1-2024

ClearSpeech: Improving voice quality of earbuds using both in-ear and out-ear microphones

Dong MA
Singapore Management University, dongma@smu.edu.sg

Ting DANG

Ming DING

Rajesh Krishna BALAN
Singapore Management University, rajesh@smu.edu.sg

Citation

MA, Dong; DANG, Ting; DING, Ming; and BALAN, Rajesh Krishna. ClearSpeech: Improving voice quality of earbuds using both in-ear and out-ear microphones. (2024). *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. 7, (4),.
Available at: https://ink.library.smu.edu.sg/sis_research/8579

ClearSpeech: Improving Voice Quality of Earbuds Using Both In-Ear and Out-Ear Microphones

DONG MA*, Singapore Management University, Singapore
TING DANG, Nokia Bell Labs, Cambridge, United Kingdom
MING DING, Data61, CSIRO, Australia
RAJESH BALAN, Singapore Management University, Singapore

Wireless earbuds have been gaining increasing popularity and using them to make phone calls or issue voice commands requires the earbud microphones to pick up human speech. When the speaker is in a noisy environment, speech quality degrades significantly and requires speech enhancement (SE). In this paper, we present ClearSpeech, a novel deep-learning-based SE system designed for wireless earbuds. Specifically, by jointly using the earbud's in-ear and out-ear microphones, we devised a suite of techniques to effectively fuse the two signals and enhance the magnitude and phase of the speech spectrogram. We built an earbud prototype to evaluate ClearSpeech under various settings with data collected from 20 subjects. Our results suggest that ClearSpeech can improve the SE performance significantly compared to conventional approaches using the out-ear microphone only. We also show that ClearSpeech can process user speech in real-time on smartphones.

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

Additional Key Words and Phrases: Speech Enhancement, Smart Earbuds, Earables, Audio Processing

ACM Reference Format:
Dong Ma, Ting Dang, Ming Ding, and Rajesh Balan. 2023. ClearSpeech: Improving Voice Quality of Earbuds Using Both In-Ear and Out-Ear Microphones. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 4, Article 170 (December 2023), 25 pages. https://doi.org/10.1145/3631409

1 INTRODUCTION

Wireless earbuds have become increasingly popular because (1) many phone manufacturers have removed the headphone jack on the newest phone models and recommend using wireless headsets, and (2) wireless earbuds offer a hands-free way to listen to music and participate in phone/video calls. As such, wireless earbuds manufacturers, such as Sony, Apple, Google, etc., have strongly touted their products' ability to provide good listening (i.e., downlink, targeting the wearer) and calling (i.e., uplink, targeting the call recipient) experience even in noisy outdoor environments. For listening tasks, a highly touted and effective function known as active noise cancellation (ANC) has been integrated into many off-the-shelf wireless earbuds (e.g., Apple Airpods, Sony WF-1000XM, etc.). ANC usually leverages both in-ear and out-ear microphones to eliminate external noise. For

[*] Corresponding author: Dong Ma, dongma@smu.edu.sg, Singapore Management University

Authors' addresses: Dong Ma, dongma@smu.edu.sg, Singapore Management University, Singapore; Ting Dang, ting.dang@nokia-bell-labs.com, Nokia Bell Labs, Cambridge, United Kingdom; Ming Ding, ming.ding@data61.csiro.au, Data61, CSIRO, Australia; Rajesh Balan, rajesh@smu.edu.sg, Singapore Management University, Singapore.

2474-9567/2023/12-ART170 $15.00
https://doi.org/10.1145/3631409

calling tasks, however, as shown in Section 2.1, the quality of voice calls is still unsatisfying even on recent high-grade wireless earbuds, as they mainly utilize the out-ear microphones to capture the wearer's speech.

The problem of improving the quality of calls is known as speech enhancement (SE). Existing SE solutions mainly take the speech captured from the out-ear microphone as input, and apply either digital signal processing (DSP) [5, 27, 32] and/or deep learning (DL) techniques [10, 26, 36, 37, 42] to improve the speech quality. The current state-of-the-art techniques use DL [43, 44]. Some commercial earbuds use dedicated bone-conduction microphones to capture human speech. However, they not only incur extra system overhead but also fail to accurately capture spoken consonants that do not generate bone vibrations [25]. Additionally, there is a stream of work performs SE with the assistance of another modality (e.g., ultrasound [26], video [36]), which requires special settings (e.g., facing to human mouth to capture the mouth movements) and is not applicable to wireless earbuds.

In this paper, we take a different approach that leverages the in-ear and out-ear microphones on wireless earbuds for SE. Note: although ANC also utilizes the two microphones, it solves the opposite and easier problem we are addressing. In particular, the target signal (i.e., music/voice sent to the earphone) is known in ANC, while the intended human speech is unavailable in SE. Our intuition for solving the SE problem builds upon two facts: (1) the out-ear microphone captures air-conduction speech, while the in-ear microphone provides complementary information by independently capturing both bone- and air-conducted speech data; (2) due to the isolation provided by the ear canal, the in-ear microphone signal is cleaner (i.e. with a much higher SNR) compared to the out-ear signal.

However, operationalizing these insights is not easy, and we face three challenges: (1) existing SE approaches generally need large datasets containing clean speech and noise, to synthesize noisy speech as input for model training. However, all the existing publically available speech and noise databases are recorded using out-ear microphones only, making in-ear noisy speech synthesis impossible. On the other hand, collecting a large-scale {in-ear, out-ear} paired dataset that covers various noise conditions from scratch is time-consuming and labor-intensive; (2) due to the occlusion effect [28, 33], speech data captured by in-ear and out-ear microphones exhibit different frequency characteristics and signal levels (Figure 2), for example, the loss of the high-frequency components coupled with the amplification of the low-frequency components for in-ear speech, as well as a lower magnitude level for out-ear speech. Thus, effectively exploiting and combining both signals to aid enhancement is challenging; (3) phase is crucial to the speech quality [36]. However, estimating the clean phase from the noisy speech is particularly challenging as the phase exhibits few structure patterns and is sensitive to noise [40]. For instance, a jumpy or non-stationary noise will incur phase deviation for all frequency components over time, leading to non-smooth phase estimation results across both frequency and time domains.

To address these challenges, we present ClearSpeech, a system that jointly combines the in-ear and out-ear microphone signals for SE on wireless earbuds. It consists of three modules to tackle the three challenges, respectively. First (**Section 4**), we propose an out-to-in (O2I) ear sound transformation model to convert the noise signals in the public noise database to their in-ear version. This allows our solution to reuse existing noise datasets to generate accurately synthesized in-ear and out-ear noisy speech and eliminate the need for in-ear noise data collection. Although we still need to collect the clean in-ear speech data for model training, the overhead is much smaller compared to noise data collection which needs to consider various noise types, collection environments, noise volumes, mixture of noises, etc., to ensure the robustness of the model. Second (**Section 5**), through deep frequency analysis of the two signals (in-ear and out-ear), we (1) design a DL structure to learn the low and high frequency components separately from in-ear and out-ear signals respectively, (2) propose a novel gate mechanism to learn their mutual information, and (3) design a new loss function that forces the model to focus on recovering the middle-frequency information lost due to the occlusion effect. Third (**Section 6**), via numerical analysis, we proved that quantized phases are sufficient for SE and only the phase of low-frequency part is critical to speech quality. We then designed a deep neural network to estimate the quantized phase and a new loss function that incorporates the ordinal and periodicity of the phase.

To evaluate ClearSpeech, we implemented it using an earbud prototype and then collected data from 20 subjects to perform rigorous micro benchmark evaluation[1]. The results show that our system can effectively remove various types of noise, outperforming existing approaches, and can be executed in real-time on both the GPU and mobile phones. In addition, ClearSpeech leverages the existing hardware used by modern wireless earbuds and can thus be implemented solely as a software update to them. **Note that we uploaded the audio clips processed by different approaches as a supplementary file so that the reviewers can have a perceptual evaluation.** Overall, our paper makes the following contributions:

- To the best of our knowledge[2], we are the first to comprehensively explain and design the joint use of in-ear and out-ear microphones in wireless earbuds for speech enhancement.
- We present ClearSpeech, a DL-based SE system on wireless earbuds. ClearSpeech is built upon the deep analysis of in-ear and out-ear speech patterns and consists of a suite of techniques including O2I sound transformation, in-ear & out-ear information exchange, quantized phase estimation, and etc.
- With comprehensive experiments under various conditions, we demonstrated the effectiveness and superior SE performance of ClearSpeech compared to state-of-the-art baselines. We also show that ClearSpeech can process user speech in real-time on smartphones.

2 MOTIVATION & BACKGROUND

2.1 Motivation

As discussed, most existing commercial wireless earbuds provide ANC to suppress the external noise to the wearer (i.e., downlink). Unfortunately, ANC does not help improve the transmitted speech quality during phone calls (i.e., uplink). Together with the wide employment of voice assistants on various forms of wearables, uplink noise cancellation (i.e., improving the recorded speaker's speech quality) on wireless earbuds is critical as it can significantly improve communication efficiency. For example, many of us have experienced occasions when the call recipient fails to hear our speech clearly when in a noisy environment (e.g., buses, bars, human crowds, etc.).

To investigate how good current wireless earbuds are at uplink noise cancellation, we recorded speech samples using two advanced commercial earbuds, Airpods Pro and Jabra Elite 7 Pro, at three real-world locations – i) office (noise levels of 30-40 dB), ii) at a crossroad intersection (90-100 dB), and iii) on a bus (95-105 dB). Figure 1 shows the spectrograms of the recorded speech. We observe that human speech can be identified in an office environment, while the harmonics and formants are almost overwhelmed by the noise in the crossroads and bus scenarios. Thus, even with high-grade earbuds, the acquired speech still suffers from poor quality under noisy conditions and therefore developing better SE techniques for wireless earbuds is important.

2.2 In-ear Mic vs Out-ear Mic

Given that many commercial earbuds have been equipped with an in-ear microphone for ANC, in this work, we explore the feasibility of employing the in-ear microphone for SE – reusing this mic requires zero hardware overhead. To understand how an in-ear microphone can be used for SE, we first present a comparison of the signals captured by the in-ear and out-ear microphones. As shown in Figure 2, we recorded the in-ear and out-ear microphone signals concurrently in three cases: (i) the wearer speaks in a quiet environment (1st row),

[1] The dataset will be released to the public when the paper is accepted.

[2] With a thorough search in Google Scholar, ACM Digital Library, and IEEE Xplore.

[3] Note that this is not a direct comparison of Airpods Pro and Jabra Elite 7 Pro as the audio is not recorded concurrently (i.e., the background noise is different), while the purpose here is to reveal that both commercial earbuds suffer from poor voice capture performance under strong noise.

