Exploring and Addressing Poor Auxiliary Modality in Earable Dual-microphone Speech Enhancement

ANONYMOUS AUTHOR(S)

To enhance the speech clarity in earable voice interaction scenarios, dual-microphone speech enhancement (SE) techniques with collaboration of in-ear and out-ear microphones have garnered significant attention from the research community. Nevertheless, existing dual-microphone SE techniques are established on a strong assumption: *high-quality in-ear speech (auxiliary modality) could provide efficient complementary information to the target airborne speech (primary modality)*, which decreases the adaptation in the real world. In our work, we explore a key observation that air pressure imbalance caused by ear canal deformation (ECD) adversely affects the in-ear auxiliary modality, subsequently leading to a significant degradation in speech enhancement performance. To address the bottleneck issue of poor auxiliary modality, we design an efficient quality-aware speech enhancement solution, named QuaSE, which efficiently and dynamically fuses complementary information from in-ear and out-ear microphones by assessing the quality variations of the in-ear speech. Additionally, based on the analysis of spectral distortion induced by ECD, a training strategy including quality-aware data selection and content-aware augmentation is designed to improve the generalization capability of QuaSE. Extensive experiments demonstrate that QuaSE outperforms state-of-the-art techniques by 9.35%, 4.55%, 17.68%, and 12.89% in terms of PESQ, STOI, SI-SDR, and SegSNR. Moreover, we also validate that the proposed quality-aware fusion strategy can be modularly integrated into other sensing tasks, improving the fusion performance.

Additional Key Words and Phrases: Earable Sensing, Speech Enhancement, Quality-aware Sensing.

ACM Reference Format:
Anonymous Author(s). 2025. Exploring and Addressing Poor Auxiliary Modality in Earable Dual-microphone Speech Enhancement. 1, 1 (October 2025), 24 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Voice interaction is increasingly prevalent on earable devices, providing intuitive, hands-free control while enhancing user convenience and accessibility [37]. Consequently, speech enhancement (SE) technology designed for earables has emerged to ensure clear and intelligible speech input, particularly in noisy environments. Compared with single-channel SE techniques, multi-modality/channel SE techniques have demonstrated excellent performance, including IMU-audio [21], ultrasound-audio [13, 14, 49], and RF-audio [28].

Among them, dual-microphone SE techniques [19, 30] with the collaboration of in-ear and out-ear microphones have become one of the alternatives with wide availability and high universality. Narrow-bandwidth bone-conducted speech captured by in-ear microphones (referred to as in-ear speech) is less susceptible to ambient noise due to special characteristics of the channel, which can provide complementary information to noisy-susceptible airborne speech. This fusion significantly improves the intelligibility and quality of airborne speech in noisy conditions. However, most existing studies have been predicated on a *strong assumption* that in-ear speech exhibits a highly stable correlation with airborne speech. In our work, a key observation is identified: the cross-channel correlation between in-ear and airborne speech may be severely degraded in practical scenarios, directly limiting the applicability of prior techniques.

XXXX-XXXX/2025/10-ART $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Key Observation. During the pronunciation process, articulatory gestures cause deformation of the ear canal by stretching or compressing facial muscles and soft tissues around the ear canal. Ear canal deformation (ECD) changes the air pressure in the sealed cavity formed by the earphone and the eardrum. Such ECD-induced pressure fluctuations alter the in-ear microphone's diaphragm response, compromising the fidelity of microphone recordings. In our in-depth investigation, we observed that ECD-induced pressure fluctuations introduce significant distortions into captured in-ear speech (see Sec. 3.3), significantly degrading the cross-channel correlation between in-ear and airborne speech. This observation underscores a critical limitation of existing speech enhancement techniques based on in-ear speech fusion. In dual-microphone speech enhancement tasks, the degradation of in-ear speech quality significantly undermines overall fusion performance, a phenomenon commonly referred to as modality imbalance [50].

Extensive research [3, 31, 38] has been conducted to characterize speech quality using objective or subjective metrics, such as Perceptual Evaluation of Speech Quality (PESQ) and Mean Opinion Score (MOS). However, these quality metrics mainly assess the global perceptual quality of speech and are inadequate for evaluating the integrity of detailed structural components within the signal. Thus, these quality metrics are difficult to guide the effective fusion of airborne speech with low-quality in-ear speech. In recent years, dynamic multi-modality fusion frameworks towards low-quality data have been widely studied [46, 50, 51]. Nevertheless, most of them mainly focus on the fusion of images, videos, or text in the classification tasks. Therefore, an urgent issue arises: "*How to dynamically fuse in-ear speech for effective and robust airborne speech enhancement?*"

To bridge this gap, several challenges need to be urgently addressed. *(i) Fine-grained assessment of time–frequency distortion of in-ear speech without matching references.* In practice, airborne speech corresponding to in-ear speech is often corrupted by external interference, making it unreliable as a reference for distortion assessment. Consequently, in the absence of prior knowledge, quantitatively evaluating the time–frequency distortion of in-ear speech remains highly challenging. *(ii) Effective and dynamic fusion of airborne speech with quality-varying in-ear speech.* The fluctuating quality of the in-ear modality can introduce imbalance and even degrade overall enhancement performance. Moreover, these unpredictable quality variations pose a significant challenge for designing adaptive fusion mechanisms. *(iii) Real-time and generalized speech enhancement across individuals and scenarios.* Individual differences (like ear canal anatomy and pronunciation habits) and scenario diversity (like noise characteristics) lead to substantial variability in signal patterns, while real-time response imposes stringent constraints on computational complexity, especially for IoT devices.

Thus, we propose QuaSE, an effective and robust dual-microphone airborne speech enhancement solution by addressing the low-quality in-ear speech. QuaSE integrates quality variations of in-ear speech into dynamic fusion to mitigate the negative impact of low-quality in-ear speech. First, we investigate and analyze the ECD-induced speech distortion phenomenon from the hardware aspect and explore its non-negligible negative impact on the speech enhancement task. Building on that, we propose a lightweight self-supervised quality assessment framework incorporating a data quality selection strategy to generate quality indicators that can represent the time-frequency structure distortion of in-ear speech. In particular, we introduce a spectral peak-to-valley matching algorithm that quantifies a reliable quality metric to perform quality-aware data selection. Then, we ingeniously transform quality indicators to high-level embeddings, which can serve as a middleware and capture inherent correlation to guide the dynamic fusion of cross-modality complementary features. Subsequently, we carefully design a computationally efficient deep learning framework integrating the above modules to achieve effective airborne speech enhancement. Additionally, a content-aware low-quality data augmentation approach that applies adaptive time masking on time-frequency structures is designed to improve the generalization.

Through extensive experiments with 32 participants in daily interaction scenarios, we validate that QuaSE outperforms the state-of-the-art baselines by 9.35%, 4.55%, 17.68%, and 12.89% in terms of PESQ, STOI, SI-SDR, and SegSNR, respectively. With the contributions of the designed quality-aware adaptation (QA) module, SI-SDR and SegSNR can be improved by up to 11.76% and 10.31%. The QA module can also serve as an intermediate

component to improve the performance of existing solutions. Lastly, we believe that our solution can provide valuable insights into multi-modal sensing tasks with the presence of low-quality modalities.

2 RELATED WORK

2.1 Multi-modal Speech Enhancement for Earables

Since earable devices have emerged as a unique voice input side, it is necessary to develop effective speech enhancement technology for earables to improve the quality of speech. Conventional single-modal speech enhancement (SE) technology [11, 25, 47] takes advantage of distribution differences between clean speech and noise to denoise. Nowadays, multi-modal SE technologies have demonstrated superior performance compared to conventional single-modal SE technologies. In general, multi-modal SE technologies augment the audio modality by leveraging a complementary modality that is less sensitive to noise, such as ultrasounds [13, 41, 49] and RF signals [28, 29, 33]. However, these solutions either require active operation by the user or rely on dedicated RF modules to complete the speech enhancement process, making them not suitable for earable devices.

Nowadays, specifically for earables, various multimodal speech enhancement solutions have been proposed. For example, He *et al.* [21] leverage motion sensors on earphones to capture bone-conducted sound vibrations for speech enhancement. However, to maintain excellent performance, such technology usually requires high-sampling motion sensors that are unavailable on earables, to recover speech's high-frequency components. Duan *et al.* [14] leverage ultrasonic leakage from earables to sense articulatory gestures. However, such technology requires spacers to be attached to the ear pads and requires earables to support the boom microphone structure, which is not suitable for special types of earables like wireless earables and in-ear earables.

As the number of microphones on earables increases, multi-channel speech enhancement technologies have been studied [7]. Most importantly, the in-ear audio modality has been explored as an effective and ubiquitous complementary modality for speech enhancement [19, 30], since it can be captured by low-cost microphones embedded inside of earables. However, existing solutions are based on the high-quality in-ear audio modality to enhance the airborne audio modality. In the real world, the quality of in-ear auxiliary modality cannot be guaranteed due to air pressure variations induced by ear canal deformation. Thus, our work aims to enable effective airborne speech enhancement with low-quality in-ear modality.

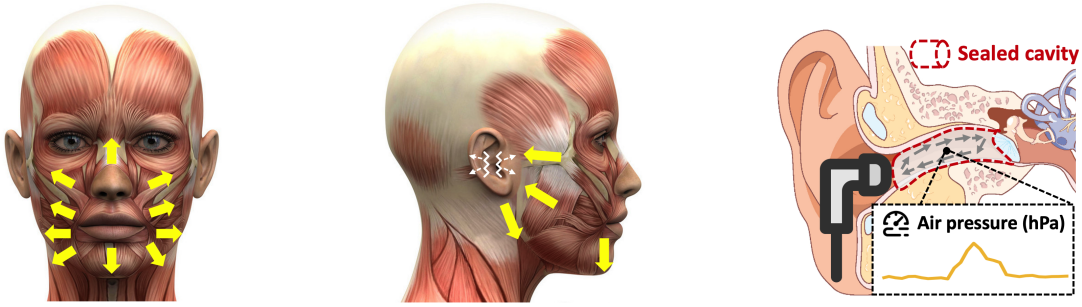
2.2 In-ear Acoustic Sensing on Earables

Since various acoustic sensors have been widely equipped on earables, in-ear audio modality has empowered sensing capability of earables [37]. promote the application of earables in various scenarios including health care [4, 17, 23], user authentication [16, 18, 24, 54], and human-computer interaction (HCI) [27, 40, 53]. Recently, ear canal deformation (ECD) measurements with earables have attracted widespread research attention. Since ECD can be influenced by various activities, accurate ECD measurements can help recognize human activities including tongue-jaw movements [5], articulatory gestures [26, 43], facial expression [1, 52], and vital signs [42]. Furthermore, some significant solutions [15, 44] have explored ECD-based biometric technologies for secure user authentication. In our work, we explore the negative impact of ear canal deformation on in-ear audio modality, thereby designing the corresponding methodology to empower the sensing capability of in-ear complementary modality.

