LargeCall: Large-Model-Assisted Phone Call Enhancement Using Smartphone's Built-in Accelerometer

***Abstract*—Achieving intelligible and noise-robust voice communication during phone calls remains a significant challenge in real-world environments, where low signal-to-noise (SNR) ratios and background conversations are prevalent. While traditional audio-only speech enhancement systems show strong performance under moderate noise, they struggle in multi-speaker or low-SNR conditions due to the lack of user-specific cues. In this work, we propose LargeCall, a dual-modality speech enhancement framework that integrates microphone audio with inertial signals from the smartphone's built-in accelerometer. Unlike visual or ultrasound-based methods, accelerometer sensing is passive, privacy-preserving, and robust in outdoor settings. Our design addresses three key challenges: (i) the scarcity of paired audio-accelerometer data is mitigated by bootstrapping from a pretrained audio-only encoder, reducing training data requirements and improving generalization; (ii) variability in phone orientation and body motion is handled through a posture-invariant transformation pipeline and a multi-dilated fusion module for robust feature extraction; and (iii) modality mismatch between audio and accelerometer streams is resolved via a cross-modal mask fusion strategy, which adaptively integrates complementary masks from both modalities without disrupting the pretrained encoder. We validate LargeCall on a cross-user evaluation protocol using realistic phone call scenarios. Experimental results show that LargeCall achieves substantial gains in both objective and subjective metrics across diverse noise conditions, demonstrating its effectiveness and real-world deployment potential. The demo of our system is available at https://anonymousers718.github.io/largecall/.**

***Index Terms*—Speech enhancement, Accelerometer, Multi-modality**

I. Introduction

Phone calls are one of the most essential and widely used methods of communication, playing a vital role in everyday communications and emergency situations due to their interactivity and immediacy. Recent statistics indicate that billions of phone calls occur globally every day, highlighting their pervasive significance [1]. Compared to other voice-based communications such as voice assistants [2] or voice messaging [3], phone calls offer immediate feedback, emotional cues, and efficient real-time interaction, making them especially valuable in urgent or complex scenarios. However, achieving clear and effective voice communication during phone calls is challenging in real-world scenarios such as noisy streets, crowded cafes, and busy public transportation stations [4], [5]. Environmental noises such as traffic, machinery, background conversations, and music severely degrade speech intelligibility in these scenarios. Additionally, users often speak softly in public to maintain privacy or avoid disturbing others, which further reduces signal intensity and leads to lower signal-to-noise ratios (SNR) [6]. Therefore, robust speech enhancement technologies that specifically address these complex and noisy phone call environments are urgently required.

Fig. 1: Real-world usage scenario of LargeCall.

Traditional speech enhancement methods primarily rely on single-modality audio data, using techniques such as spectral subtraction, Wiener filtering, and more recently, deep learning-based models like Conv-TasNet [7] and PHASEN [8]. While effective under moderate noise, they often fail in low-SNR or multi-speaker scenarios. A key limitation of single-modality methods is that, based solely on acoustic input, it is inherently difficult to distinguish the target speaker from interfering voices. Moreover, many speaker separation methods require the number of speakers to be predefined [9], which is impractical in real-world phone call scenarios where the number of interfering speakers is unknown and may vary dynamically. This makes source separation particularly challenging in realistic phone call scenarios, where background conversations are common. These limitations motivate the need for more robust methods that can incorporate complementary, noise-resilient information beyond the audio channel.

To overcome these limitations, recent research has explored multi-modal speech enhancement. Audio-visual approaches combine microphone input with lip movements [10], while ultrasound-based approaches track articulatory gestures using inaudible signals [11]. Despite their promise, both methods face practical obstacles. Visual systems raise privacy concerns and require front-facing cameras, which are rarely feasible during phone calls. Ultrasound-based systems demand active signal emission and are sensitive to wind interference [12], limiting outdoor reliability.

These constraints motivate us to seek an alternative modality that can robustly address these practical challenges. In this

paper, we propose a speech enhancement system that integrates microphone audio with the smartphone's built-in accelerometer to improve voice clarity during phone calls in noisy and complex environments. This cross-modal design allows the system to better isolate the user's speech compared to audio-only approaches. It offers several advantages for real-world use. Unlike visual or ultrasound-based methods, the accelerometer works passively without emitting signals or capturing video, avoiding privacy issues and reducing power and hardware demands. It is also inherently low-power and resistant to airflow, making it reliable in outdoor conditions. During calls, the device is typically in contact with the user's cheek, allowing the accelerometer to capture speech-induced vibrations transmitted through bone and skin. These vibration signals are highly correlated with the user's own speech and largely immune to environmental noise or interfering speakers. By integrating this user-specific signal with audio input, this system achieves more robust and accurate speech enhancement, particularly under low-SNR and multi-speaker conditions. We envision a future in which users, whether indoors or outdoors, can speak naturally during phone calls while the listener on the other end hears clean, uninterrupted speech, free from environmental interference, as illustrated in Fig. 1.