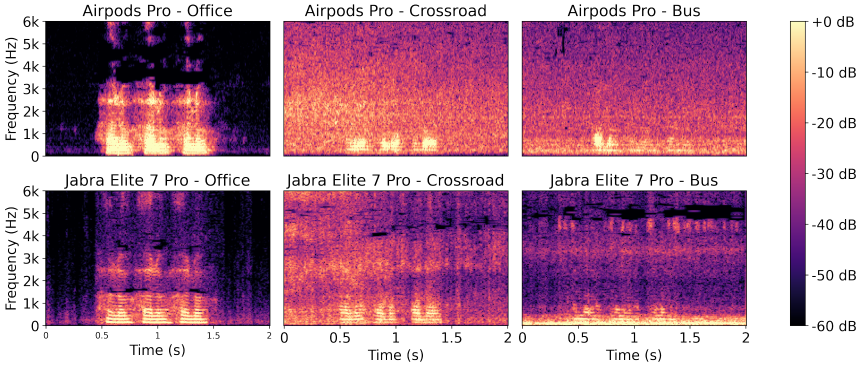


Fig. 1. Comparison of the speech spectrograms of two commercial earbuds (Airpods Pro and Jabra Elite 7 Pro) under quiet (office) and noisy (crossroad and bus) environment[3]. The wearer speaks 'hello' for three times.

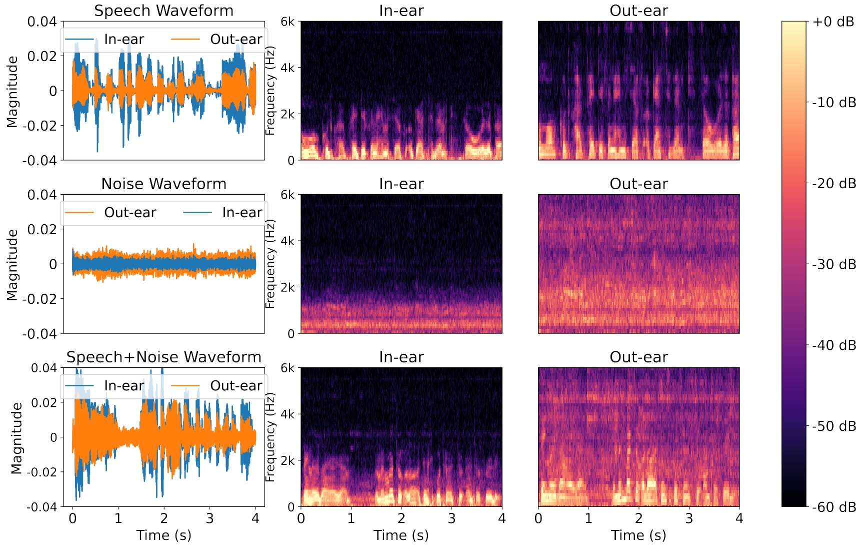


Fig. 2. Comparison of the in-ear and out-ear signals for clean speech, noise, and noisy speech, respectively.

(ii) the wearer stays in a noisy environment without speaking (2nd row), and (iii) the wearer speaks in a noisy environment (3rd row). Overall, we observe that the spectrograms for in-ear and out-ear signals are distinctly different in all three cases.

In particular, for case (i), it is clear that in-ear speech has a higher amplitude than out-ear speech in the time domain. This occurs due to two reasons: (1) the in-ear microphone captures both bone-conducted and air-conducted speech, while the out-ear microphone only captures air-conducted speech; (2) a phenomenon known as the occlusion effect[4] [28, 33] would not only amplify the low-frequency components (< 1 kHz [23]) but

[4] When the ear canal is sealed by the earbud, a closed chamber is formed between the eardrum and the earbud. As a result, the low-frequency bone-conducted speech would be amplified due to the occlusion effect.

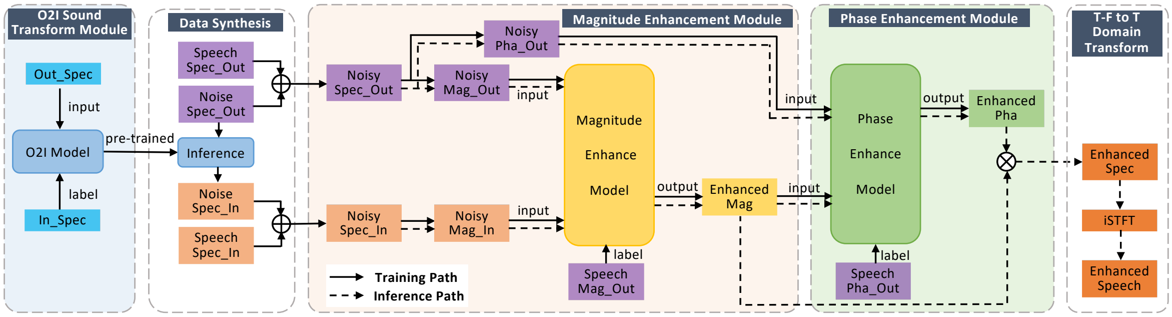


Fig. 3. Flowchart of ClearSpeech. Training goes through the solid arrows, while run-time inference flows through dashed arrows. *Spec*, *Mag*, and *Pha* represent spectrogram, magnitude, and phase. *In* and *Out* represent in-ear and out-ear signals.

also suppress the high-frequency components of the bone-conducted speech inside the ear canal. Consequently, the energy in the spectrogram is higher at low frequencies, while almost completely missing at higher frequencies. From this insight, we obtain the following observation.

Observation 1*: As the occlusion effect only enhances the low-frequency speech components, the SNR of the in-ear signal is higher than that of the corresponding out-ear signal at low frequencies (<1 kHz).*

For case (ii), the out-ear microphone captures the full spectrogram of the external noise, while the high-frequency components of the in-ear signal are eliminated by the occlusion effect. As there is no bone-conducted speech, the amplitude of the out-ear microphone is higher in this case. For case (iii), the spectrograms reveal that the out-ear microphone is severely affected by the external noise, while the in-ear microphone encounters a limited impact. The above discoveries demonstrate that (1) the in-ear microphone is more resilient to external noise; and (2) in-ear and out-ear microphones provide complimentary speech information at different frequency bands, suggesting that SE performance improvements are possible by combining both signals.

3 SYSTEM OVERVIEW

In this section, we present ClearSpeech, a dual-microphone-based SE system on wireless earbuds. Figure 3 illustrates the flowchart of ClearSpeech, which consists of five modules. Specifically, *the three color-shaded modules are designed to tackle the three challenges described in the Introduction respectively*. Next, we briefly describe each of the modules.

O2I Sound Transformation Module (Section 4): This module aims to address the first challenge - there is no publicly-available {out-ear, in-ear} paired noise dataset, by generating accurate in-ear noise from existing out-ear noise datasets. To do this, we first collected a small-scale dataset that contains concurrently recorded in-ear and out-ear sounds of different noise types. Then, the out-ear and in-ear sounds are respectively used as the input and label to train a deep learning model (O2I model) that learns the mapping between them.

Data Synthesis: After obtaining the pre-trained O2I model, we generate the in-ear version of existing large-scale out-ear noise datasets by performing model inference. Then, we synthesize noisy out-ear speech using clean out-ear speech and noise, and noisy in-ear speech using corresponding clean in-ear speech and converted noise.

Magnitude Enhancement Module (Section 5): This module is designed to tackle the second challenge by enhancing the magnitude of the Time-Frequency (T-F) spectrogram. To do this, we use the two noisy out-ear and in-ear magnitude spectrograms as inputs and the clean speech (i.e., from out-ear) as the label to train a DL-based magnitude enhancement model. Specifically, based on the frequency characteristics of the in-ear and out-ear speech data, we devised a deep neural network that (1) builds upon Observation 1 and captures the unique characteristics of in-ear and out-ear speech respectively by emphasizing the low- and high-frequency components from the two streams separately; (2) exploits the correlation between the in-ear and out-ear speech via a gate mechanism; and (3) enhances the learning of salient frequency components via a customized loss function according to Observation 2 (Section 5.6).

Phase Enhancement Module (Section 6): This module aims to estimate the clean phase from the noisy speech. In particular, we feed both the noisy phase from the out-ear microphone and the enhanced magnitude from the former module to the phase enhancement model, as phase is closely associated with the magnitude [44]. With the supervision of the phase from clean out-ear speech, the trained model will output the enhanced phase. Specifically, based on our findings that (1) the quantized phase can achieve reliable speech reconstruction instead of the continuous phase; and (2) the phase of high-frequency components (>4 kHz) has minimal impact on speech quality, we propose three new mechanisms for the phase enhancement model, consisting of (1) phase quantization to reduce the estimation difficulty by transforming a regression problem to a classification problem; (2) frequency truncation during phase estimation to reduce the computation cost while maintaining the accuracy; (3) phase-nature-aware loss function, which is specifically designed to account for the ordinal and periodicity of phase.

T-F to T (Time) Domain Transform: After acquiring the enhanced magnitude and phase, this module combines them and uses inverse Short-time Fourier Transform (iSTFT) to recover the enhanced speech in the time domain.

Note that ClearSpeech performs different steps during training and inference, as indicated in the flowchart with solid and dashed arrows. Concretely, during real-world inference, no noise transformation and data synthesis is required. Instead, the noisy in-ear and out-ear speech data is directly processed by the SE modules to obtain the clean speech waveform in the time domain.

4 O2I SOUND TRANSFORMATION

4.1 Design Rationale

Collecting a large-scale in-ear noise dataset for SE model training requires tremendous labor and time overhead. To avoid this, we propose transforming existing out-ear noise datasets (captured with out-ear microphones) to their in-ear versions. This type of transformation builds upon the hypothesis that *there exists a correlation between the in-ear and out-ear sounds*, which is verified in Figure 2 (2nd row). Specifically, in-ear signals show similar patterns as their out-ear counterpart, even though they lack the high-frequency (>1 kHz) components.

4.2 O2I Model Design

To exploit this correlation, we propose a DL-based approach (light blue box in Figure 3). Particularly, the U-Net [31] architecture is used due to its great potential in extracting feature representations with the use of residual structure [17]. In particular, tt can transform the out-ear sound signals into a low-dimensional space to capture the correlated information with respect to the in-ear sound. This information is then further reconstructed into the corresponding in-ear sound representation. Next, we describe the model input and structure.