3 MOTIVATION AND PRELIMINARY

3.1 Motivation: Speech Enhancement with In-ear and Out-ear Microphones

Nowadays, earphones have become a novel speech interaction or voice input platform. A key issue is how to ensure speech intelligibility in noisy environments. Since most earphones are equipped with not only out-ear microphones but also in-ear microphones that can capture the sound transmitted into the ear canal through



(a) Articulatory gestures. (b) Ear canal deformation. (c) Air pressure variations.

Fig. 1. Articulatory gestures cause geometrical deformation of the ear canal, which in turn causes air pressure variations inside the ear canal.

bone conduction, dual-microphone-based speech enhancement techniques that leverage in-ear and out-ear microphones have been widely studied [19, 22, 30]. Due to the particularity of the channel, prior literature has explored that in-ear speech is non-susceptible to ambient noise, making it an ideal auxiliary modality to improve the intelligibility of airborne speech. However, existing techniques are studied based on the following assumption:

- *There is a high correlation between the auxiliary modality (i.e., in-ear speech) and the target modality (i.e., airborne speech), which ensures that the auxiliary modality can offer complementary information to the target modality.*

This strong assumption, however, may not be met in real-world scenarios. Based on our practical observations in Sec. 3.3, we find that in-ear speech will experience severe interference, correspondingly decreasing the correlation between airborne speech. Existing dual-microphone-based speech enhancement techniques [19, 22, 30] ignore such negative impact of low-quality in-ear auxiliary modality on the target modality, making it difficult to ensure their robustness and effectiveness in real scenarios.

3.2 Air Pressure Variations Induced By ECD

3.2.1 ***Ear canal deformation (ECD) induced by articulatory gestures.*** The ear canal, also known as the external auditory canal, is a tubular structure that extends from the outer ear (pinna) to the tympanic membrane (eardrum). Soft tissues surrounding the ear canal are situated between the mandible and the mastoid and are highly deformable. As shown in Fig. 1(a), during the pronunciation process, speech organs (such as the mandible, lips, tongue, jaw, and velum) cause facial muscles to stretch or compress. These articulatory gestures may activate movements of the mandible and facial muscles and pull the soft tissues around the ear canal to move, changing the shape of the ear canal, *i.e.*, ear canal deformation (ECD) [10, 12], as shown in Fig. 1(b). In general, the ear canal model can be discretized into 11 cross sections [12]. The morphological parameters of each cross section (including diameter, circumference, area, curvature, and angulation) contribute to ear canal deformation. In addition to articulatory gestures, head movements, insertion depth of the earphones, *etc.*, can also cause slight ear canal deformation [2]. In our work, we mainly address the impact of ECD induced by articulatory gestures.

3.2.2 ***Air pressure variations inside the ear canal.*** As shown in Fig. 1(c), we can observe that the sealed space is created when the earphone fully seals the ear canal. In this fully sealed cavity, the ear canal deformation alters the available volume for the air inside. According to basic principles of fluid mechanics, when the space volume decreases because the ear canal walls press inward, the air molecules are forced closer together, resulting

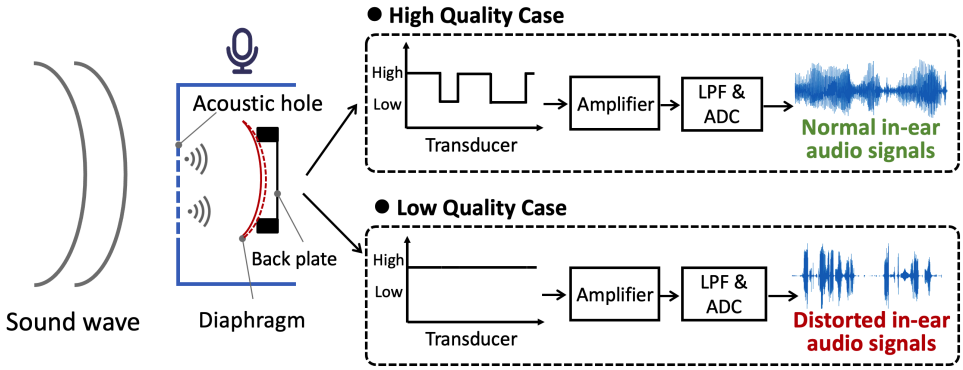


Fig. 2. A basic workflow illustration of the electret condenser microphone and MEMS microphone. When the ear canals are fully sealed by earphones, audio signals captured by in-ear microphones are distorted by ECD-induced air pressure variations.

in an increase in air pressure. Conversely, if the space volume increases slightly, the air pressure may drop. Therefore, when the user is speaking, ear canal deformation induced by articulatory gestures will alter the air pressure in the ear canals. For instance, Ando *et al.* [2] have explored the feasibility of leveraging air pressure in ear canals for face-related movement recognition.

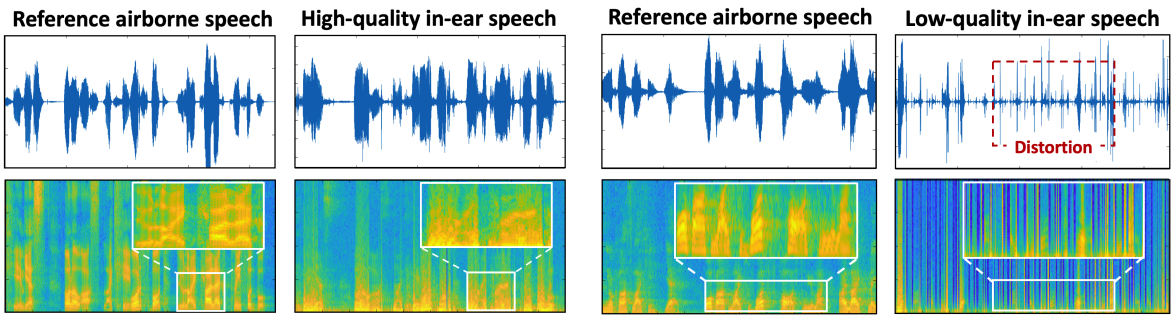
3.3 Understanding the Phenomenon of In-Ear Modality Distortion

3.3.1 ***Basic workflow of Microphone Chips.*** Common microphone types on earphones are Electret Condenser Microphones (ECM) and Micro Electro Mechanical Systems (MEMS) microphones [48]. Regardless of the type of microphone, the basic principle is to convert sound waves into digital signals. Sound waves consist of a series of air pressure variations. A microphone diaphragm responds to sound pressure variations, causing its displacement relative to the back plate, as shown in Fig. 2. The transducer component in the microphone translates mechanical movements to analog electrical signals. Then, digital audio signals are output after being processed by the signal amplifier, low-pass filter (LPF), and analog-to-digital conversion (ADC) circuit.

3.3.2 ***In-ear speech signal distortion induced by pressure imbalance.*** As introduced in Sec. 3.2, articulatory gestures cause ear canal deformation (ECD). The deformation alters the air pressure within the ear canal, often creating a constant (or slowly varying) pressure difference across the diaphragm. As shown in Fig. 2, in the sealed ear canal, the pressure imbalance caused by ear canal deformation counteracts or "clamps" the intended oscillatory forces of the incoming sound waves and inhibits the movement of the diaphragm. According to the prior literature [35], the relationship between the diaphragm movement $d_d(t)$ and the transducer output voltage $u(t)$ can be expressed as follows:

$$u(t) = S_e * (d_d(t) - d_0(t)) \tag{1}$$

where S_e and $d_0(t)$ represent the electronic sensitivity and the initial displacement of the microphone package, respectively. Based on Eq. 1, the output voltage signal $u(t)$ depends on the displacement difference between $d_d(t)$ and $d_0(t)$. When the ear canal deforms and causes changes in the air pressure inside the ear canal, the relative movement of the diaphragm caused by the sound waves will be restricted. Such a phenomenon may cause the microphone sensor to encounter a *stuck-at-low fault*, where the output audio signal remains permanently biased at a low level. Under this condition, the sensor fails to respond to acoustic pressure variations, resulting in a



(a) Speech in the non-fully sealed ear canal. (b) Speech in the fully sealed ear canal.

Fig. 3. In-ear speech recording in the two cases: (a) non-fully sealed ear canal and (b) fully sealed ear canal. Notably, airborne speech is also recorded as the reference signal during the process of in-ear speech recording.

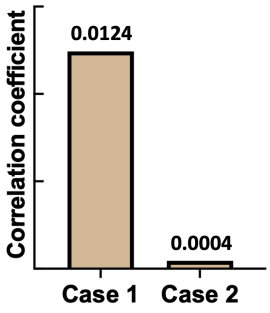
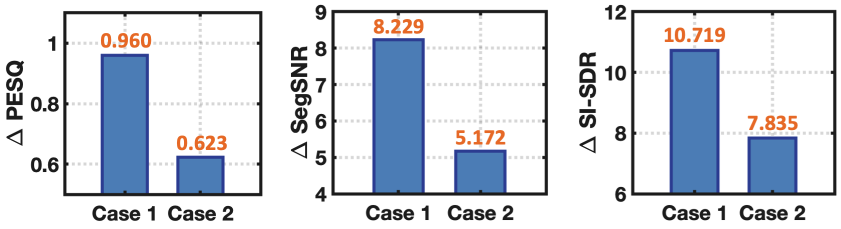


Fig. 4. Correlation analysis of in-ear channel and airborne channel.



(a) Δ PESQ (b) Δ SegSNR (c) Δ SI-SDR

Fig. 5. Negative impact of low-quality auxiliary modality (*i.e.*, in-ear speech) on the dual-microphone-based speech enhancement task.

constant low-level output. Such a malfunction prevents effective sound acquisition and severely compromises the integrity of the recorded in-ear audio signals.

3.3.3 ***Negative impact analysis in the speech enhancement task.*** We designed a proof-of-concept hardware prototype to validate the impact of ECD-induced air pressure variations on in-ear speech quality. Notably, we embedded two AS-B6027AL30-RC electret microphone chips [32] inside the earphones. Then, we required a participant to wear earphones in the following cases:

- *Case 1*: the eartips of earphones fit tightly into the ear canal (insertion depth is about 15 mm, referred to as the fully sealed ear canal)
- *Case 2*: the eartips of earphones did not fit tightly into the ear canal (insertion depth is about 8 mm, referred to as the non-fully sealed ear canal).

