Kui Ren. 2021. Wavoice: A noise-resistant multi-modal speech recognition system fusing mmwave and audio signals. In *Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems*. 97–110.
- [31] 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*. 175–187.
- [32] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1368–1396.
- [33] Vimal Mollyn, Riku Arakawa, Mayank Goel, Chris Harrison, and Karan Ahuja. 2023. IMUPoser: Full-Body Pose Estimation using IMUs in Phones, Watches, and Earbuds. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [34] Online. 2022. Basic English Speaking. <https://basicenglishspeaking.com>
- [35] Online. 2023. Alango Technologies - Making Digital Sound Better. Technologies. Voice Enhancement. OnlyVoice™ with In-Ear Microphone. <https://www.alango.com/technologies-onlyvoice-in-ear-microphone.php>.

- [36] Online. 2023. Earphones And Headphones Market Size, Share & Trends Analysis Report By Price Band (>100, 50-100, <50), By Product (Earphones And Headphones), By Technology, By Application, By Region, And Segment Forecasts, 2023 - 2030. <https://www.researchandmarkets.com/reports/4118850/earphones-and-headphones-market-size-share-and>. (Accessed on 10/30/2023).
- [37] Online. 2023. PyTorch. <https://pytorch.org/>.
- [38] Online. 2023. Transcribe Audio to Text | Notta. <https://www.notta.ai/en>. (Accessed on 11/09/2023).
- [39] Online. 2024. AS-B6027AL30-RC microphone. <http://www.aospow.com/Products/znjqry6mmm.html>. (Accessed on 04/21/2024).
- [40] Online. 2024. WH-1000XM4 | Help Guide | Speaking with someone while wearing the headset (Speak-to-Chat). <https://helpguide.sony.net/mdr/wh1000xm4/v1/en/contents/TP0002754732.html>. (Accessed on 04/17/2024).
- [41] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, Min Wu, and KJ Ray Liu. 2023. Radio SES: mmWave-Based Auditoradio Speech Enhancement and Separation System. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 1333–1347.
- [42] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. 2011. The importance of phase in speech enhancement. *speech communication* 53, 4 (2011), 465–494.
- [43] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5206–5210.
- [44] Ashutosh Pandey and DeLiang Wang. 2019. TCNN: Temporal convolutional neural network for real-time speech enhancement in the time domain. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6875–6879.
- [45] Se Rim Park and Jinwon Lee. 2016. A fully convolutional neural network for speech enhancement. *arXiv preprint arXiv:1609.07132* (2016).
- [46] Santiago Pascual, Antonio Bonafonte, and Joan Serra. 2017. SEGAN: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452* (2017).
- [47] Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1015–1018.
- [48] Timothy Pommée, Mathieu Balaguer, Julien Pinquier, Julie Mauclair, Virginie Woisard, and Renée Speyer. 2021. Relationship between phoneme-level spectral acoustics and speech intelligibility in healthy speech: a systematic review. *Speech, Language and Hearing* 24, 2 (2021), 105–132.
- [49] Tobias Röddiger, Christian Dinse, and Michael Beigl. 2021. Wearability and comfort of earables during sleep. In *Proceedings of the 2021 ACM International Symposium on Wearable Computers*. 150–152.
- [50] Dongjing Shan, Xiongwei Zhang, Chao Zhang, and Li Li. 2018. A novel encoder-decoder model via NS-LSTM used for bone-conducted speech enhancement. *IEEE Access* 6 (2018), 62638–62644.
- [51] Premjeet Singh, Manoj Kumar Mukul, and Rajkishore Prasad. 2018. Bone conducted speech signal enhancement using LPC and MFCC. In *Intelligent Human Computer Interaction: 10th International Conference, IHCI 2018, Allahabad, India, December 7–9, 2018, Proceedings 10*. Springer, 148–158.
- [52] David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. [arXiv:1510.08484](https://arxiv.org/abs/1510.08484) [arXiv:1510.08484v1](https://arxiv.org/abs/1510.08484v1).
- [53] Stefan Stenfeldt. 2012. Transcranial attenuation of bone-conducted sound when stimulation is at the mastoid and at the bone conduction hearing aid position. *Otology & neurotology* 33, 2 (2012), 105–114.
- [54] Stefan Stenfeldt and Sabine Reinfeldt. 2007. A model of the occlusion effect with bone-conducted stimulation. *International journal of audiology* 46, 10 (2007), 595–608.
- [55] 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*. 160–173.
- [56] Tayseer MF Taha, Ahsan Adeel, and Amir Hussain. 2018. A survey on techniques for enhancing speech. *International Journal of Computer Applications* 179, 17 (2018), 1–14.
- [57] Tomoki Toda, Alan W Black, and Keiichi Tokuda. 2007. Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing* 15, 8 (2007), 2222–2235.
- [58] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. 2021. Multi-task learning for dense prediction tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence* 44, 7 (2021), 3614–3633.
- [59] Mahesh Viswanathan and Madhubalan Viswanathan. 2005. Measuring speech quality for text-to-speech systems: development and assessment of a modified mean opinion score (MOS) scale. *Computer speech & language* 19, 1 (2005), 55–83.
- [60] Heming Wang, Xueliang Zhang, and DeLiang Wang. 2022. Fusing Bone-Conduction and Air-Conduction Sensors for Complex-Domain Speech Enhancement. *IEEE/ACM transactions on audio, speech, and language processing* 30 (2022), 3134–3143.
- [61] Zhong-Qiu Wang, Gordon Wichern, Shinji Watanabe, and Jonathan Le Roux. 2022. STFT-domain neural speech enhancement with very low algorithmic latency. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2022), 397–410.
- [62] 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 *IEEE INFOCOM 2022-IEEE Conference on Computer Communications*. IEEE, 1789–1798.
- [63] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyaoy Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International*

- Conference on Mobile Systems, Applications, and Services*. 14–26.
- [64] 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*, Vol. 34. 9458–9465.
 - [65] 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* 5, 3 (2021), 1–30.
 - [66] Xiao-Lei Zhang and DeLiang Wang. 2016. A deep ensemble learning method for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 24, 5 (2016), 967–977.
 - [67] Changyan Zheng, Tiejong Cao, Jibin Yang, Xiongwei Zhang, and Meng Sun. 2019. Spectra restoration of bone-conducted speech via attention-based contextual information and spectro-temporal structure constraint. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences* 102, 12 (2019), 2001–2007.