Model input (signal preprocessing): We collected a small-scale {out-ear, in-ear} paired dataset for O2I model training (explained in Section 7.2). The collected audio signals undergo three steps before feeding to the DL model for training: (1) the time series are segmented into one-second clips using the sliding window technique with 50% overlap; (2) each clip is converted to a T-F spectrogram using Short-time Fourier Transform (STFT). The parameters for STFT are n_fft = 512, hop_length = 160, $window\_length$ = 400, and *window='hann'*, resulting in a 2D vector of shape (257,101), where 257 is the number of frequency bins covering a frequency range of 0 to 8 kHz and 101 is the number of windows in one second; (3) the magnitude of the spectrogram is extracted and

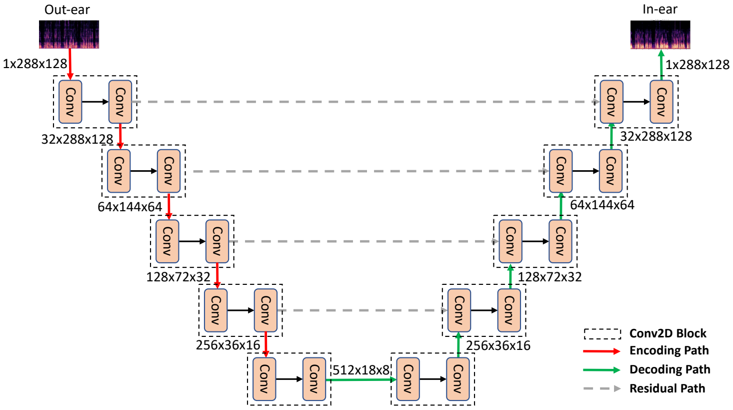


Fig. 4. U-Net based O2I model.

then normalized between 0-1. Note that we follow the same pre-processing steps for the magnitude and phase enhancement model presented in Section 5 and Section 6.

U-Net structure: Figure 4 illustrates the design details of the U-Net model. Specifically, it follows the auto-encoder architecture [20] and consists of an encoding and decoding path. Each step is a Conv2D block with two convolution layers using a kernel size of 3 × 3. Each convolution layer is followed by a Batch Normalization layer and a Rectified Linear Unit (Relu) [29] activation function. The number of channels at each Conv2D block doubles in the encoding path (with output size halves), and halves in the decoding path (with output size doubles), namely, [32, 64, 128, 256, 512] for the encoding path and [512, 256, 128, 64, 32] for the decoding path. Since the original magnitude spectrogram has dimensions of 257 × 101, both dimensions cannot be halved (as an integer) with a depth of 4. Thus, we extend the input size to 288 × 128 using zero padding. However, only the original 257 × 101 elements of the output are used when computing the loss.

5 MAGNITUDE ENHANCEMENT

With the pre-trained O2I sound transformation model, we now are able to convert the existing out-ear noise datasets to their in-ear versions, thereby creating the {out-ear, in-ear} paired noise datasets. This allows us to generate synthesized noisy speech for model development. ClearSpeech enhances the magnitude and phase of the speech separately and sequentially, and in this section, we focus on magnitude enhancement that has a dominant impact on the intelligibility of speech [36].

5.1 Model Overview

Figure 5 illustrates the proposed magnitude enhancement model. The model takes as input two streams of noisy magnitude spectrogram data (in-ear and out-ear) and outputs enhanced magnitude spectrogram data. It does this using three main stages: (1) coarse-grained feature extraction, which aims to extract high-level representations of the speech data, (2) fine-grained feature extraction and information exchange, which aims to collaboratively learn the mutual information between out-ear and in-ear signals in a fine-grained manner, and (3) in-ear & out-ear fusion, which aims to fuse the output from each stream and capture the temporal dynamics over time. This design is able to capture the complementary information in in-ear and out-ear speech data, and more importantly, is able to learn the salient information of in-ear and out-ear sounds in the time-frequency domain by taking into account their unique characteristics.

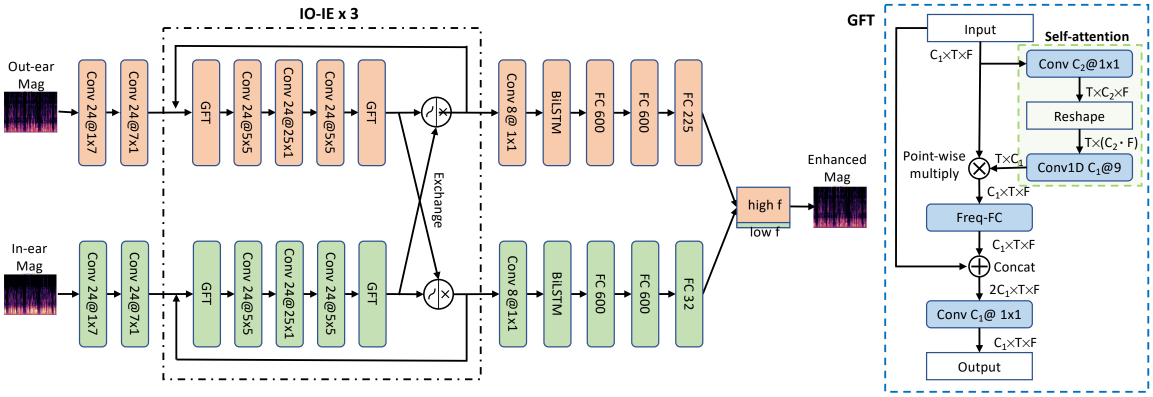


Fig. 5. Magnitude enhancement model.

5.2 Signal Pre-processing

With the created {out-ear, in-ear} paired noise datasets, we synthesize noisy in-ear and out-ear speech data by superimposing the noise to the corresponding clean speech signal. Next, the noisy speech data is segmented into one-second clips with an overlapping ratio of 50%. We then transform each noisy clip into the time-frequency domain using STFT, resulting in a magnitude spectrogram and a complex phase spectrogram. The two noisy magnitude spectrograms with a size of $F \times T$ ($F = 257, T = 101$) are the inputs to the magnitude enhancement model.

5.3 Coarse-grained Feature Extraction

Each of the input noisy spectrograms is first fed into two cascaded convolutional layers for high-level feature extraction. Specifically, instead of using 3×3 convolution kernels that are typically used in image classification to capture local features, we leverage two one-dimensional kernels (with a size of 1×7 and 7×1) to enhance the effective learning of long-term time and frequency correlations respectively. Moreover, the sequential combination of both enables precise and comprehensive learning across a wider range. Both of the convolution layers consist of 24 channels.

5.4 In-ear & Out-ear Information Exchange

Inspired by PHASEN's [42] design for extracting and exchanging information between magnitude and phase components, we propose an IO-IE block that (1) captures high-level local and global information for in-ear and out-ear speech data independently; and (2) enables mutual information exchange between in-ear and out-ear streams. The first module in our IO-IE is a Global Frequency Transmission (GFT) block, which is designed to learn the global information of the magnitude spectrogram (e.g., the harmonics and formants). The flowchart of the GFT is presented in the right part of Figure 5 and consists of three main steps. First, with the input feature, the self-attention module leverages 2D and 1D convolution layers to estimate an attention map, which is then multiplied with the input feature in a point-wise manner. The kernel size of the 1D convolution is $C_1 = 9$ and the channel size of the 2D convolution is $C_2 = 5$. Second, the resulted feature is passed to the Freq-FC layer that contains a trainable frequency transformation matrix (FTM) to capture the information from all frequency bands. Third, the output of Freq-FC and the input feature are concatenated based on the residual mechanism and fused with a 1×1 convolution.

The second module in the IO-IE block stacks the three convolution layers used to capture the local information. Specifically, a small kernel size of 5×5 is used the first and third convolution layers with a kernel size of 25×1

for the second convolution layer to capture long-range time-domain correlation as speech and noise can exhibit distinct characteristics in the time domain due to their temporal patterns, envelope fluctuations, etc. Finally, another GFT module is used after the second module to elastically learn the higher level representations of local and global information. These two modules (second module and GFT) are applied independently on the in-ear and out-ear speech data for fine-grained feature extraction.

The mutual information exchange is realized by the third module - a gate mechanism. The underlying rationale is that in-ear and out-ear signals actually capture the same source (human speech) but experience distinct propagation channels, so there exists shared information. To extract and exchange the information, we propose a gate mechanism, which performs the following operations:

$$f(x_1, x_2) = x_1 \circ Tanh(x_2), \tag{1}$$

where $\circ$ denotes element-wise multiplication. (x_1=in-ear, x_2=out-ear) for the in-ear stream, and vice versa for the out-ear stream. The purpose of the $Tanh$ operation is to restrict the latter term within [-1, 1], i.e., applying a mask on the former term. The IO-IE block is repeated three times to allow deeper in-ear & out-ear information exchange with multiple rounds of feature extraction. Note that each 2D convolution layer is followed by batch normalization (BN) and activation function ReLU.

5.5 In-ear & Out-ear Fusion

After the collaborative fine-grained feature extraction, we utilize a BiLSTM layer to further capture the continuous nature of speech signals by learning the temporal correlations between T-F bins [10, 36, 42] by using three fully connected (FC) layers to estimate the magnitude spectrogram for each stream. In addition, a 1×1 convolution layer is applied before the BiLSTM layer to reduce the channel size from 24 to 8, thereby reducing the computation of the BiLSTM layer. The size of the hidden state in the BiLSTM layer is set to 100. The number of neurons in the first two FC layers is 600 and the activation function is ReLU.

For the final in-ear & out-ear fusion, we propose a novel learning strategy to accommodate the different characteristics of the two signals. Concretely, based on Observation 1, we propose to estimate the low-frequency and high-frequency portions of the magnitude spectrogram as separate in-ear and out-ear streams respectively. In particular, the in-ear stream will only reconstruct the frequency components within the 0-1 kHz range using the last FC layer of 32 neurons, while the out-ear stream will reconstruct the frequency components within the 1-8 kHz range using the last FC layer of 225 neurons. Finally, we will concatenate both streams in the frequency dimension to form the full enhanced magnitude spectrogram.

5.6 Magnitude Loss Function

By analyzing the characteristics of the in-ear and out-ear speech data, we discovered another observation.