During the process of in-ear speech recording, we also used recorded airborne speech as the reference to explore the impact of ECD-induced air pressure variations on in-ear speech. In the non-fully sealed ear canal, airborne speech and in-ear speech are shown in Fig. 3(a). We can observe that in-ear speech aligns with airborne speech,

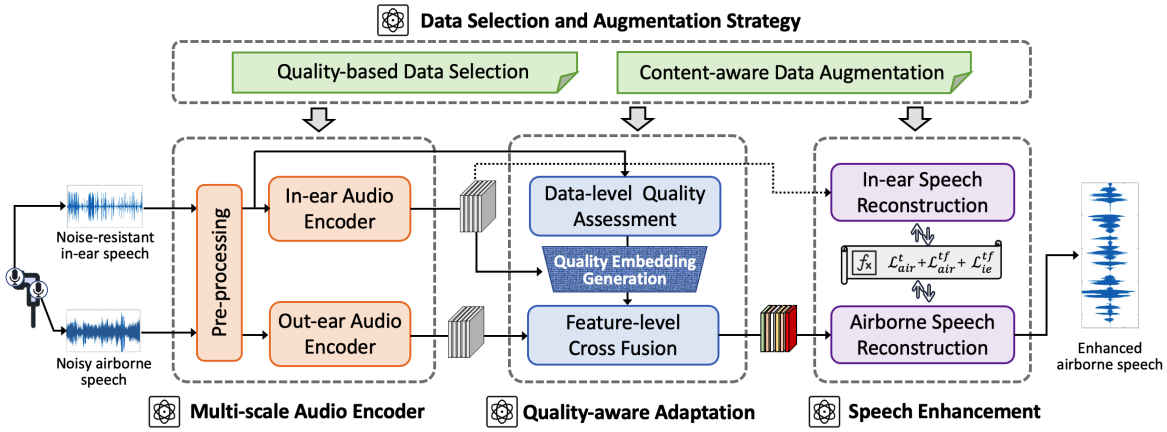


Fig. 6. QuaSE consists of three modules: modality quality assessment, modality quality improvement, and dual-channel speech enhancement.

and the temporal and spectral structures of in-ear speech are fully preserved. For example, the spectrogram of high-quality in-ear speech displays rich and well-defined frequency components. On the contrary, as shown in the white box of Fig. 3(b), low-quality in-ear speech recorded in the fully sealed ear canal is severely distorted, losing temporal and spectral information compared with reference speech. That is because the in-ear microphone's *stuck-at-low fault* causes the microphone output to remain at the "0" level.

The correlation between airborne speech and high-quality in-ear speech is about *30 times* greater than the correlation between airborne speech and low-quality in-ear speech, as shown in Fig. 4, indicating that ECD-induced air pressure variations severely decrease the correlation between in-ear speech and airborne speech. Furthermore, we explore the negative impact of low-quality in-ear speech on the airborne speech enhancement task. Since *Earspeech* designed by Han *et al.* [19] is the state-of-the-art dual-channel-based speech enhancement solution, we reproduce it as our baseline model to conduct the negative impact of low-quality in-ear speech. As shown in Fig. 5, we can clearly observe that low-quality in-ear speech (*Case 2*) decreases the fusion performance in terms of PESQ, SegSNR, and SI-SDR. For instance, with the assistance of high-quality in-ear speech, the fusing performance could bring 0.960, 8.229, and 10.719 improvements on PESQ, SegSNR, and SI-SDR, respectively. However, the fusing performance with low-quality in-ear speech only yields 0.623, 5.172, and 7.835 improvements on PESQ, SegSNR, and SI-SDR, respectively. These experimental results indicate that ECD-induced air pressure variations not only degrade the quality of in-ear speech but also affect the fusion performance.

Insight. Although prior solutions have validated the effectiveness of in-ear speech as an auxiliary modality in airborne speech enhancement, they are all based on the assumption that in-ear speech is of high quality. From the above observation and analysis, we can recognize that ear canal deformation induced by articulatory gestures degrades the quality of in-ear speech by altering the air pressure in the ear canal. Subsequently, the low-quality in-ear auxiliary modality decreases the fusing performance. Therefore, how to ensure the effectiveness and robustness of dual-microphone-based speech enhancement with low-quality in-ear speech is an issue that cannot be ignored.

4 SYSTEM OVERVIEW

To mitigate the negative impact of low-quality in-ear auxiliary modality, we design a deep-learning-based efficient speech enhancement framework that can adaptively assess the contributions of auxiliary modality to the target modality. Fig. 6 demonstrates the technical workflow of QuaSE.

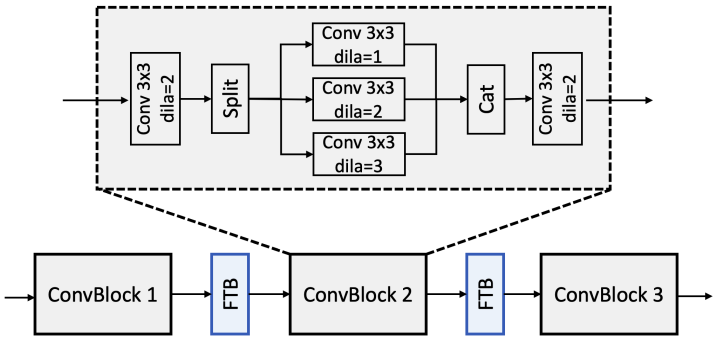


Fig. 7. Model structure of the multi-scale feature extraction.

◊ ***Multi-scale Deep Feature Extraction*** (§5.1): The dual-channel speech collected by in-ear and out-ear microphones is pre-processed and then transformed into time-frequency (T-F) representations. A dual-channel multi-scale audio encoder module consisting of two symmetric audio encoders is designed to capture multi-scale feature representations, achieving a balance between receptive field expansion and detail preservation.

◊ ***Quality-aware Adaptation*** (§5.2): Considering the in-ear modality quality's uncertainty, a self-supervised data-level quality assessment module and a quality embedding generation module are designed to quantify the contribution of the auxiliary modality. Subsequently, a quality-aware cross fusion module combines weighted in-ear feature representations and out-ear feature representations to capture complementary information.

◊ ***Dual-channel Speech Enhancement*** (§5.3): The fused feature representations are fed into the airborne speech reconstruction module and then transformed to T-domain speech signals. It should be noted that the in-ear speech reconstruction module is added to eliminate the problem of modality imbalance, so it only participates in the training process. Additionally, we design a customized multi-dimensional loss function to optimize the model, significantly enhancing its learning capabilities.

◊ ***Data Selection and Augmentation*** (§6): To ensure the generalization and robustness of QuaSE, we construct a large-scale synthetic dataset for pre-training. Based on the cross-channel correlation between airborne and in-ear speech, a data selection strategy is designed to identify the modality quality. Then, based on the spectral characteristics of low-quality in-ear speech, a content-aware data augmentation strategy is designed to expand the scale of the dataset.

5 DESIGN OF QUASE

5.1 Multi-scale Audio Encoder

Because the in-ear microphone also captures motion-induced and heartbeat-induced body sounds, we need to perform the pre-processing operations on the in-ear audio signals. To do this, we apply a high-pass filter with a cut-off frequency of 100 Hz is leveraged to remove the low-frequency components of in-ear and out-ear audio signals. Then, to maintain computational efficiency, we also downsample the two-channel audio signals to 16 kHz. After that, in-ear speech samples and airborne speech samples are transformed into Time-Frequency (T-F) representations via the short-time Fourier transform (STFT).

Taking into account the uncertainty of speech quality, we propose a multi-scale audio encoder model to capture fine-grained features. In addition, the in-ear audio encoder and the out-ear audio encoder are designed to have the same structure to ensure the consistency of extracted in-ear and airborne features. Fig. 7 demonstrates the detailed model structure of the multi-scale audio encoder. At the core, the network is composed of multiple ConvBlocks

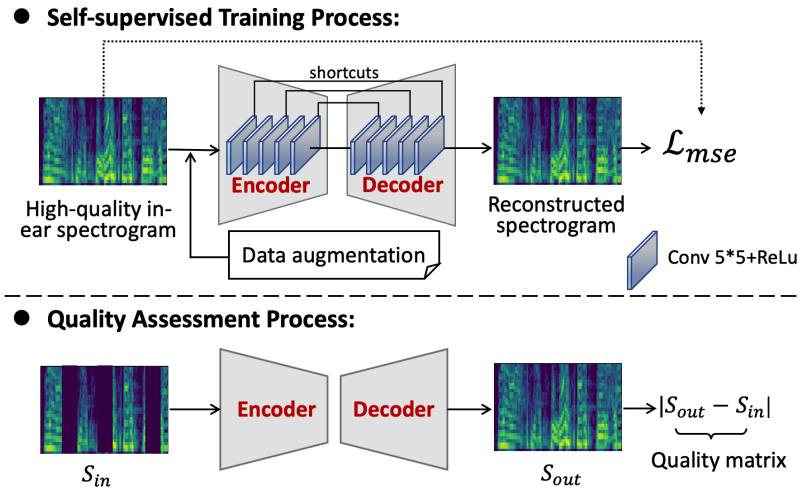


Fig. 8. Self-supervised in-ear speech quality assessment network.

arranged sequentially, interleaved with Frequency Transformation Blocks (FTBs) [47]. Each ConvBlock follows a hierarchical design. An initial 3×3 dilated convolution (dilation = 2) is applied to expand the receptive field, followed by a split operation that distributes features into three parallel branches. These branches apply 3×3 convolutions with different dilation rates (1, 2, and 3), enabling the block to capture spatial information at multiple scales. The outputs of the branches are then concatenated and fused through a final 3×3 dilated convolution (dilation = 2), producing enriched feature maps that integrate local and contextual information.

5.2 Quality-aware Adaptation

5.2.1 ***Data-level Quality Self-assessment.*** As shown in Fig. 3, we can observe that the low-frequency components (≤1000 Hz) of the in-ear speech are significantly enhanced, but the spectral structure is still similar to that of airborne speech. That is because the occlusion effect increases the acoustic impedance inside the ear canal, correspondingly enhancing the gain of low-frequency components and attenuating the gain of high-frequency components [6]. Intuitively, we can identify the in-ear speech quality based on the similarity of the low-frequency spectral structure between in-ear speech and airborne speech. Unfortunately, ambient noise interferes with airborne speech in the real world, making it difficult to obtain clean airborne speech for reference. Therefore, an acute issue arises: *how to accurately characterize in-ear speech quality in the frequency domain without reference speech?*

To solve the above issue, we carefully design a self-supervised in-ear speech quality assessment network with an autoencoder backbone structure. Fig. 8 presents the workflow of the quality assessment network. In order to preserve key information and keep computational efficiency, we adopt fully-connected convolutional layers without down-sampling operations to construct the encoder module. Each convolutional layer is configured with a kernel size of 5×5, a stride of 1, and no padding. The decoder module consists of stacked deconvolutional layers that are symmetrically structured in relation to the convolutional layers of the encoder module. To preserve more structural details in the output of deconvolutional layers, shortcut connection operations are applied to map the convolutional and corresponding deconvolutional layers [8].