Observation 2*: due to the decrease of salient information in 1-4 kHz for in-ear speech, overall, the in-ear and out-ear inputs provide non-equivalent speech information across different frequency ranges.*

To validate Observation 2, we calculate the energy distribution of in-ear and out-ear speech across different frequencies. Specifically, we apply the STFT to the in-ear and out-ear speech respectively and average the energy values along the time axis in the resulting magnitude spectrogram. As illustrated in Figure 6, the results show that although in-ear speech still contains some information between 1-2.5 kHz, the amount of energy of in-ear speech is much less than that of out-ear speech, translating to the reduction of information. Thus, instead of applying the loss function on the estimated magnitude spectrogram uniformly [36, 42], we designed a customized loss function that incorporates the above observation to enhance the learning of the medium frequency of 1-4 kHz. Concretely,

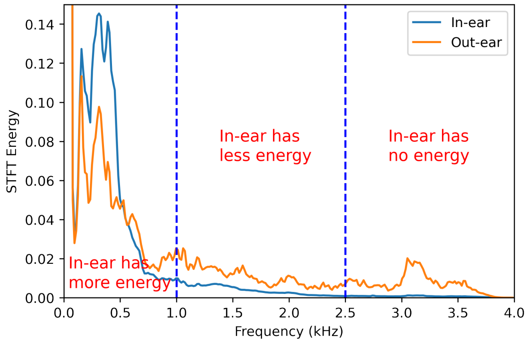


Fig. 6. Energy distribution of in-ear and out-ear speech across different frequencies.

we split the 0-8 kHz[5] frequency range into three bands: $f_l \in [0, 1]$ kHz where both in-ear and out-ear speech data contains abundant information, $f_m \in [1, 4]$ kHz where out-ear speech contains much more information than in-ear speech, and $f_h \in [4, 8]$ kHz where both signals contain very limited speech information and only a few harmonics exist (as typical human speech is below 4 kHz [30]). Then, we apply different weights to the Mean Square Error (MSE) loss of the three bands. Based on Observation 2, we should apply a larger weight to the loss L_{f_m} to force the model to pay more attention to recovering the middle-frequency range as it forms the majority of speech frequency range while the inputs contain less information within it. Similarly, the smallest weight should be assigned to L_{f_h} as there is only little speech information[6] in the high-frequency region. The final loss function is a weighted combination of the three losses:

$$L = \alpha_1 L_{f_l} + \alpha_2 L_{f_m} + (1 - \alpha_1 - \alpha_2) L_{f_h} \tag{2}$$

$$L_{f_*} = MSE(\hat{y}_{f_*}, y_{f_*}) \tag{3}$$

where α_* represents the weights for each band, and L_{f_*} represents the MSE loss between predicted enhanced spectrogram $\hat{y}_{f_*}$ and true clean spectrogram y_{f_*} for subband f_*. Empirically, we found that $\alpha_1 = 0.2, \alpha_2 = 0.7$ achieves the optimal SE performance.

6 PHASE ENHANCEMENT

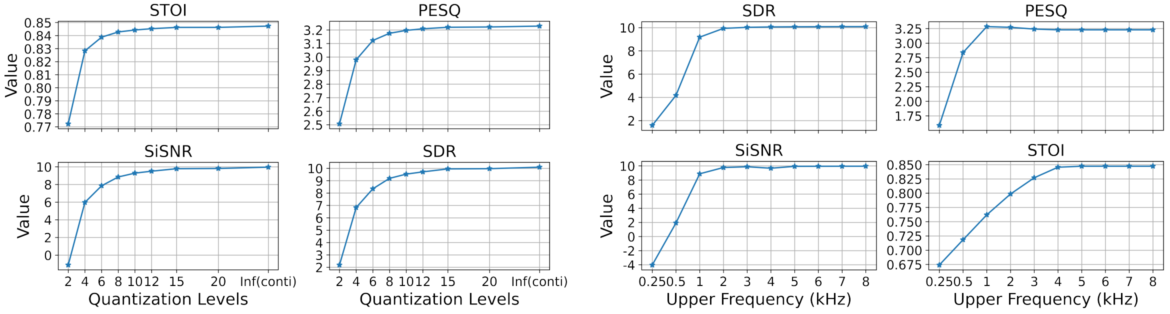
Phase is an important factor affecting speech quality [36]. But due to the lack of structural patterns, denoising phase from the noisy speech is extremely challenging. In this section, we first analyze the impact of different phase-related factors on speech reconstruction using numerical experiments, and then propose a novel phase estimation mechanism consisting of three components: phase quantization, frequency truncation, and phase-nature-aware loss function.

6.1 Impact of Phase Quantization and Truncation

6.1.1 Continuous Phase vs. Quantized Phase. Phase is defined as the offset of a wave from a given point. It has periodicity and its value can be depicted in radians ranging from -π to π continuously. However, is a continuous phase really necessary for SE? To answer this question, we compare the SE performance with continuous phase and quantized phase prediction. Specifically, we apply different quantization levels (from 2-20) to generate discrete phase values. For a quantization level of N, we divide [-π, π] to N segments uniformly. The median value of each

[5] Typical speech recognition and enhancement techniques are applied on speech with a 16 kHz sampling rate. Therefore, the frequency range of the speech is 0-8 kHz based on the Nyquist sampling theory.

[6] However, we cannot assign zero weight to L_{f_h} as the model would learn randomly for the high-frequency band and therefore affect the SE performance.



(a)

(b)

Fig. 7. (a) Impact of continuous and quantized phase. (b) Impact of phase at different frequencies.

segment is regarded as the quantized phase, and any phase that falls within this segment will be replaced with the quantized phase. For example, 4 quantization levels result in 4 segments [-π, -$\pi/2$], [-$\pi/2$, 0], [0, $\pi/2$], and [$\pi/2$, π], and the corresponding quantized phases are -3$\pi/4$, -$\pi/4$, $\pi/4$, and 3$\pi/4$. A phase of value $\pi/8$ falls within [0, $\pi/2$] and therefore would be quantized to $\pi/4$. Using the enhanced magnitude from the model developed in Section 5 and the ground truth phase, we evaluate and compare the speech reconstruction performance with continuous and quantized phases respectively. Four commonly adopted metrics are used to quantify the reconstructed speech quality (as described in Section 8.1.)

The results are presented in Figure 7(a), where Inf on the X axis refers to the continuous phase and higher Y-axis values indicate better quality of the reconstructed speech. We can see that with the increase in the quantization levels, all four metrics show improvement. A plateau is observed when the quantization level is above 15 because the misalignment of phase is sufficiently small so that its impact on speech quality and intelligibility is negligible. The plateau extends to Inf, which indicates that the reconstruction performance using quantized phase is comparable to that of the continuous phase. Thus we derive the following observation, which would be useful to guide our phase enhancement model design.

Observation 3*: with a sufficient quantization level (e.g., ≥15), the impact of using discrete phase on speech reconstruction is negligible compared to using continuous phase.*

Observation 3 also agrees with the common practice in digital communications, where the decoding (e.g., 16-QAM) is based on quantized phases [34].

6.1.2 Low-frequency Components vs. High-frequency Components of Phase. For a T-F spectrogram of size (257, 101), we can obtain its phase spectrogram of the same shape, where each value is the phase at a certain frequency at a given time. Here, we investigate how the phase at different frequency ranges affects the SE performance. To do this, we define a masked phase spectrum, by replacing the phase values above frequency f_u (i.e., [f_u, 8 kHz]) with a constant value (e.g., 0, $\pi/2$, or any) over time. Then, we utilize the masked phase to reconstruct the speech and compute the four SE metrics for the different upper frequencies f_u.

As shown in Figure 7(b), all the four metrics remain stable when upper frequency f_u decreases from 8 kHz, until they drop sharply when f_u is lower than a certain threshold (2 kHz for the first three metrics and 4 kHz for the last metric). From this, we obtain another observation.

Observation 4*: Only the phases of the low-frequency components (<4 kHz) of the phase spectrogram are critical to SE.*

The rationale behind Observation 4 is that high-frequency components of human speech have near-zero energy on the magnitude spectrogram. Thus, changes in the high-frequency phase will not affect the enhanced speech.

6.2 U-Net based Phase Classification

Design Rationale: Using Observation 3, we proposed to estimate the quantized phase instead of the continuous one. As such, we transform a *regression* problem to a *classification* problem, which is simpler intuitively. The quantization level is set to 15 according to Figure 7(a). Using Observation 4, we proposed to estimate the truncated phase only, i.e., for frequencies below 4 kHz. As such, we can further reduce the complexity of phase classification and also save memory as well as reduce computation/latency.

Model Architecture: To estimate the quantized phase, the model must classify each element in the phase spectrogram simultaneously. This is a similar problem to semantic segmentation [39] in computer vision, where each phase element corresponds to a pixel of the image. Thus, we adopt the mainstream semantic segmentation network, U-Net [31], for phase estimation.

Specifically, we utilize the enhanced magnitude and noisy out-ear phase (stacked as different channels of the convolution layer) to predict a clean quantized phase [44]. In particular, the input noisy phase is represented as complex values and their real parts and imaginary parts are regarded as different channels. The reason for using complex phases instead of real phases is to avoid the effect of phase wrapping [42, 44]. The output quantized phase is an integer between 0-14, and will be converted to a phase in radians between -π to π. The architecture of the phase model is almost the same as the O2I model shown in Figure 4 except for two differences: (1) the input size changes from $3 \times 288 \times 128$ to $3 \times 128 \times 128$ and the output size changes from $3 \times 288 \times 128$ to $1 \times 128 \times 128$; (2) at the end of the decoding path, the phase model adds a Softmax layer to convert the output of each class to 0-1 and then assigns the index of the maximal value as the quantized phase label (0-14). The number of channels for the convolution blocks in the encoding path is [32, 64, 128, 256, 512] and [512, 256, 128, 64, 32] for the decoding path. Each 2D convolution layer is followed by a Batch Normalization layer and a Rectified Linear Unit (Relu) activation function.

6.3 Phase-nature-aware Loss Function

Phase has two properties that impact the loss function: (1) quantized phase is ordinal with the ranking order, and misclassification of the quantized levels should take into account the ordinal nature. For example, given a ground truth phase of level 3, predicting it to level 4 and level 8 incurs different levels of misclassification – which is is not considered in the widely used Cross Entropy (CE) loss; (2) phase is periodic with a period of 2π, meaning that predicting to -π and π is identical, and -π to $3\pi/4$ is closer than -π to -$\pi/4$. However, the widely used traditional CE loss for classification can not capture these characteristics. Thus, we propose a new loss function called Ordinal and Periodicity aware Cross Entropy (OPCE) loss L_{opce}, defined as follows:

$$L_{opce} = (1 + \beta) L_{CE}(\hat{q}, q) \tag{4}$$

$$\beta = min(|\hat{q} - q|, Q - 1 - |\hat{q} - q|) \tag{5}$$

where L_{CE} represents the standard Cross Entropy loss between predicted phase labels $\hat{q}$ and true labels q. β refers to a scaling factor (the extent of misclassification) that accounts for the ordinal nature. When computing β, we use the *min* operation to incorporate phase periodicity, where Q is the quantization level.

7 PROTOTYPE AND DATA COLLECTION

In this section, we present the design of the earbuds prototype, the data collection and data synthesis procedures, as well as the model training settings.

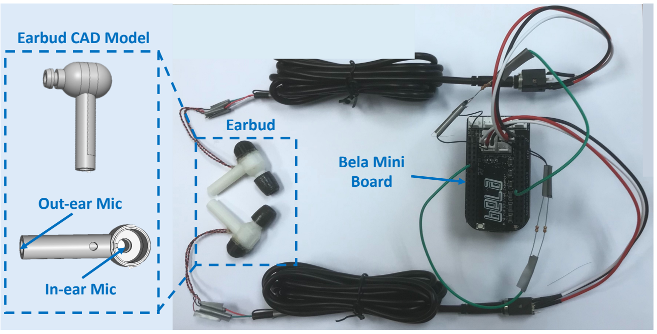


Fig. 8. The designed earbuds prototype and data recording board.

7.1 Earbuds Prototype

Although many commercial earbuds, such as Airpods Pro and Honor Earbuds 3, have already integrated an in-ear microphone for ANC, there is no API available to access the raw signals heard by the microphone. Thus, we designed and built a pair of earbuds to evaluate the proposed system. As shown in Figure 8, we designed a CAD model and fabricated a 3D printed earbud case. Specifically, following the arrangement of microphones on commercial earbuds, one microphone was embedded inside the earbud case and faced inward to capture in-ear speech (*in-ear mic*), with another microphone integrated at the end of the handle and faced downward to pick up out-ear speech (*out-ear mic*). Both microphones are analog microphones from CUI Devices (CMC-4015-40L100 [2]) and are similar to commercial earbuds. The microphones are connected to a Bela Mini Board [1] using 3.5mm audio jacks. The Bela Mini Board records the two audio signals simultaneously at 44.1 kHz using a browser-based IDE, and then forwards the data to a laptop for processing. A foam ear tip is integrated to ensure user comfort and sealing quality.

7.2 Data Collection

With the designed earbuds prototype, we collected two datasets for our evaluation. Specifically, the *External Audio Dataset* is a small-scale dataset used to learn the mapping between in-ear and out-ear noise (i.e., for the O2I model training). With the trained model, we then generate the in-ear version of existing public noise datasets to create {in-ear, out-ear} paired noise datasets. The *Speech dataset* contains clean speech recorded with both the in-ear and out-ear microphones simultaneously, which is used to (1) synthesize noisy in-ear and out-ear speech, and (2) establish the ground truth speech (out-ear speech only) to supervise the SE models training. 20 subjects (fluent English speakers comprising of 12 males and 8 females between 18 and 28 years of age) were recruited for the IRB-approved data collection.

External Audio Dataset: We selected 9 different types of noise[7] that cover a wide range of frequencies. This allowed the O2I model to learn the in-ear and out-ear correlations at any frequency. The subjects wore the earbuds and were asked to remain still while a laptop played each sound at a reasonable volume. At the same time, the in-ear and out-ear signals were recorded. Each sound was played for 2 minutes with 2 minutes × 9 × 20 = 360 minutes, in total, was recorded.

Speech Dataset: We selected 300 sentences from the TIMIT corpus [14] and displayed them on a screen in front of the subject. The subject was asked to remain still and read the sentences sequentially, while the in-ear

[7] Male song, female song, air conditioning, piano, traffic noise, restaurant background noise, ocean wave sounds, male speech, and female speech.

and ou-ear signals were recorded. The average length of each sentence, across all subjects, was 2.82 seconds. In total, 2.82 seconds × 300 × 20 × 2 ≈ 9.4 hours of clean speech were collected.

7.3 Data Synthesis and Model Training

To create noisy in-ear and out-ear speech for model training, we selected 10 different types of common sounds[8] from AudioSet [15], which allows the magnitude and phase model to capture the characteristics of real-world noise, and added them to the collected speech signals. Specifically, to synthesize noisy out-ear speech, the public noise was directly added to the collected clean out-ear speech (balanced among different noise types). To synthesize noisy in-ear speech, we first converted the same public noise to its in-ear version using the O2I model, and then added it to the collected clean in-ear speech. The resulting SNR for the noisy samples during training was uniformly distributed between -10 to 10 dB. The signals were segmented into 1-second clips with a 50% overlapping ratio, resulting in 34,005 samples for evaluation.

We split the 20 subjects into three groups, namely, 14 (7 males + 7 females) for training (23,729 samples), 2 (1 male + 1 female) for validation (3,411 samples), and 4 (2 males + 2 females) for testing (6,865 samples)[9]. All the models were trained for 50 epochs with the Adam optimizer. The learning rate was set to 0.001 for the first 40 epochs and 0.0001 for the remaining 10 epochs. The model with the lowest validation loss within the 50 epochs was saved as the trained model. The loss function for the O2I model was mean square root (MSE) loss, while the magnitude and phase enhancement model utilized the customized loss function as presented in previous sections. All the models were implemented in PyTorch and trained with a GeForce GTX 1080 Ti GPU.

8 PERFORMANCE EVALUATION

In this section, we first evaluate the performance of the O2I module and the overall SE framework. Then, we assess the impact of different parameters and conditions, before presenting the system performance.

8.1 Metrics and Baselines

We consider four typical SE metrics in the evaluation. Since our target is to recover the clean out-ear speech, we use it as the reference signal to calculate these metrics.

- **STOI**: short-time objective intelligibility, which measures the intelligibility of speech. Its values range from 0 to 1, with higher values implying better intelligibility.
- **PESQ**: perceptual evaluation of speech quality. It is an objective measure of speech quality considering multiple audio characteristics such as audio sharpness, variable latency, clipping and etc. Its values range from 0 to 5, with a higher value indicating better quality.
- **SiSNR**: scale-invariant signal-to-noise ratio. Compared to conventional SNR, the signals are normalized to a mean value of zero to ensure scale-invariance.
- **SDR**: signal-to-distortion ratio, where the distortion refers to unwanted signals correlated with the speech.

To compare the performance of ClearSpeech with state-of-the-art techniques, we also consider diverse deep learning based baselines including T-F domain, time domain, and speech separation approaches. The following briefly describes the main idea and technology proposed in each baseline.

- PHASEN [42]: it is a T-F domain SE framework that estimates the clean magnitude and phase concurrently using the noisy out-ear spectrogram. The core of PHASEN is a two-stream network that only learns a

[8] [Male speech, man speaking], [Female speech, woman speaking], [Rhythm and blues], [Pop music], [Motorcycle], [Aircraft engine], [Traffic noise, roadway noise], [Dishes, pots, and pans], [Environmental noise], [Wind noise].

[9] Note that although the training, validation, and testing are conducted on the same speech content, we introduced diverse simulated noises during the training and testing stages to ensure that the model can effectively handle varying and previously unseen noisy conditions. Moreover, the frequency characteristics of the same speech from different subjects are also distinct due to the unique individual's voiceprint.

complex ideal ratio mask with special awareness of the phase and harmonics of the clean speech. Specifically, an information exchange mechanism is designed so that the two streams can share mutual information to facilitate phase estimation, and a frequency transformation block that can catch long-range correlations along the frequency axis is designed to learn the harmonic correlation.

- FullSubNet+ [9]: it is a three-branch SE framework that takes T-F domain features (i.e., magnitude spectrorgam, real spectrorgam, and imaginary spectrorgam of noisy speech for each branch) as inputs to learn the complex ideal ratio mask. Each branch consists of three stages. First, a lightweight multi-scale time-sensitive channel attention module that adopts multi-scale convolution and channel attention mechanism is designed to help the subsequent networks focus on more discriminative frequency bands for noise reduction. Then, a full-band model that takes the whole spectrogram as input is connected to capture the global spectral context and the long-distance crossband dependencies. To learn the signal stationarity and local spectral pattern, a sub-band model that processes each frequency independently is stacked thereafter.
- InterSubNet [8]: it is a recent SE framework that works on T-F domain features. To reduce the complexity and number of parameters of the full-band model employed in FullSubNet+, InterSubNet only utilizes the sub-band model to reduce the computation overhead. To overcome the limitation of the sub-band model, i.e., lack of global spectral information, InterSubNet designs a sub-band interaction module to capture the cross-band dependencies and global spectral patterns. In detail, instead of processing the parallel sub-band units independently, a two-LC (linear layer) based mapping function is applied across multiple sub-band units.
- Conv-Tasnet [22]: it is a time-domain speech separation framework that split the speech of multiple speakers from the mixed audio. It can be also used for SE by regarding speech and noise as audio from different speakers. Conv-Tasnet adopts an auto-encoder structure, where a linear encoder is designed to generate a representation of the speech waveform and a linear decoder is used to transform the separated speech data back into the waveforms. In between the encoder and decoder, a separation module, which consists of 1-D dilated convolutional blocks to model the long-term dependencies of the speech signal, applies a set of weighting functions to the encoder outputs for speaker separation.
- DPRNN [21]: it is also a time-domain speech separation framework with optimizations upon the Conv-Tasnet model. Concretely, the 1-D dilated convolutional block used in Conv-Tasnet cannot perform utterance-level sequence modeling when its receptive field is smaller than the sequence length. Thus, DPRNN (dual-path recurrent neural network) splits the long sequential input into smaller chunks and designs two RNNs, an intra-chunk RNN and an inter-chunk RNN, for local and global modeling, respectively. In detail, the intra-chunk RNN will be applied first to process the local chunks independently, after which the output information from all the chunks will be aggregated by the inter-chunk RNN to perform utterance-level processing.

8.2 O2I Sound Transformation

The O2I sound transformation module plays a critical role in the proposed pipeline as the performance of the O2I noise conversion directly affects the quality of the synthesized in-ear data and thus the subsequent SE performance. Figure 9 presents the spectrograms of the original out-ear signal, in-ear signal, and the converted in-ear signal, and provides a visual validation of the transformation performance. We can observe that the converted signal is almost identical to the original in-ear signal, indicating that the O2I model successfully captures the correlation between the out-ear and in-ear signals (i.e., suppressing the high-frequency components while amplifying the low-frequency parts). Moreover, using the phase of the in-ear signal, we reconstructed the time domain signal and calculated the four SE metrics using the original in-ear signal as a reference. The computed STOI, PESQ,

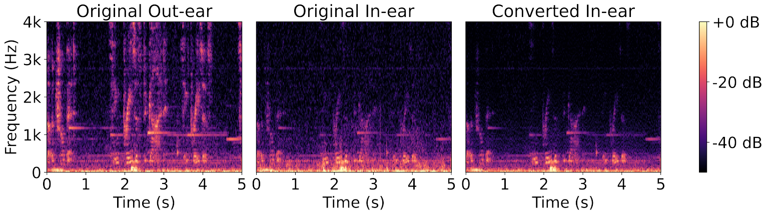


Fig. 9. O2I model performance. The 1st spectrogram is tranformed to the 3rd spectrogram with the O2I model, which is almost identical to the 2nd spectrogram.