In the model training process, we use high-quality in-ear spectrograms to train the network. The high-quality data selection strategy is introduced in Sec. 6.1. The encoder module is responsible for compressing the input high-quality in-ear spectrogram $S_{in}^{T\times F}$ into the high-level feature map $F_{im}^{C\times T\times F}$. Then, the decoder module reconstructs

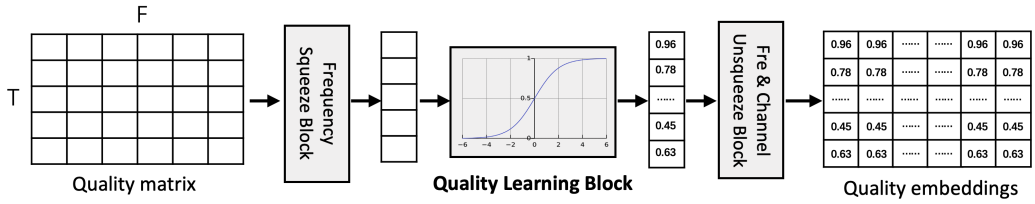


Fig. 9. The workflow of quality embedding generation.

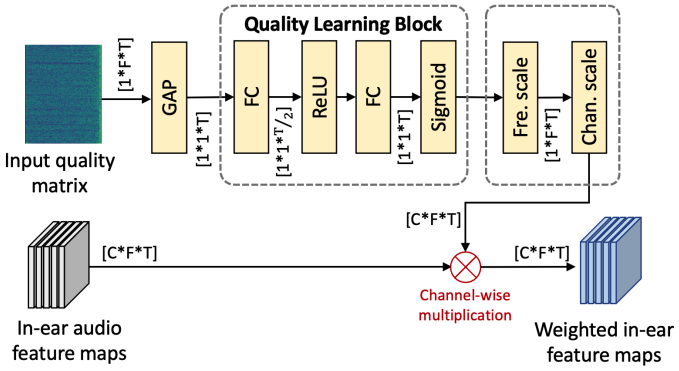


Fig. 10. The process of in-ear feature map weighting and the detailed structure of the quality embedding generation module.

it back into the spectrogram. The entire training process is in unsupervised way. Constrained by the MSE loss function L_{mse}, the feature extraction network can achieve accurate reconstruction of the input spectrogram. In the feature extraction process, only high-quality in-ear spectrograms can be reconstructed with minimal error. Low-quality in-ear spectrograms cannot be accurately reconstructed because their spectral structures have been distorted by ECD-induced air pressure variations. Therefore, we calculate the mean absolute error between the input $S_{in}^{T\times F}$ and output $S_{out}^{T\times F}$ as quality metrics.

5.2.2 ***Quality Embedding Generation.*** If the quality matrix is directly fed into the subsequent modules, the entire network may struggle to converge. Thus, we design a quality embedding generation module that consists of a frequency squeeze block, a quality learning block, and an unsqueeze block. Fig. 9 demonstrates the workflow of quality embedding generation. For a given quality matrix $Q_m \in \mathbb{R}^{1\times F\times T}$, if we perform element-by-element calculations, it will incur high computational overhead and storage pressure. As analyzed in Sec. 3.3, we can find that if the signal is distorted at a certain time T, then the frequency components $F = \{f_1, f_2, ...f_n\}$of the signal at that time will suffer from frequency-independent severe attenuation. Since such attenuation mainly results from the abnormal signal acquisition of the hardware, the attenuation coefficients in different frequency bands are considered to be the same.

Based on this key observation and signal distortion characteristics, we adopt squeeze and unsqueeze operations to improve computing efficiency. Specifically, for a given input quality matrix $Q_m \in \mathbb{R}^{1\times F\times T}$, we leverage a convolutional layer with a kernel size of 1×1 followed by frequency-dimensional global average pooling (GAP)

to squeeze along the frequency dimension:

$$U_{squ}(j) = \frac{1}{n}\sum_{i=1}^{n} Q_m(i,j) \tag{2}$$

$U \in \mathbb{R}^{1\times1\times T}$ represents the squeezed quality matrix. In this way, the complex matrix operations are transformed into efficient vector operations, improving computing efficiency by F times. Subsequently, a quality learning block is designed to fully leverage the information aggregated in the above squeeze operation. To achieve this objective, two fully-connected layers with ReLU (ψ) and Sigmoid (σ) functions are combined, which can learn the underlying nonlinear relationship. It is worth noting that the first fully connected layer reduces the input dimension to 1/2 of the original, balancing the model complexity and learning capability. Through the mapping of the Sigmoid activation function, the value of each element will be in the range of 0-1, representing the weight of quality. The output of the quality learning block can be expressed as follows:

$$U_l = \sigma(W_2\ \psi(W_1 U_{squ})) \tag{3}$$

where W_1 and W_2 represent the weights of two fully-connected layers. Then, a frequency and channel unsqueeze block is added to expand the scale of the feature map along the frequency dimension and the channel dimension, respectively. Finally, the input quality matrix $Q_m \in \mathbb{R}^{1\times F\times T}$ is transformed to the quality embeddings $Q_e \in \mathbb{R}^{C\times F\times T}$ that can be directly multiplied by the in-ear audio feature maps. The detailed quality embedding generation module structure and the process of in-ear feature map weighting are demonstrated in Fig. 10.

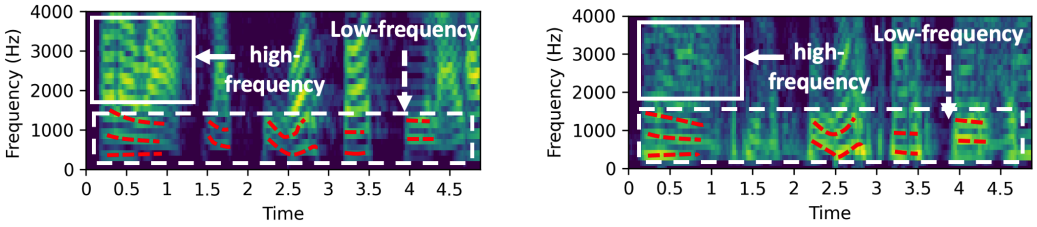
5.2.3 ***Feature-level Cross Fusion.*** After the processing of the multi-scale audio encoder, the input in-ear speech spectrogram ($S_{ie}^{T\times F}$) and the airborne speech spectrogram ($S_{air}^{T\times F}$) are transformed into feature maps $F_{ie}^{C\times T\times F}$ and $F_{air}^{C\times T\times F}$, respectively. The quality embedding $Q_e^{C\times T\times F}$ is multiplied by $F_{ie}^{C\times T\times F}$. Then, the weighted in-ear feature map $Q_e^{C\times T\times F} * F_{ie}^{C\times T\times F}$ and the airborne feature map $F_{air}^{C\times T\times F}$ are concatenated along the channel dimension. To avoid information loss, we do not perform downsampling operations on these feature maps, which brings challenges to real-time and effective feature fusion. Therefore, a lightweight attention mechanism, *i.e.*, Convolutional Block Attention Module (CBAM) [45], is performed on $Cat(Q_e^{C\times T\times F} * F_{ie}^{C\times T\times F}, F_{air}^{C\times T\times F})$, which can sequentially apply channel and spatial attention to refine feature representations, enabling the network to focus on informative regions while suppressing irrelevant features.

5.3 Speech Enhancement Decoder

In the dual-channel fusion-based speech enhancement module, the output of the in-ear audio encoder (*i.e.*, $F_{ie}^{C\times T\times F}$) and the fused feature map (*i.e.*, $F_{fese}^{2\times C\times T\times F}$) are fed into the in-ear speech reconstruction model and the airborne speech reconstruction model, respectively.

5.3.1 ***Airborne Speech Reconstruction.*** In the airborne speech reconstruction model, the fused feature map $F_{fese}^{2\times C\times T\times F}$ is reconstructed to clean airborne speech signals. The core components consist of three 3×3 dilated convolution (dilation = 2) blocks with 72, 48, 24 filters. Each convolution block is followed by batch normalization and the ReLU function. In addition, skip connections are leveraged to directly transfer feature maps from the encoder to the corresponding decoder layers, which helps preserve fine-grained time-frequency details. The final convolution block applies a 3×3 dilated convolution (dilation = 2) layer, batch normalization, and a sigmoid activation function to generate a single-channel spectrogram mask, *i.e.*, $Mask_{air}^{T\times F}$. The final output airborne speech is generated by $\hat{s}_{air} = \text{iSTFT}(Mask_{air}^{T\times F} * S_{air}^{T\times F})$.

5.3.2 ***In-ear Speech Reconstruction.*** Inspired by prior literature [19], we also add an auxiliary in-ear speech reconstruction module to force the model not to forget in-ear branch information during learning. Specifically, it consists of five 3×3 dilated convolution (dilation = 2) blocks with 72, 48, 24, 12, and 1 filters. Each convolution



(a) Airborne speech (b) In-ear speech

Fig. 11. Spectrogram comparison between airborne speech and in-ear speech.

block is followed by batch normalization and the ReLU function. Lastly, a reconstructed in-ear spectrogram $\hat{S}_{ie}^{T\times F}$ will be output. It is worth noting that the in-ear speech reconstruction model only participates in the training process and will be frozen during the inference process.

5.3.3 ***Customized Loss Function***. The customized loss function consists of three components: in-ear time-frequency domain loss, airborne time-frequency domain loss, and airborne time domain loss, which are defined as follows:

$$\mathcal{L}_{air}^{t} = \left\|\hat{s}_{air} - s_{air}^{c}\right\|^2 \tag{4}$$

$$\mathcal{L}_{air}^{tf} = \left\|\hat{S}_{air} - S_{air}^{c}\right\|^2 \tag{5}$$

$$\mathcal{L}_{ie}^{tf} = \left\|\hat{S}_{ie} - S_{ie}\right\|^2 \tag{6}$$

where s_{air}^{c} and S_{air}^{c} represent the clean airborne speech signals and the corresponding spectrogram. $\hat{S}_{air}$ represent the reconstructed airborne speech spectrogram, denoted by $\hat{S}_{air} = Mask_{air}^{T\times F} * S_{air}^{T\times F}$.

6 DATA SELECTION AND AUGMENTATION

Since the quality of in-ear modality varies, we need to design a training strategy including high-quality data selection, masking-based data augmentation, and noise mixing to effectively complete the training process of the designed deep learning model.