Table 1. SE performance vs. baselines and variants.

Baselines/Variants	**Year**	**Feature Domain**	**STOI**		**PESQ**		**SiSNR**		**SDR**	
			Mag Only	Mag+ Phase	Mag Only	Mag+ Phase	Mag Only	Mag+ Phase	Mag Only	Mag+ Phase
Conv-Tasnet [22]	2019	Time	-	0.71	-	2.36	-	8.28	-	4.44
DPRNN [21]	2020	Time	-	0.71	-	2.37	-	**8.35**	-	5.27
PHASEN [42]	2020	T-F	-	0.69	-	2.26	-	1.36	-	3.64
FullSubNet+ [9]	2022	T-F	-	0.73	-	2.39	-	1.95	-	4.12
Inter-SubNet [8]	2023	T-F	-	0.73	-	2.41	-	2.16	-	4.42
In Only	-	T-F	0.61	0.61	1.96	1.95	-22.66	-22.01	-1.13	-1.10
Out Only	-	T-F	0.69	0.69	2.38	2.42	3.99	4.27	5.38	5.61
In+Out(ClearSpeech)	-	T-F	0.79	**0.79**	2.84	**2.89**	6.22	6.42	7.00	**7.16**

SiSNR, and SDR for the entire test set are 0.831, 3.666, 7.202, and 7.829 respectively[10], further suggesting the excellent performance of our U-Net based O2I model. Consequently, we can eliminate the burden of collecting a new dataset and reuse the existing ones instead.

8.3 Overall SE Performance

Table 1 presents a performance comparison of the proposed ClearSpeech *In+Out* with its variants (*In Only*, *Out Only*) and some recent SE baselines introduced in Section 8.1 [8, 9, 21, 22, 42]. For *In Only*, the model only contains the stream for the in-ear signal (i.e., the green block in Figure 5), without the out-ear branch and the exchange between in-ear and out-ear signal. Similarly, *Out Only* model only contains the orange stream in Figure 5. In addition, the final FC layer of each model contains 256 neurons as it recovers all the frequency bands. For the phase enhancement model of *In only*, the input is the noisy phase of the in-ear signal, instead of the out-ear signal.

The SNR of the testing samples follows the same distribution as the training samples (i.e., -10 to 10 dB). Since ClearSpeech enhances magnitude and phase sequentially, we present two values - *Mag Only* and *Mag+Phase* - for each metric, from which we can assess how the magnitude and phase enhancement models contribute to the overall SE performance. For *Mag Only*, we reuse the noisy out-ear phase to reconstruct the time domain speech. For PHASEN, only *Mag+Phase* is provided as it estimates magnitude and phase concurrently. The baselines are implemented based on the authors' GitHub releases. All the variants and baselines were trained and tested with the same data for a fair comparison.

[10] These values are considered as excellent performance for speech similarity [36, 42].

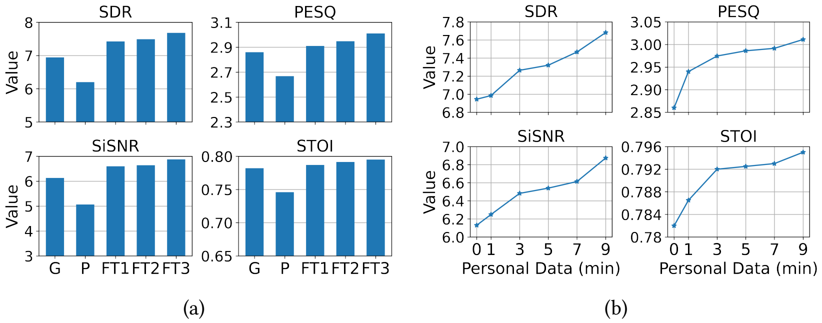


Fig. 10. (a) Performance when adding personal data (G means generally pre-trained model, P means model trained with a single user data from scratch, FT*n* (n=1,2,3) show results after fine-tuning the last n linear layers. (b) Performance when adding different amounts of personal data using FT3.

First, for baseline comparison, ClearSpeech significantly outperforms all the baselines due to two reasons: (1) the use of in-ear speech provides additional information compared to the use of out-ear signals only; and (2) the proposed techniques (e.g., information exchange, high-low frequency separation, and phase quantization) effectively utilize this additional information. In detail, existing T-F domain approaches (PHASEN, FullSubNet+, and Inter-SubNet) yield slightly better performance than the time domain based approaches (Conv-Tasnet and DPRNN) in terms of STOI and PESQ. However, Conv-Tasnet and DPRNN achieve the best performance (even higher than ClearSpeech) in terms of SiSNR. This is because SiSNR is used as the loss function for model training in the two approaches and therefore this metric is particularly optimized. Second, for the comparison with variants, both variants (*In Only* and *Out Only*) show poorer performance than ClearSpeech as less information is used by the variants. The *In Only* variant yields the worst performance as in-ear speech is frequency-distorted and contains limited speech information compared to out-ear speech. Third, comparing the *Mag Only* and *Mag+Phase* models, we observe that the main SE gain comes from the magnitude enhancement model, with the phase enhancement model further improving the performance by correcting the phase shifts caused by noise.

8.4 Model Personalization

As mentioned in Section 7.3, we split the 20 subjects into 14 for training, 2 for validation, and 4 for testing. Thus, the above results shown so far were obtained using a universal model trained with other people's speech. Generally, adding personal data for model fine-tuning (known as personalized speech enhancement) would improve the SE performance as personal data provides additional cues about the speaker (such as more details about the structure of their vocal system etc.) that allows for better extraction of the clean speech [11]. Moreover, the use of the in-ear microphone in our system provides another personal attribute to improve performance. Specifically, due to the unique geometry of each individual's ear canal, the correlation between the out-ear and in-ear speech is slightly different between people [4].

To investigate the possible improvement when using personalized SE models, we conducted two additional experiments: (1) we fine-tuned the universal model by freezing the feature extraction layers of the magnitude enhancement model and re-training the final linear layers; (2) we trained a model from scratch with a single user's data (for each of the four testing subjects), referred to as *P* in Figure 10(a). Since our magnitude enhancement model contains three linear layers, we consider three variants during the fine-tuning, namely, *FT1* - tune the last linear layer only, *FT2* - tune the last two linear layers, and *FT3* - tune all three linear layers. As shown in Figure 10(a), personalized models indeed improve the SE performance, and re-training all three linear layers (i.e.,

Table 2. SE performance vs. SNR.

Metric		**SNR (dB)**				
		-10	-5	0	5	10
STOI	Noisy	0.34	0.43	0.54	0.64	0.71
	Out	0.50	0.62	0.71	0.77	0.82
	In+Out	0.66	0.74	0.80	0.84	0.88
	Gain(%)	100	63	53	54	55
PESQ	Noisy	1.10	1.28	1.62	2.02	2.29
	Out	1.74	2.13	2.51	2.84	3.07
	In+Out	2.22	2.63	2.98	3.27	3.50
	Gain(%)	*75*	*59*	*53*	*52*	*55*
SiSNR	Noisy	-6.78	-5.29	-1.31	0.76	1.22
	Out	-1.47	2.13	5.56	8.04	9.53
	In+Out	0.89	4.67	7.56	10.08	11.94
	Gain(%)	44	34	29	28	29
SDR	Noisy	-3.78	-2.79	-0.21	1.13	1.71
	Out	1.90	4.18	6.62	8.62	9.89
	In+Out	3.39	5.67	8.06	10.09	11.57
	Gain(%)	26	21	21	20	21

FT3) yields the best performance. Note that because we collected limited data from each individual, we found that developing fully personalized models from scratch (i.e., *P*) resulted in even worse performance compared to the universal models trained with a much large volume of data.

Finally, with the *FT3* variant, we investigated the impact of the amount of personal data by adding different amounts of personal data ranging in length from 1 minute to 9 minutes. From Figure 10(b), we observe that the SE performance continuously improves with more personal data and with 3 minutes of personal data, the SE performance can be improved by around 5%.

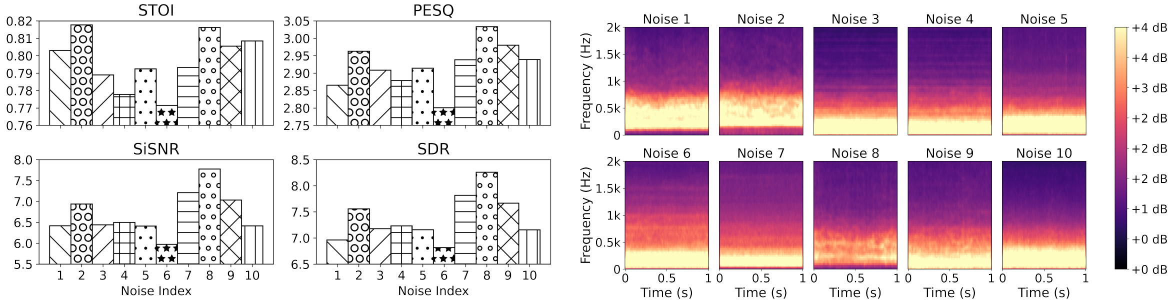
8.5 Impact of SNR

As discussed in Section 2.2, due to the occlusion effect, the in-ear microphone is more resilient to external noise compared to the out-ear microphone. Thus, we would expect more performance gain from using in-ear speech when the SNR of out-ear speech is low. To validate this, we test ClearSpeech at different SNRs, i.e., all the testing samples are synthesized to a particular SNR (e.g., -10 dB, -5 dB, 0 dB, 5 dB, and 10 dB) by controlling the intensity of the noise. Note that the overall performance presented in Table 1 was obtained by setting the SNRs of the testing data to be normally distributed within the range of -10 dB to 10 dB.

Table 2 presents the performance for *Noisy Input*, *Out Only*, and *In+Out*, where the gain is computed as *((In+Out)-(Out Only))/((Out Only)-(Noisy Input))*×100%. We observe that (1) with the increase in SNR, the quality of the enhanced speech improves; (2) at lower SNR scenarios, the in-ear microphone can offer more gain. Thus, ClearSpeech shows great potential in enhancing speech under a variety of noise conditions.

8.6 Impact of Noise Type

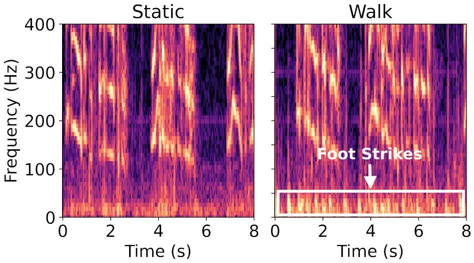
As mentioned in Section 7.3, we selected 10 common noise signals for SE model training and evaluation. However, due to the frequency characteristics of in-ear speech (e.g., amplifying low-frequency sounds), we are uncertain whether it offers equivalent improvement across different noisy characteristics (e.g., high-frequency,



(a)

(b)

Fig. 11. (a) SE performance vs. noise type. (b) Average spectrograms of different noise types.



Metric	Static	Walk
STOI	0.78	0.80
PESQ	3.02	2.95
SiSNR	7.53	6.76
SDR	5.14	4.60

Fig. 12. SE performance vs. human motion.

low-frequency, wide-band, narrow-band). Figure 11(a) compares the SE performance of the 10 noise types [11], and we observe that there are indeed performance variations across different types of noise. To further investigate the root cause of such variation, Figure 11(b) plots the average of spectrograms across all test samples for each of the 10 noise signals. We observe that Noise 8 (Dishes, pots, and pans) manifests lower energy within the frequency band of speech and has a smaller impact on SE. SE with Noise 8 is thus an easier task and achieves the best SE performance among the 10 noise types. On the contrary, Noise 6 (Aircraft engine) exhibits high energy within a wider band in the low-frequency range that highly overlaps with human speech frequency, and thus has relatively poorer SE performance.

8.7 Impact of Human Motion

In addition to human speech, the in-ear microphone, as demonstrated in [12, 23, 24], is able to capture other human-generated sounds (such as heartbeats, finger taps on the face, foot strikes, etc.) that are propagated to the ear canal via bone conduction. To investigate the impact of these human-generated sounds on SE performance, we recruited one subject to read sentences from the TIMIT corpus while they were walking (the strongest interference) and recorded both the in-ear and out-ear microphone data. The left part of Figure 12 compares the spectrograms of in-ear speech collected when the subject was stationary and while walking. We observe that human speech is above 100 Hz and there is a clear foot strike signal (one spike per step as annotated in the white

[11] The 10 noise types and associated indexes are: 1-[Male speech, man speaking], 2-[Female speech, woman speaking], 3-[Rhythm and blues], 4-[Pop music], 5-[Motorcycle], 6-[Aircraft engine], 7-[Traffic noise, roadway noise], 8-[Dishes, pots, and pans], 9-[Environmental noise], 10-[Wind noise].

Table 3. Latency of ClearSpeech on Desktop and Pixel 3.

Device	Model	Pre-P	Mag	Pha	Post-P	Total
Desktop	FP32	4.9 ms	19.9 ms	3.5 ms	7.7 ms	36 ms
Pixel 3	FP32	8.9 ms	797 ms	229 ms	11.2 ms	1046.1 ms
	INT8	8.9 ms	506 ms	99 ms	11.2 ms	625.1 ms

rectangle) at below 50 Hz from the walking spectrogram, which does not exist in the static spectrogram. This indicates that walking indeed introduces additional noise to the in-ear microphone signal. Then, we processed both the in-ear and out-ear data and fed them into the pre-trained SE models. As shown in the table, compared to the static scenario, walking just leads to a slight decrease in SE performance. This is because the majority of human speech is above 100 Hz so the SE models mainly learn features at those frequencies, implying that human motion artifacts have a limited impact on ClearSpeech. Since the sound produced by heartbeats is much weaker than that of foot strikes and heartbeat always exists even in static scenarios, its impact on SE performance would be negligible.

8.8 In-the-wild Study

All the above evaluations were performed using synthesized data. To validate the effectiveness of ClearSpeech in real-world scenarios, we conducted an in-the-wild study. Specifically, we recorded the noisy in-ear and out-ear speech under three conditions: (1) office with background music; (2) office with background people chatting; and (3) crossroads with busy traffic. Since the clean speech baseline is unavailable, we cannot objectively measure our system with the four SE metrics. Thus, we invited 14 subjects for an IRB-approved subjective assessment. Specifically, the subjects were instructed to listen to the noisy speech (as a reference) and enhanced speech (produced by ClearSpeech) and rate the enhancement quality of ClearSpeech using a 5-point scale (1 - poor, 2 - fair, 3 - good, 4 - very good, 5 - excellent). The average scores for the three scenarios were 3.83, 3.51, and 4.20, respectively, demonstrating the effectiveness of ClearSpeech across different real-world scenarios (note that even the lowest value of 3.51 is perceived between good and very good). Looking deeper, condition 2 - office with background people chatting yielded a relatively worse performance as the noise is spontaneous and similar to the intended signal (speech). On the other hand, traffic noise, while being loud, has a relatively flat energy distribution across all frequencies, and thus achieves the best performance. Moreover, the results also suggest that the quality of our synthesized data (derived from the O2I model) is close to real data distributions.

8.9 Latency

We evaluated the latency of ClearSpeech on two platforms: a desktop with a GPU (GeForce GTX 1080 Ti) and a smartphone (Google Pixel 3). We decomposed the whole SE pipeline into 4 stages: pre-processing (spectrograms extraction, phase quantization, etc.), magnitude, phase enhancement (model inference), and post-processing (phase mapping, T-domain transformation, etc.). As shown in Table 3, the desktop only takes 36 ms to process a one-second sample, and could act as a cloud-based SE system when the network is good or to process large-scale offline audio recordings. However, when running ClearSpeech locally on a smartphone [12], the inference time of the magnitude and phase enhancement models increases by 40-60× due to the lower computation capability on smartphones, lifting the overall processing time to 1046 ms. Leveraging recent achievements in efficient model execution on mobile devices [3], we first convert the model weights from floating-point (FP32) to integer (INT8)

[12] Note when making calls or issuing voice commands, the speech collected by the earbud has to be offloaded to a smartphone anyway before sending to the recipient.

by applying quantization and then fine-tune the INT8 model with a small learning rate of 0.0001 for 50 epochs to alleviate the impact of the reduced weight resolution. Compared to the FP32 model (STIO=0.79, PESQ=2.89, SiSNR=6.42, SDR=7.16), the INT8 model experiences a slight performance drop with STIO=0.78, PESQ=2.81, SiSNR=6.02, SDR=6.67. As shown in the table, the quantized model only requires 625 ms to process a one-second sample, allowing ClearSpeech to run in real-time. With the adoption of mobile GPU, the latency is expected to be further shortened.

9 DISCUSSION

In this section, we discuss some design details, limitations, and potential future work.

O2I model: To synthesize the training data, we pre-trained an O2I model to convert the external noise to its in-ear version. Then, the converted noises are added to self-collected speech data to train the SE models. One might question the necessity of the O2I model and ask what if we use the collected noise signals to synthesize the noisy speech data? To answer this question, we re-train the SE models using collected noise and speech data. The final SE performance drops by around 10% compared to using the O2I transformation. The reason is that public datasets contain much more variations (e.g., recording environment, type and placement of microphone, etc.) than our self-collected noise dataset. However, in practice, collecting a large-scale dataset with different noise types and recording conditions is extremely labor-intensive and time-consuming.

One might also ask why not train another model to convert external speech to the in-ear version so that existing public speech datasets can be used for training as well? We explored this method initially but found it very challenging to obtain a model with satisfactory performance in terms of magnitude transformation. The reason is that in-ear speech is composed of bone-conducted (dominant) and air-conducted speech, while out-ear speech consists of air-conducted speech only. Moreover, different people have different bone structures, making the generalization of the model problematic. This was also demonstrated in Section 8.4 where model personalization can further improve the performance. In contrast, both in-ear and out-ear noises are composed of air-conducted sound only, so it would be possible and easier to learn their correlation with DL.

Phase quantization: As shown in Figure 7(a), more quantization levels yield closer SE performance compared to the continuous phase. However, such analysis is based on a perfect classification of the quantized phase. As the classification becomes less accurate with more classes, there exists a trade-off between quantization level and classification accuracy (thereby affecting the SE performance). Thus, we trained another two phase estimation models with quantization levels of 10 and 20 respectively, and calculated the SE metrics. The results of both models were worse than using a quantization level of 15 – implying that the SE performance is not only determined by the resolution of the quantized phases but also by the complexity of classification. Moreover, to demonstrate the effectiveness of the proposed quantized phase based estimation, we maintained the same model structure and trained another phase reconstruction model with continuous phase (i.e., U-Net based regression, where MSE is used as the loss function). This continuous phase model yields SE performance of (STIO=0.75, PESQ=2.63, SiSNR=5.53, SDR=3.55), which is consistently lower than the quantized phase model of (STIO=0.79, PESQ=2.89, SiSNR=6.42, SDR=7.16).

Noisy phase - in-ear vs. out-ear : In the phase reconstruction model, we utilize the noisy phase from the out-ear signal as input because we aim to recover the clean out-ear speech. Another possible design choice is to use the noisy phase from the in-ear signal as it has better signal quality. However, there is also a phase shift between the in-ear and out-ear speech data due to the occlusion effect. We posit that this double phase interference (by noise and occlusion effect) makes it much tougher to perform accurate phase reconstruction. To demonstrate this, we retrain the phase prediction model with the noisy in-ear phase, and this model using noisy in-ear phase obtains SE performance of (STIO=0.76, PESQ=2.65, SiSNR=-9.29, SDR=-4.57). Overall, the performance is poorer compared to our chosen model that uses out-ear noisy phase data. In particular, the SDR and SiSNR of the in-ear phase data model are significantly poorer as these metrics are very sensitive to phase shifts.

Overhead of the in-ear microphone : The in-ear microphone on existing earbuds is only used for active noise cancellation (ANC), which requires an extremely short processing delay and therefore the ANC algorithm is executed on the onboard audio chip. In our current design, the DNN models are expected to be executed on the smartphone that is paired with the earbuds because the smartphone is more powerful in terms of computation, memory, and battery capacity. More importantly, human speech during calls has to be streamed to the phone anyway before it can be forwarded to the recipient through WiFi or cellular connections.

In addition to sending the out-ear data to the phone (which is then sent to the other side of the conversation), the overhead added by our scheme is that the in-ear data also has to be sent so that it can be processed by our solution to perform SE – usually this in-ear data is processed by the earbud and then discarded. The data is sent by the earbud to the phone using BLE. Even though sending the in-ear data as well as the regular out-ear data could double the amount of data sent between the earbud and the phone, this is still acceptable, in terms of transmission delays and power consumption, as the actual magnitude of data sent is still small. Specifically, the sampling rate of the in-ear microphone can be down to 8 kHz for SE; then, with 2 Bytes (16 bits) per sample, the additional in-ear data rate is 2 mics × 8 kHz × 16 bit = 256 kbps. This is much smaller than the BLE data rate which is measured in Mbps – thus there will be no significant impact on the transmission time as BLE has more than sufficient bandwidth for the additional data.