6.1 High-quality Data Selection Strategy

The training dataset comprises multiple pairs of airborne and in-ear speech samples. To train the self-supervised quality assessment model, it is essential to select high-quality in-ear speech samples together with their corresponding airborne speech samples. However, the lack of ground-truth high-quality references for in-ear speech poses a major challenge for quality evaluation. To address this issue, we propose an automatic strategy to identify the quality of the in-ear speech

Since in-ear speech and airborne speech are audio signals transmitted along different channels from the same sound source, there is a cross-channel correlation between in-ear speech and airborne speech [19, 20]. Based on the electro-acoustic (EA) model proposed in [6], the cross-channel correlation along the frequency domain can be expressed as follows: (referring to [19] for the detailed derivation)

$$F_{tf} = \frac{S_{ie}(f)}{S_{air}(f)} = \frac{F_{OE}(f) * H_{bone}(f)}{H_{air}} \tag{7}$$

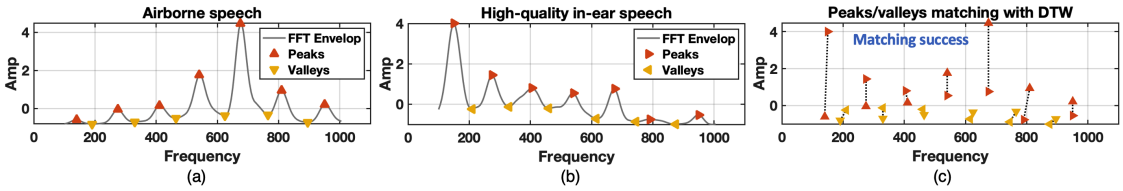


Fig. 12. Peak-valley matching in high-quality data selection. (a) The FFT spectrum envelop of airborne speech;(b) The FFT spectrum envelop of high-quality in-ear speech;(c) Peaks and valleys matching using DTW.

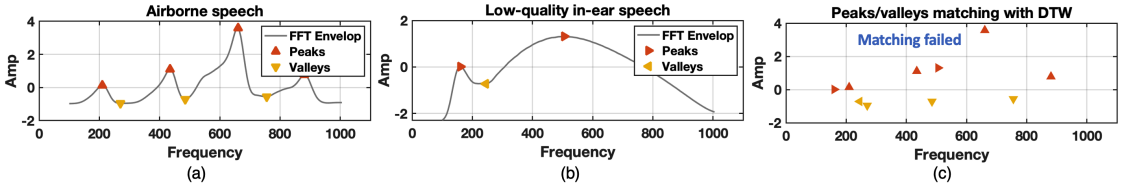


Fig. 13. Peak-valley matching in low-quality data selection. (a) The FFT spectrum envelop of airborne speech; (b) The FFT spectrum envelope of low-quality in-ear speech; (c) Peaks and valleys matching using DTW.

where H_{air} and H_{bone} are frequency-dependent gains of air conduction and bone conduction, respectively. As shown in Fig. 11, we can observe that the low-frequency components (≤1000 Hz) of the in-ear speech are significantly enhanced, but the spectral structure (like fundamental frequency and its corresponding harmonics) is still similar to that of airborne speech. Thus, we design a spectral peak-to-valley matching algorithm to identify the in-ear speech quality based on the similarity of the low-frequency structure between in-ear speech and airborne speech.

Because the high-frequency components of in-ear speech are severely attenuated, we mainly focus on the contribution of low-frequency components. Thereby, we use a bandpass filter with passband frequencies of 100 Hz and 1000 Hz to preserve the low-frequency components of in-ear speech and airborne speech. Then, we split the input speech sample into multiple clips using a sliding window with a length of 100 ms and a step of 50 ms. At each audio clip, the Fast Fourier Transform (FFT) is applied to extract the magnitude spectrum:

$$M(m,\omega)) = abs(\int_{-\infty}^{\infty} x(t)\, w(t-mT)\, e^{-j\omega t}\, dt) \tag{8}$$

where m and ω represent the indexes in the time domain and frequency domain, respectively. To eliminate the interference of residual noise, we adopt a Gaussian smoothing and Z-score normalization to extract the magnitude spectrum envelope of each frame, denoted by $Env(m,\omega)$. Then, we start to find the peaks and valleys of the spectrum envelope based on local extrema detection. Considering the characteristics of the spectrum envelope, the minimum difference between two adjacent peaks or valleys is set to 0.1*N, where N is the length of the spectrum envelope. In addition, a minimum prominence parameter is set to 0.1*std($Env(m,\omega)$) to keep robustness under drifting baselines. After processing, the peak values and valley values of the airborne spectrum envelope and the in-ear spectrum envelope are denoted by $P_{air}, V_{air}, P_{ie}, V_{ie}$, and the peak locations and valley locations are correspondingly denoted by $L^P_{air}, L^V_{air}, L^P_{ie}, L^V_{ie}$. Based on the timing, ordering, and rhythm of peaks and valleys, we quantify similarity by greedy matching and DTW alignment.

◇ ***Greedy matching.*** For the given peak locations and valley locations, we design a greedy matching approach to calculate the matching rate and the matching error, which is given by Alg. 1. As for high-quality in-ear speech and corresponding airborne speech, the peaks and valleys of the spectral envelope should appear at approximately

the same locations. Yet, the peak and valley locations of the low-quality in-ear spectral envelope mismatch with those of the airborne spectral envelope, as shown in Fig. 12 and Fig. 13.

Algorithm 1: Peak/Valley Greedy Matching

Input: Loc_{ie}, Loc_{air}: the peak locations of the in-ear spectrum envelope and the airborne spectrum envelope. τ : matching tolerance.
Output: R: the peak-to-peak matching rate. Err : the absolute error of matched peak locations.

$N_{match} = 0,\ Err = 0$;
for $i \leftarrow 1$ **to** $len(Loc_{ie})$ *and* $j \leftarrow 1$ **to** $len(Loc_{air})$ **do**
  $d = Loc_{ie}(i) - Loc_{air}(j)$;
  if $abs(d) \leq \tau$ **then**
    $N_{match} = N_{match} + 1$;
    $Err = Err + \text{abs}(d)\ i = i + 1,\ j = j + 1$;
  end
  else if $d < 0$ **then**
    $i = i + 1$;
  end
  else
    $j = j + 1$;
  end
end
$R = \frac{N_{match}}{(\max(\text{len}(Loc_{ie}),\text{len}(Loc_{air}))}$;
$Err = Err/N_{match}$;

◊ ***First-order difference DTW alignment***. Then, we calculate the interval sequences of adjacent peaks (or valleys) and align them with DTW (Dynamic Time Warping), which can represent the rhythm of peaks and valleys. For the peak locations of the in-ear spectrum envelope and the airborne spectrum envelope, the DTW alignment algorithm is given by Alg. 2.

Algorithm 2: First-order Difference Alignment

Input: Loc_{ie}, Loc_{air}: the peak locations of the in-ear spectrum envelope and the airborne spectrum envelope.
Output: D: normalized DTW distance.

$D = 0$;
$\Delta_{ie} = \text{diff}(Loc_{ie}), \Delta_{air} = \text{diff}(Loc_{air})$;// First-order difference calculation.
$[\text{D}, y_{ie}, y_{air}] = \text{dtw}(\Delta_{ie}, \Delta_{air})$; // y_{ie} and y_{air} represent the alignment indexs of Loc_{ie} and Loc_{air}.
$D = \frac{1}{1+\frac{D}{len(y_{ie})}}$; // Mapping D into $[0, 1]$;

◊ ***Similarity Score Calculation***. Then, we calculate the similarity score by combining the weighted matching rates and first-order difference distances. If the similarity score is larger than the threshold ξ, the in-ear audio clip is considered to be of high quality. The complete algorithm flow is shown in Alg. 3. Fig. 12 represents peak and valley matching results of the high-quality in-ear speech and corresponding airborne speech. However, as shown in Fig. 13, peaks and valleys of low-quality in-ear speech cannot be accurately matched with those of airborne speech, resulting in large errors.

Algorithm 3: High-quality Data Selection Strategy

Input: s_{ie} and s_{air}: input in-ear and airborne speech signals.
Output: Q_m: the quality metrics of s_{ie}, where 1 represents the high quality and 0 represents the low quality.

Splitting s_{ie} and s_{air} into multiple audio clips $x_{ie}^{i=1,\ldots,M}$ and $x_{air}^{i=1,\ldots,M}$;
for $i \leftarrow 1$ **to** M **do**
  $F_{ie}(\omega) = \mathrm{FFT}(x_{ie}^{i}, f_s),\ F_{air}(\omega) = \mathrm{FFT}(x_{air}^{i}, f_s)$; // 100 Hz $\leq \omega \leq$ 1000 Hz.
  $Env_{ie} = \mathrm{norm}(\mathrm{smooth}(M_{ie}))$; // Extracting the magnitude spectrum envelope.
  $Env_{air} = \mathrm{norm}(\mathrm{smooth}(M_{air}))$;
  $[P_{ie}, L_{ie}^{P}] = \mathrm{findPeaks}(Env_{ie}, \Theta)$;
  $[V_{ie}, L_{ie}^{V}] = \mathrm{findPeaks}((-1) * Env_{ie}, \Theta)$;
  $[P_{air}, L_{air}^{P}] = \mathrm{findPeaks}(Env_{air}, \Theta)$;
  $[V_{air}, L_{air}^{V}] = \mathrm{findPeaks}((-1) * Env_{air}, \Theta)$;
  $[R_{pk}, Err_{pk}] = \mathrm{greedyMatching}(L_{ie}^{P}, L_{air}^{V}, \tau)$; // see Alg. 1.
  $[R_{vl}, Err_{vl}] = \mathrm{greedyMatching}(V_{ie}^{P}, V_{air}^{V}, \tau)$;
  $[D_{pk}] = \mathrm{dtwDistance}(L_{ie}^{P}, L_{air}^{V})$; // see Alg. 2.
  $[D_{vl}] = \mathrm{dtwDistance}(V_{ie}^{P}, V_{air}^{V})$;
  $Score(i) = \alpha_1 * R_{pk} + \alpha_2 * R_{vl} + \alpha_3 * \frac{(D_{pk}+D_{vl})}{2} - \alpha_4 * \frac{Err_{pk}+Err_{vl}}{2}$;
  $Score(i) = \max(0, \min(1, Score(i)))$; // Mapping to [0,1].
end
if $avg(Score) \leq \xi$ **then**
  $Q_m = 0$; // Low quality.
end
else
  $Q_m = 1$; // High quality.
end

6.2 Content-aware Low-quality Data Augmentation

Data augmentation approaches can expand the scale of the dataset, thereby improving the robustness and generalization. Different from prior solutions [19, 20] that focus on the augmentation of high-quality speech, our work needs to expand the scale of low-quality in-ear speech. This work additionally designs a highly interpretable masking-based data augmentation approach, as shown in Fig. 14.