In addition, as SE will be only activated during phone calls performed in a noisy environment (which only occupies a small portion of the earbuds usage), the impact of these additional transmissions on power consumption is also limited. Moreover, with the advancement in efficient DNN execution, some existing commercial earbuds (e.g., Huawei Freebuds, JBL LIVE PRO+) can already support DNN-based noise cancellation, which provides some promising insights to execute SE models on the earphone in the future.

Using additional microphones: Currently, we only exploited the in-ear and out-ear microphones (i.e., two channels) from the left earbud to train the SE models. However, commercial wireless earbuds usually have a symmetric microphone deployment on both earbuds. Therefore, a future research direction is to exploit more onboard microphones (e.g., 3 or 4) from both earbuds for SE. A significant advantage of this scheme is that microphones from the left and right ears can provide spatial information, which would be very useful for SE as demonstrated in [7, 35].

10 RELATED WORK

10.1 Speech Enhancement

Speech enhancement, which aims to extract a clean signal from a noisy source, has been investigated for decades. Traditional signal processing (SP) techniques are analysed either in the time or frequency domain, including spectral subtraction [5], Wiener filtering [32], nonnegative matrix factorization [27], etc. However, these methods generally fail when SNR is low or when the noise is complex and non-stationary. Recently, with the advances in DL, a number of studies have explored DL for speech enhancement. Different neural network structures have been investigated, including the restricted Boltzmann machine [41], deep neural network with skip connections [37], ensemble learning with multiple networks [19]. Some other works also attempted SE with multi-model information, such as audio with video [26] or audio with ultrasound [36]. However, the speech signals in these approaches are usually from a single out-ear microphone and the performance is moderate. By using multiple out-ear microphones, the SE performance can be improved due to the beamforming gain from direction/spatial information [7, 16, 38]. Instead, our work differs and advances in (1) proposing the joint use of in-ear and out-ear microphones for SE; (2) proposing the O2I model to enable the reuse of existing public datasets; (3) extracting global features across different frequencies to account for the frequency differences of in-ear and out-ear speech data; (4) exchanging mutual information during feature extraction using gate mechanism and learning complementary information at different frequencies using independent streams; (5) estimating clean phase with frequency truncation and phase quantization.

10.2 In-ear Microphone Based Sensing

In addition to ANC, the in-ear microphone has been employed for various sensing applications in the research community, which can be grouped into four categories: (1) Human-computer interaction. Ma et al. presented OESense [23], a gesture recognition system that utilizes an in-ear microphone to capture the bone-conducted vibrations generated by finger tapping on the face. Jin et al. [18] proposed to detect silent speech commands using a speaker and an in-ear microphone; (2) Authentication. Similarly, with the ultrasound sensing technique, EarEcho [13] distinguishes different earbud users based on the fact that the shape of the human ear canal is unique for each person. EarGate [12] authenticates the earbud user purely with an in-ear microphone to record the vibrations generated during walking (similar to gait); (3) Vital signs monitoring. Due to the occlusion effect, some internal body sounds (e.g., heartbeat) are amplified inside the ear canal. Thus, researchers exploited the in-ear microphone to detect human heart rate and respiratory rate [6, 24]; (4) Behaviors sensing. Similarly, external stimuli (e.g., walking and chewing) produced body sounds can be also measured by the in-ear microphone. Therefore, human activity recognition and step counting can be also realized, as demonstrated in [23]. Compared to the above works, we focus on analyzing the properties of in-ear speech and exploiting the in-ear microphone for SE on wireless earbuds.

11 CONCLUSION

We presented ClearSpeech, a novel SE system on wireless earbuds with the joint use of the in-ear and out-ear microphones. We conducted an in-depth analysis of the characteristics of in-ear and out-ear speech. Based on the derived observations, we designed a set of techniques to effectively utilize the two signals for SE. With the developed prototype and collected data, we demonstrated the superior SE performance of ClearSpeech. ClearSpeech does not require additional hardware and can be readily deployed on future wireless earbuds with minimal engineering and economic overhead.

ACKNOWLEDGMENTS

This research was supported by the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant (Grant ID: 21-SIS-SMU-036).

REFERENCES

[1] Online. Bela Mini Board. https://learn.bela.io/products/bela-boards/bela-mini/. (Accessed on Dec 4, 2022).

[2] Online. Microphone. https://www.cuidevices.com/product/resource/cmc-4015-40l100.pdf. (Accessed on Dec 4, 2022).

[3] Online. Pytorch Quantization. https://pytorch.org/docs/stable/quantization.html. (Accessed on Dec 4, 2022).

[4] Takayuki Arakawa, Takafumi Koshinaka, Shohei Yano, Hideki Irisawa, Ryoji Miyahara, and Hitoshi Imaoka. 2016. Fast and accurate personal authentication using ear acoustics. In *2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)*. IEEE, 1–4.

[5] Steven Boll. 1979. Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on acoustics, speech, and signal processing* 27, 2 (1979), 113–120.

[6] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2021. Motion-resilient heart rate monitoring with in-ear microphones. *arXiv preprint arXiv:2108.09393* (2021).

[7] 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.

[8] Jun Chen, Wei Rao, Zilin Wang, Jiuxin Lin, Zhiyong Wu, Yannan Wang, Shidong Shang, and Helen Meng. 2023. Inter-Subnet: Speech Enhancement with Subband Interaction. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[9] Jun Chen, Zilin Wang, Deyi Tuo, Zhiyong Wu, Shiyin Kang, and Helen Meng. 2022. FullSubNet+: Channel attention fullsubnet with complex spectrograms for speech enhancement. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 7857–7861.

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

[11] Sefik Emre Eskimez, Takuya Yoshioka, Huaming Wang, Xiaofei Wang, Zhuo Chen, and Xuedong Huang. 2022. Personalized speech enhancement: New models and comprehensive evaluation. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 356–360.

[12] 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.

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

[14] John S Garofolo. 1993. Timit acoustic phonetic continuous speech corpus. *Linguistic Data Consortium, 1993* (1993).

[15] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 776–780.

[16] Lloyd Griffiths and CW Jim. 1982. An alternative approach to linearly constrained adaptive beamforming. *IEEE Transactions on antennas and propagation* 30, 1 (1982), 27–34.

[17] Mattias P Heinrich, Maik Stille, and Thorsten M Buzug. 2018. Residual U-net convolutional neural network architecture for low-dose CT denoising. *Current Directions in Biomedical Engineering* 4, 1 (2018), 297–300.

[18] Yincheng Jin, Yang Gao, Xuhai Xu, Seokmin Choi, Jiyang Li, Feng Liu, Zhengxiong Li, and Zhanpeng Jin. 2022. EarCommand: " Hearing" your silent speech commands in ear. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–28.

[19] Pavan Karjol, M Ajay Kumar, and Prasanta Kumar Ghosh. 2018. Speech enhancement using multiple deep neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5049–5052.

[20] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. 2013. Speech enhancement based on deep denoising autoencoder.. In *Interspeech*, Vol. 2013. 436–440.

[21] Yi Luo, Zhuo Chen, and Takuya Yoshioka. 2020. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 46–50.

[22] Yi Luo and Nima Mesgarani. 2019. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 27, 8 (2019), 1256–1266.

[23] 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.

[24] Alexis Martin and Jérémie Voix. 2017. In-ear audio wearable: Measurement of heart and breathing rates for health and safety monitoring. *IEEE Transactions on Biomedical Engineering* 65, 6 (2017), 1256–1263.

[25] Héctor A Cordourier Maruri, Paulo Lopez-Meyer, Jonathan Huang, Willem Marco Beltman, Lama Nachman, and Hong Lu. 2018. V-Speech: noise-robust speech capturing glasses using vibration sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 4 (2018), 1–23.

[26] 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.

[27] Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon. 2013. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing* 21, 10 (2013), 2140–2151.

[28] H Gustav Mueller, Kathryn E Bright, and Jerry L Northern. 1996. Studies of the hearing aid occlusion effect. In *Seminars in Hearing*, Vol. 17. Copyright© 1996 by Thieme Medical Publishers, Inc., 21–31.

[29] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Icml*.

[30] Nikolai Novitski, Minna Huotilainen, Mari Tervaniemi, Risto Näätänen, and Vineta Fellman. 2007. Neonatal frequency discrimination in 250–4000-Hz range: Electrophysiological evidence. *Clinical Neurophysiology* 118, 2 (2007), 412–419.

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

[32] Pascal Scalart et al. 1996. Speech enhancement based on a priori signal to noise estimation. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 2. IEEE, 629–632.

[33] Roman Schlieper, Song Li, Stephan Preihs, and Jürgen Peissig. 2019. The relationship between the acoustic impedance of headphones and the occlusion effect. In *Audio Engineering Society Conference: 2019 AES International Conference on Headphone Technology*. Audio Engineering Society.

[34] Stefania Sesia, Issam Toufik, and Matthew Baker. 2011. *LTE - The UMTS Long Term Evolution: From Theory to Practice*. John Wiley & Sons Ltd.

[35] Irtaza Shahid, Yang Bai, Nakul Garg, and Nirupam Roy. 2022. VoiceFind: Noise-resilient speech recovery in commodity headphones. In *Proceedings of the 1st ACM International Workshop on Intelligent Acoustic Systems and Applications*. 13–18.

[36] 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.

[37] Ming Tu and Xianxian Zhang. 2017. Speech enhancement based on deep neural networks with skip connections. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 5565–5569.

[38] Barry D Van Veen and Kevin M Buckley. 1988. Beamforming: A versatile approach to spatial filtering. *IEEE assp magazine* 5, 2 (1988), 4–24.

[39] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. 2018. Understanding convolution for semantic segmentation. In *2018 IEEE winter conference on applications of computer vision (WACV)*. Ieee, 1451–1460.

[40] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. 2015. Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing* 24, 3 (2015), 483–492.

[41] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. 2013. An experimental study on speech enhancement based on deep neural networks. *IEEE Signal processing letters* 21, 1 (2013), 65–68.

[42] 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.

[43] Asri Rizki Yuliani, M Faizal Amri, Endang Suryawati, Ade Ramdan, and Hilman Ferdinandus Pardede. 2021. Speech enhancement using deep learning methods: a review. *Jurnal Elektronika dan Telekomunikasi* 21, 1 (2021), 19–26.

[44] 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.