As introduced in Sec. 3.3, in-ear air pressure variations induced by ear canal deformation (ECD) suppress the movements of the microphone diaphragm, thus affecting the sound capture. From the perspective of the spectrogram, it looks like adding masks to cover the original information. First, we leverage a Gaussian Mixture Model (GMM) to generate high-quality in-ear speech from the existing large airborne speech corpus, similar to a prior solution [19]. Then, we add random masks to the synthesized in-ear speech to generate low-quality in-ear speech. Since the ECD-induced air pressure imbalance only interferes frequency component at a certain moment, we perform the time masking on the in-ear spectrogram. Unlike prior methods that apply random masks, we introduce content-aware adaptive time masking, where mask locations are conditioned on the spectral energy distribution. The in-ear speech spectrogram is denoted by $S(T, F)$, where T and F represent the time and frequency bins, respectively. For the m-th time bin, the probability of being masked is defined as:

$$P_m = \frac{\sum_{f=f_0}^{F} S(m, f)}{\sum_{t=t_0}^{T} \sum_{f=f_0}^{F} S(t, f)} \tag{9}$$

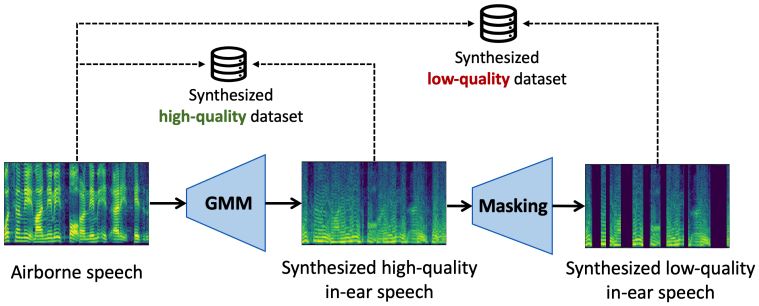


Fig. 14. Workflow of the low-quality data augmentation.

Finally, airborne speech samples combined with their synthesized high-quality in-ear samples constitute the high-quality pre-training dataset, whereas pairing with synthesized low-quality in-ear samples constitutes the low-quality pre-training dataset.

7 EXPERIMENTAL SETUP

7.1 Dataset Description

7.1.1 ***Clean Dual-channel Dataset.*** In the real-world quiet environments, we collect dual-channel speech from 32 participants using three types of earphones that are equipped with in-ear and out-ear microphones, as shown in Fig. 15. Among these participants, 27 are native Chinese speakers and 5 are native English speakers. Each participant is required to wear earphones and read the English material for approximately 20 to 30 minutes. To facilitate subsequent signal processing, all speech signals are split into multiple 5-second clips. If the signal length is less than 5 seconds, it will be automatically padded with 0. In total, a real-world dual-channel speech dataset including about 2800 pairs of airborne speech samples and in-ear speech samples is constructed.

7.1.2 ***High-quality Dual-channel Dataset.*** As introduced in Sec. 5.2.1, a high-quality dual-channel dataset including high-quality in-ear speech and corresponding airborne speech needs to be constructed to train the self-supervised quality assessment model. Based on the design in Sec. 6.1, we need to select high-quality in-ear speech from the clean dual-channel dataset. Specifically, based on the empirical analysis, the similarity score threshold (*i.e.*, ξ) is set to 0.5. α_1, α_2, α_3, and α_4 in Alg. 3 are set to 0.4, 0.4, 0.1, and 0.1, respectively.

7.1.3 ***Noisy Dual-channel Dataset.*** To construct the noisy dual-channel dataset, we need to add noise to the clean dual-speech dataset. We adopt the same noise addition strategy as in the prior literature [19]. Considering the complexity of noise in the real world, ambient noise [36], competing speaker noise [34], and music noise [39] are chosen to be added into the clean speech. Lastly, the SNR of noisy speech ranges from -5 dB to 10 dB, which can cover most noisy scenarios.

7.1.4 ***Synthetic Dual-channel Dataset.*** As introduced in Sec. 6.2, a data augmentation strategy is designed to expand the scale of high-quality and low-quality datasets. Therefore, an existing airborne speech corpus (*i.e.*, LibriSpeech [34]) is leveraged to generate high-quality in-ear speech and low-quality in-ear speech, simultaneously. In total, we construct a synthetic dual-channel dataset that includes about 8000 pairs of airborne speech and in-ear speech.

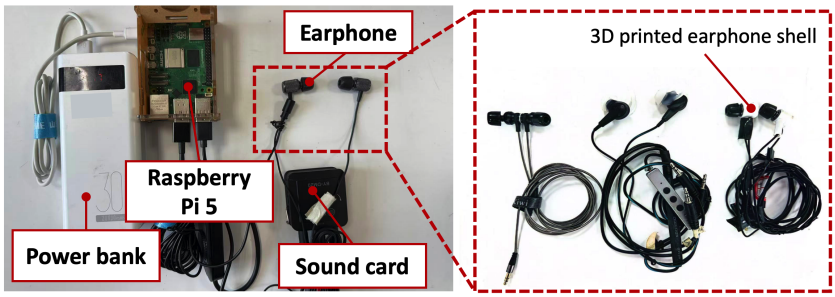


Fig. 15. Hardware implementation.

7.2 Model Training

We implement the proposed model using the PyTorch framework. The complete model training process is conducted on a server that is equipped with 144 Intel(R) Xeon(R) Platinum 8352V@2.10GHz CPUs and 3 NVIDIA RTX A6000 GPUs. During the training process, the noisy airborne speech signals and corresponding in-ear speech signals are fed into the model, and the clean airborne speech signals serve as the ground truth for computing the loss. First, we use the high-quality dual-channel dataset to train the quality self-assessment model. After that, the model parameters are frozen. Then, the synthetic dual-channel dataset is used to pre-train the entire model of QuaSE. Afterwards, we fine-tune the pre-trained model with the noisy dual-channel dataset with a maximum epoch of 32 and a batch size of 12.

7.3 Evaluation Metrics

We divide all participants into eight groups and adopt leave-one-group-out cross-validation for evaluation. In our study, we use four widely used objective metrics to comprehensively evaluate the quality and intelligibility of reconstructed speech:

- *PESQ*. The Perceptual Evaluation of Speech Quality compares the perceptual characteristics of reconstructed and reference signals, producing a score from -0.5 to 4.5. A higher PESQ value reflects better perceived naturalness and overall speech quality.
- *STOI*. The Short-Time Objective Intelligibility estimates how much of the linguistic content in the speech can be understood by a human listener. Its score ranges from 0 to 1, with larger values indicating clearer and more intelligible speech.
- *SI-SDR*. The Scale-Invariant Signal-to-Distortion Ratio evaluates the level of distortion between reconstructed speech and reference speech while being invariant to overall amplitude scaling. Higher values suggest that the reconstructed signal is acoustically closer to the original reference speech.
- *SegSNR*. The Segmental Signal-to-Noise Ratio computes the average SNR across short overlapping frames. Larger values correspond to stronger noise suppression while retaining speech components.

8 EVALUATION ANALYSIS

8.1 Baseline Comparison

To evaluate the overall performance of our solution, we conduct a baseline comparison with several solutions, including Phasen [47], Inter-Subnet [9], and EarSpeech [19]. Among them, Phasen and Inter-Subnet are typical single modality-based speech enhancement solutions. EarSpeech is the state-of-the-art dual-microphone speech enhancement solutions that incorporate in-ear and out-ear microphones. Tab. 1 summarizes speech enhancement

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

Method	PESQ				STOI				SI-SDR (dB)				SegSNR (dB)			
	EN	MN	SN	Avg	EN	MN	SN	Avg	EN	MN	SN	Avg	EN	MN	SN	Avg
Noisy Speech	2.62	2.26	2.31	2.45	0.85	0.78	0.77	0.81	5.03	4.91	5.03	5.00	6.52	1.98	8.92	5.93
Phasen	3.21	3.10	3.01	3.11	0.84	0.81	0.80	0.82	10.85	10.13	8.46	9.81	10.97	10.17	9.51	10.22
Inter-Subnet	3.35	3.14	3.09	3.19	0.85	0.82	0.81	0.83	11.07	10.72	9.46	10.41	11.02	10.65	10.39	10.69
EarSpeech	3.27	3.13	2.89	3.10	0.88	0.87	0.87	0.88	12.47	12.03	11.48	11.99	12.76	10.58	11.59	11.64
QuaSE	3.58	3.43	3.17	3.39	0.93	0.91	0.91	0.92	15.24	12.98	13.04	14.11	13.47	12.53	13.43	13.14
EarSpeech w/ QA	3.37	3.26	3.19	3.27	0.91	0.88	0.86	0.88	13.75	13.46	12.98	13.40	13.55	12.73	12.24	12.84
QuaSE w/o QA	3.19	3.06	3.01	3.09	0.87	0.85	0.85	0.86	14.20	12.05	12.03	13.11	12.34	11.93	11.47	11.91

metrics under different noise conditions, where EN, MN, and SN represent environmental noise, music noise, and speech noise, respectively. QuaSE consistently outperforms existing baselines across all reported metrics. Specifically, QuaSE achieves 9.35%, 4.55%, 17.68%, and 12.89% improvements over the best-performing baseline in PESQ, STOI, SI-SDR, and SegSNR, respectively. Compared with EarSpeech, the significant improvement of QuaSE mainly lies in its capability to address the negative impact of low-quality in-ear speech on the airborne speech enhancement by dynamically fusing the complementary information based on quality variations. For some user groups, QuaSE and EarSpeech demonstrate a comparable performance. That is because the deep learning model in EarSpeech may also exhibit a generalization capability towards low-quality in-ear speech when the training dataset involves low-quality in-ear speech.

8.2 Quality-aware Adaptation Effectiveness

The primary novelty of our solution is the quality-aware adaptation of cross-modal fusion. Next, we evaluate the effectiveness of the quality-aware adaptation (QA) module designed in Sec. 5.2. As shown in Tab. 1, with the assistance of the QA module, the enhancement performance of QuaSE can be improved by 9.71%, 6.98%, 7.62%, and 10.33% in terms of PESQ, STOI, SI-SDR, and SegSNR, respectively. In addition, we also try to integrate the QA module into EarSpeech to explore its feasibility as an auxiliary component to improve performance. From the comparison in Tab. 1, PESQ, SI-SDR, and SegSNR are improved by up to 5.48%, 11.76%, and 10.31%, respectively, revealing the effective gains of the QA module. The performance improvements of EarSpeech indicate that the QA module can also be used as an intermediate component to facilitate the effectiveness of multi-modality sensing tasks.

8.3 Data Augmentation Effectiveness

Next, we assess the effectiveness of data augmentation (Sec. 6.2) in terms of generalization across unseen sentences and unseen languages.

8.3.1 ***Generalization Across Unseen Sentence.*** First, we additionally collect dual-channel speech of previously unseen sentences to evaluate the generalization capability of QuaSE. As shown in Fig. 16, the proposed data augmentation strategy consistently improves performance across multiple metrics (PESQ, STOI, and SI-SDR), demonstrating enhanced robustness to lexical and syntactic variations. For instance, compared with the baseline, QuaSE with data augmentation achieves relative gains of about 10% in terms of SI-SDR. These results confirm that augmentation not only prevents overfitting to sentence-specific acoustic patterns but also strengthens the capability to generalize to diverse sentence structures encountered in real-world scenarios.

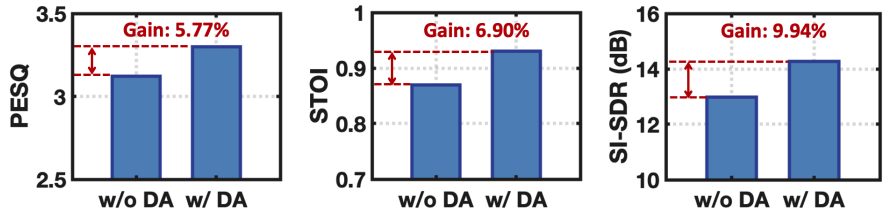


Fig. 16. Generalization among unseen sentences.

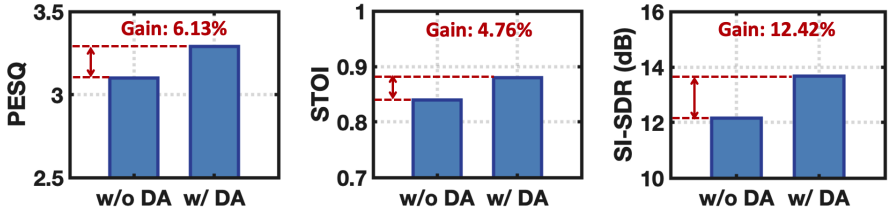


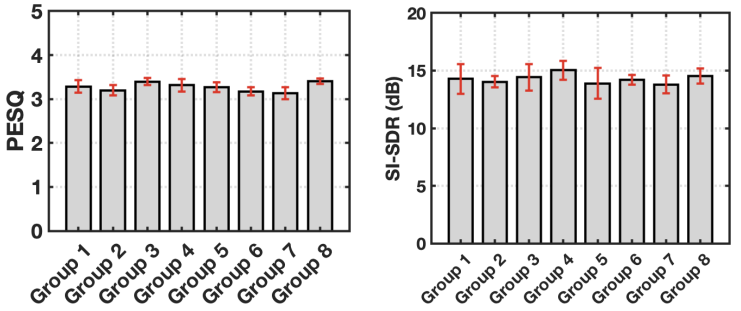
Fig. 17. Generalization among different languages.

8.3.2 ***Generalization Across Unseen Language.*** Next, we further investigate cross-language generalization by evaluating performance on languages that are not included during training. In our work, we select the Mandarin language, which has phonetic or tonal deviations from English. As illustrated in Fig. 17, QuaSE data augmentation scheme maintains competitive performance even under substantial phonetic and tonal mismatches. This demonstrates that the augmentation strategy facilitates learning language-agnostic acoustic cues, thereby enabling the system to operate effectively in cross-language deployment scenarios without requiring language-specific retraining.

8.3.3 ***Generalization Across User Group.*** The leave-one-group-out cross-validation is adopted for the evaluation. It ensures that all speech samples of users in the test dataset have not appeared in the training dataset before. Fig. 18 demonstrates the enhancement performance of QuaSE across different user groups. The slight differences of PESQ and SI-SDR between various user groups indicate that QuaSE also shows a great generalization capability among user groups. That is because the synthetic dual-channel dataset has included 60 speakers, making QuaSE learn user-independent features for speech enhancement.

8.4 Impact of Earphone Type

Different types of earphones may result in different degrees of seal in the ear canal, which directly brings different degrees of in-ear speech distortion. Thus, for the three earphone types, we assess the performance of QuaSE across different earphone types. As shown in Fig. 19, performance variations are observed due to differences in acoustic sealing, microphone placement, and resonance characteristics. However, compared with the baseline, QuaSE still shows outperforming performance. For the Earphone-2 type, EarSpeech and QuaSE show only a slight performance difference. This is because the earplugs of Earphone-2 do not form a closed space with the ear canal, so the air pressure inside the ear canal and the air pressure outside the ear canal remain balanced, which will not cause severe speech distortion. As for Earphone-1 and Earphone-3, we can observe that QuaSE can effectively deal with the issue of ECD-induced in-ear speech distortion.



(a) Airborne speech (b) In-ear speech

Fig. 18. Enhancement performance across user groups.

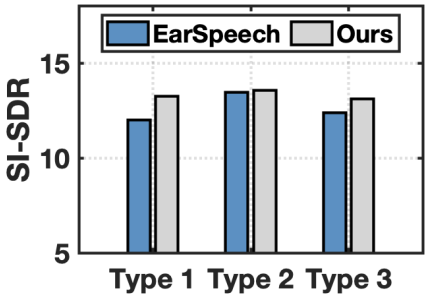


Fig. 19. Impact of earphone type.

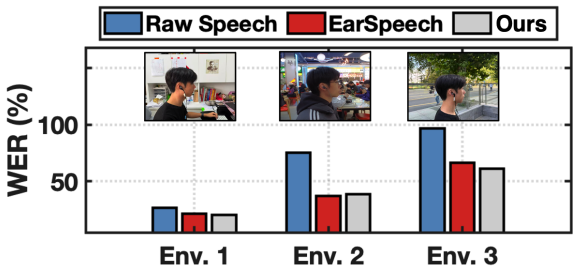


Fig. 20. WER of real-world scenarios.

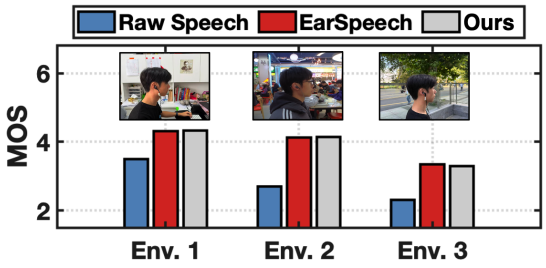


Fig. 21. MOS of real-world scenarios.

8.5 Real-world Study

Objective Measuring. In three typical voice interaction scenarios, a real-world study is conducted to assess the robustness of QuaSE. Specifically, three typical voice interaction scenarios include an office room with the noise level of 57.43 dB, a canteen with the noise level of 63.76 dB, and the outside street with the noise level of 75.12 dB. Since the noisy speech collected in the real scenarios cannot be accurately measured by objective evaluation metrics such as SNR and SDR, we use the Word Error Rate (WER) to quantify the enhancement performance of QuaSE. The enhanced airborne speech is converted into text with *Whisper Transcription GPT – AI Audio Converter*. As shown in Fig. 20, QuaSE reduces the WER across all three scenarios. In the more challenging canteen scenario with strong overlapping competitive speech, QuaSE still achieves a significant relative reduction of 35% in WER, demonstrating robustness against multi-speaker interference. Similarly, in the outdoor street scenario with dynamic noise sources such as traffic, QuaSE still lowers the WER by 30%, confirming its adaptability to highly non-stationary noise conditions. Since the Whisper Transcription GPT model is trained to be robust to ambient noise, several noisy speech samples can also be correctly identified.

Subjective Measuring. Next, we recruit 10 volunteers to rate the enhanced speech samples using the Mean Opinion Score (MOS) rating 0-5. A higher MOS reflects improved intelligibility and listening comfort, signifying that the enhancement method produces speech closer to natural human perception. The results are shown in Fig. 21, indicating that QuaSE has made great progress in both subjective and objective evaluation.

Table 2. Averaged run-time latency of 5-second speech clip on different platforms.

	GPU Platform	CPU Platform
Quality Adaptation	0.221 (± 0.172 s)	3.276 (± 0.081 s)
Encoder & Decoder	0.008 (± 0.017 s)	1.230 (± 0.071 s)
Overall Pipeline	0.217 (± 0.269 s)	3.974 (± 0.117 s)

8.6 Latency Evaluation

To assess the efficiency of the proposed model in practical deployment, we conduct inference latency measurements under different hardware platforms. Because the designed hardware prototype does not have computing capability, we leverage a client-server computing mode to measure inference latency, *i.e.*, measurement of the time from the end of earphone recording to the completion of speech enhancement. We input the 5-second airborne speech clip and the corresponding in-ear speech clip into QuaSE, repeat the calculation 200 times, and calculate the average latency. As summarized in Tab. 2, the average latency of the quality-aware adaptation module (Sec. 5.2) is higher than other modules regardless of CPU or GPU platforms. That is because it involves complex matrix operations to generate quality embeddings, as introduced in Sec. 5.2.2. After computational optimization, the inference latency of QuaSE only achieves 3.97 s with a real-time factor of 0.79 (< 1.00) even on the CPU platform. These results confirm that QuaSE meets real-time requirements, which is essential for latency-sensitive applications such as voice calls.

9 CONCLUSION

In our work, we identify that ear canal deformation induces air pressure fluctuations, degrading in-ear speech quality and limiting existing methods. To address that, we present QuaSE, a quality-aware dual-microphone speech enhancement framework that can dynamically fuse quality-varying in-ear speech and noisy airborne speech. Extensive experiments demonstrate that our work significantly improves the robustness and effectiveness of speech enhancement performance in real-world scenarios and provides meaningful guidance for future studies.

REFERENCES

[1] Takashi Amesaka, Hiroki Watanabe, and Masanori Sugimoto. 2019. Facial expression recognition using ear canal transfer function. In *Proceedings of the 2019 ACM International Symposium on Wearable Computers*. 1–9. https://doi.org/10.1145/3341163.3347747

[2] Toshiyuki Ando, Yuki Kubo, Buntarou Shizuki, and Shin Takahashi. 2017. Canalsense: Face-related movement recognition system based on sensing air pressure in ear canals. In *Proceedings of the 30th Annual ACM Symposium on User Interface Software and Technology*. Association for Computing Machinery, New York, NY, USA, 679–689. https://doi.org/10.1145/3126594.3126649

[3] Anderson R Avila, Hannes Gamper, Chandan Reddy, Ross Cutler, Ivan Tashev, and Johannes Gehrke. 2019. Non-intrusive speech quality assessment using neural networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 631–635. https://doi.org/10.1109/ICASSP.2019.8683175

[4] Kayla-Jade Butkow, Ting Dang, Andrea Ferlini, Dong Ma, and Cecilia Mascolo. 2023. hEARt: Motion-resilient Heart Rate Monitoring with In-ear Microphones. In *2023 IEEE International Conference on Pervasive Computing and Communications*. 200–209. https://doi.org/10.1109/PERCOM56429.2023.10099317

[5] Yetong Cao, Huijie Chen, Fan Li, and Yu Wang. 2021. CanalScan: Tongue-jaw movement recognition via ear canal deformation sensing. In *IEEE INFOCOM 2021-IEEE Conference on Computer Communications*. IEEE, 1–10. https://doi.org/10.1109/INFOCOM42981.2021.9488852

[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. https://doi.org/10.1121/10.0001237

[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. https://doi.org/10.1145/3498361.3538933

[8] Hu Chen, Yi Zhang, Mannudeep K Kalra, Feng Lin, Yang Chen, Peixi Liao, Jiliu Zhou, and Ge Wang. 2017. Low-dose CT with a residual encoder-decoder convolutional neural network. *IEEE transactions on medical imaging* 36, 12 (2017), 2524–2535. https://doi.org/10.1109/TMI.2017.2715284

[9] 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. https://doi.org/10.1109/ICASSP49357.2023.10094858

[10] Sune Darkner, Rasmus Larsen, and Rasmus R Paulsen. 2007. Analysis of deformation of the human ear and canal caused by mandibular movement. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 801–808.

[11] Alexandre Defossez, Gabriel Synnaeve, and Yossi Adi. 2020. Real Time Speech Enhancement in the Waveform Domain. arXiv:2006.12847 [eess.AS]

[12] Michel Demuynck, Aidin Delnavaz, and Jérémie Voix. 2024. Morphological analysis of the human earcanal deformations during face-related activities. *Applied Ergonomics* 116 (2024), 104195. https://doi.org/10.1016/j.apergo.2023.104195

[13] 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. https://doi.org/10.1145/3550303

[14] Di Duan, Yongliang Chen, Weitao Xu, and Tianxing Li. 2024. EarSE: Bringing Robust Speech Enhancement to COTS Headphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–33. https://doi.org/10.1145/3631447

[15] 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. https://doi.org/10.1145/3351239

[16] Feiyu Han, Panlong Yang, Yuanhao Feng, Haohua Du, and Xiang-Yang Li. 2024. Exploring Earable-Based Passive User Authentication via Interpretable In-Ear Breathing Biometrics. *IEEE Transactions on Mobile Computing* 23, 12 (2024), 15238–15255. https://doi.org/10.1109/TMC.2024.3453412

[17] 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. https://doi.org/10.1145/3659595

[18] 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. https://doi.org/10.1109/INFOCOM53939.2023.10229037

[19] Feiyu Han, Panlong Yang, You Zuo, Fei Shang, Fenglei Xu, and Xiang-Yang Li. 2024. Earspeech: Exploring in-ear occlusion effect on earphones for data-efficient airborne speech enhancement. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 3 (2024), 1–30. https://doi.org/10.1145/3678594

[20] Feiyu Han, You Zuo, Weiwei Jiang, Dawei Yan, Yuxin Zhao, Panlong Yang, and Yubo Yan. 2025. EAROE: Enabling Body-Channel Voice Interaction Interface on Earphones via Occlusion Effect. *IEEE Internet of Things Journal* (2025). https://doi.org/10.1109/JIOT.2025.3529912

[21] 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. https://doi.org/10.1145/3581791.3596832

[22] Jens Heitkaemper, Joe Caroselli, Max McKinnon, Arun Narayanan, and Nathan Howard. 2025. Bone Conducted Signal Guided Speech Enhancement For Voice Assistant on Earbuds. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5. https://doi.org/10.1109/ICASSP49660.2025.10889416

[23] Changshuo Hu, Thivya Kandappu, Yang Liu, Cecilia Mascolo, and Dong Ma. 2024. BreathPro: Monitoring breathing mode during running with earables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 2 (2024), 1–25. https://doi.org/10.1145/3659607

[24] Changshuo Hu, Xiao Ma, Xinger Huang, Yiran Shen, and Dong Ma. 2024. LR-Auth: Towards Practical Implementation of Implicit User Authentication on Earbuds. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8, 4 (2024), 1–27. https://doi.org/10.1145/3699793

[25] Yanxin Hu, Yun Liu, Shubo Lv, Mengtao Xing, Shimin Zhang, Yihui Fu, Jian Wu, Bihong Zhang, and Lei Xie. 2020. DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement. arXiv:2008.00264 [eess.AS]

[26] 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. https://doi.org/10.1145/3534613

[27] Ke Li, Ruidong Zhang, Bo Liang, François Guimbretière, and Cheng Zhang. 2022. Eario: A low-power acoustic sensing earable for continuously tracking detailed facial movements. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 2 (2022), 1–24. https://doi.org/10.1145/3534621

[28] 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. https://doi.org/10.1145/3485730.3485945

[29] Tiantian Liu, Feng Lin, Chao Wang, Chenhan Xu, Xiaoyu Zhang, Zhengxiong Li, Wenyao Xu, Ming-Chun Huang, and Kui Ren. 2023. WavoID: Robust and secure multi-modal user identification via mmWave-voice mechanism. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–15. https://doi.org/10.1145/3586183.3606775

[30] Dong Ma, Ting Dang, Ming Ding, and Rajesh Balan. 2024. ClearSpeech: Improving Voice Quality of Earbuds Using Both In-Ear and Out-Ear Microphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 7, 4 (2024), 1–25. https://doi.org/10.1145/3631409

[31] Pranay Manocha, Buye Xu, and Anurag Kumar. 2021. NORESQA: A framework for speech quality assessment using non-matching references. *Advances in neural information processing systems* 34 (2021), 22363–22378.

[32] Online. 2024. AS-B6027AL30-RC microphone. http://www.aospow.com/Products/znjqry6mmm.html. (Accessed on 04/21/2024).

[33] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, Min Wu, and KJ Ray Liu. 2023. Radio SES: mmWave-Based Audioradio Speech Enhancement and Separation System. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), 1333–1347. https://doi.org/10.1109/TASLP.2023.3250846

[34] 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. https://doi.org/10.1109/ICASSP.2015.7178964

[35] Michael Pedersen, Wouter Olthuis, and Piet Bergveld. 1997. A silicon condenser microphone with polyimide diaphragm and backplate. *Sensors and Actuators A: Physical* 63, 2 (1997), 97–104. https://doi.org/10.1016/S0924-4247(97)01532-X

[36] Karol J Piczak. 2015. ESC: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*. 1015–1018. https://doi.org/10.1145/2733373.2806390

[37] Tobias Röddiger, Christopher Clarke, Paula Breitling, Tim Schneegans, Haibin Zhao, Hans Gellersen, and Michael Beigl. 2022. Sensing with earables: A systematic literature review and taxonomy of phenomena. *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies* 6, 3 (2022), 1–57. https://doi.org/10.1145/3550314

[38] Kailai Shen, Diqun Yan, Jing Hu, and Zhe Ye. 2024. Non-intrusive speech quality assessment: A survey. *Neurocomputing* 580 (2024), 127471. https://doi.org/10.1016/j.neucom.2024.127471

[39] David Snyder, Guoguo Chen, and Daniel Povey. 2015. MUSAN: A Music, Speech, and Noise Corpus. arXiv:1510.08484 arXiv:1510.08484v1.

[40] Xingzhe Song, Kai Huang, and Wei Gao. 2022. FaceListener: Recognizing Human Facial Expressions via Acoustic Sensing on Commodity Headphones. In *2022 21st ACM/IEEE International Conference on Information Processing in Sensor Networks (IPSN)*. 145–157. https://doi.org/10.1109/IPSN54338.2022.00019

[41] 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*. Association for Computing Machinery, New York, NY, USA, 160–173. https://doi.org/10.1145/3447993.3448626

[42] Xue Sun, Jie Xiong, Chao Feng, Wenwen Deng, Xudong Wei, Dingyi Fang, and Xiaojiang Chen. 2023. Earmonitor: In-ear motion-resilient acoustic sensing using commodity earphones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 6, 4 (2023), 1–22. https://doi.org/10.1145/3569472

[43] Xue Sun, Jie Xiong, Chao Feng, Haoyu Li, Yuli Wu, Dingyi Fang, and Xiaojiang Chen. 2024. EarSSR: Silent Speech Recognition via Earphones. *IEEE Transactions on Mobile Computing* 23, 8 (2024), 8493–8507. https://doi.org/10.1109/TMC.2024.3356719

[44] Zi Wang, Sheng Tan, Linghan Zhang, Yili Ren, Zhi Wang, and Jie Yang. 2021. EarDynamic: An ear canal deformation based continuous user authentication using in-ear wearables. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–27. https://doi.org/10.1145/3448098

[45] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. 2018. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*. 3–19. https://doi.org/10.1007/978-3-030-01234-2_1

[46] Yixuan Wu, Jintai Chen, Lianting Hu, Hongxia Xu, Huiying Liang, and Jian Wu. 2025. OmniFuse: A general modality fusion framework for multi-modality learning on low-quality medical data. *Information Fusion* 117 (2025), 102890. https://doi.org/j.inffus.2024.102890

[47] 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. https://doi.org/10.1609/aaai.v34i05.6489

[48] Guoming Zhang, Chen Yan, Xiaoyu Ji, Tianchen Zhang, Taimin Zhang, and Wenyuan Xu. 2017. Dolphinattack: Inaudible voice commands. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*. Association for Computing Machinery, New York, NY, USA, 103–117. https://doi.org/10.1145/3133956.3134052

[49] Qian Zhang, Kaiyi Guo, Yifei Yang, and Dong Wang. 2025. WearSE: Enabling Streaming Speech Enhancement on Eyewear Using Acoustic Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 9, 1 (2025), 1–30. https://doi.org/10.1145/3712288

[50] Qingyang Zhang, Yake Wei, Zongbo Han, Huazhu Fu, Xi Peng, Cheng Deng, Qinghua Hu, Cai Xu, Jie Wen, Di Hu, et al. 2024. Multimodal fusion on low-quality data: A comprehensive survey. *arXiv preprint arXiv:2404.18947* (2024). https://doi.org/arXiv:2404.18947

[51] Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. 2023. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*. PMLR, 41753–41769. https://doi.org/10.5555/3618408.3620161

[52] Shijia Zhang, Taiting Lu, Hao Zhou, Yilin Liu, Runze Liu, and Mahanth Gowda. 2023. I Am an Earphone and I Can Hear My User's Face: Facial Landmark Tracking Using Smart Earphones. *ACM Transactions on Internet of Things* 5, 1 (2023), 1–29. https://doi.org/10.1145/3614438

[53] Peizhao Zhu, Yuzheng Zhu, Wenyuan Li, Yanbo He, Yongpan Zou, Kaishun Wu, and Victor C. M. Leung. 2025. CHAR: Composite Head-Body Activities Recognition With a Single Earable Device. *IEEE Transactions on Mobile Computing* 24, 7 (2025), 6532–6549. https://doi.org/10.1109/TMC.2025.3548647

[54] Yongpan Zou, Jianhao Weng, Haibo Lei, Danyang Wang, Victor C. M. Leung, and Kaishun Wu. 2024. EarPrint: Earphone-Based Implicit User Authentication With Behavioural and Physiological Acoustics. *IEEE Internet of Things Journal* (2024), 1–1. https://doi.org/10.1109/JIOT.2024.3417622