While integrating speech and accelerometer signals offers clear advantages, it also introduces several key challenges.

The first challenge concerns how to effectively train a dual-modality model under limited availability of paired audio-accelerometer data. Most deep learning-based speech enhancement systems rely on large-scale training corpora to generalize across speakers, environments, and noise types [13]–[15]. However, collecting synchronized audio and inertial signals at such scale is impractical in real-world scenarios. While large-scale audio-only and audio-visual datasets are widely available, datasets with aligned audio-accelerometer recordings remain extremely scarce, limiting their use in dual-modality learning. To address this, we build on a key insight: audio-only pretrained models already exhibit strong denoising capabilities (as shown in Section II-A). Instead of training the entire system from scratch, we bootstrap our framework using a pretrained audio encoder. By freezing the encoder and training only the accelerometer and fusion modules, we significantly reduce the number of trainable parameters and mitigate overfitting risks under limited supervision (Section III-B).

The second challenge lies in the variability of accelerometer signals caused by changes in phone orientation and user motion during calls. Users may hold the phone at different angles, adjust its position while speaking, or engage in activities such as walking or head movement. As shown in preliminary study (Sec. II-B), these variations lead to inconsistent signal distributions across the three sensing axes, complicating the extraction of stable and speaker-relevant articulatory features. To address this issue, we first normalize accelerometer signals by rotating them into a global reference frame using orientation estimates from the Madgwick filter. We further apply a high-pass filter to suppress low-frequency motion artifacts, retaining only speech-induced vibrations. In addition, we propose the Multi-Dilated Articulatory Fusion Module (Section III-C) to extract robust inertial features across varying temporal scales and axes. Together, these components enhance the model's ability to capture reliable articulatory patterns under diverse phone-holding postures and motion conditions.

The third challenge lies in effectively fusing speech and accelerometer signals without introducing modality interference. As demonstrated in the preliminary study (Section II), the speech stream provides rich spectral information for denoising, while the accelerometer stream offers user-specific vibration patterns that enhance speaker discrimination. However, directly combining features from the two modalities can lead to degraded performance due to distribution mismatch. This mismatch arises because the speech encoder is pretrained on large-scale corpora, whereas the accelerometer encoder is trained from scratch, making joint learning unstable and less effective. To address this, we adopt a late fusion strategy via the proposed Cross-Modal Mask Fusion Module (Section III-C). This module allows each stream to generate a semantic mask aligned with its strength and dynamically weights the contributions of each based on contextual reliability. This design ensures that the speech encoder retains its pretrained denoising capacity while effectively incorporating speaker-related cues from the accelerometer.

Our contributions are summarized as follows:

- We propose a dual-modality speech enhancement framework that leverages a pretrained audio-only model to address the challenge of limited paired audio-accelerometer data. This transfer learning strategy enables data-efficient training while maintaining strong generalization capability.
- We design an orientation-agnostic accelerometer encoder that addresses variability in phone placement and body motion. The module includes rotation normalization, multi-axis fusion, and motion-denoising to ensure robust feature extraction under diverse real-world conditions.
- We introduce a novel mask-based fusion mechanism that integrates speaker-discriminative accelerometer cues with denoising masks from the speech. This design preserves the structure of the pretrained encoder while enhancing the model's ability to separate the target speaker from interfering voices.
- We conduct extensive experiments under a cross-user evaluation protocol, in which training and testing users are disjoint. Results demonstrate strong generalization to unseen speakers and consistent performance across diverse environmental conditions, validating the effectiveness of our design.

II. Preliminary Study

A. Effectiveness and Limitations of Pretrained Model

To assess the feasibility of leveraging pretrained audio-only models for speech enhancement, we conduct a comparative evaluation across several state-of-the-art (SOTA) models. As shown in Table I, MossFormerGAN [15] consistently outperforms other large-scale models, exhibiting clear advantages in both speech quality and intelligibility. Based on this observation, we adopt MossFormerGAN as the core audio enhancement

Model	SISNR	STOI	PESQ	MOSSIG	MOSBACK
SepFormer [13]	-1.42	0.12	-0.04	0.33	0.74
FRCRN [14]	0.77	0.02	0.01	0.12	0.11
MossFormer2 [15]	3.34	0.27	0.06	0.55	0.98
MossFormerGAN [15]	**5.53**	**0.59**	**0.09**	**0.77**	**1.21**

TABLE I: Pretrained model enhancement comparison.

backbone in our system. Metric definitions are provided in Sec. IV-B.

To further understand the limitations of MossFormerGAN backbone under realistic conditions, we perform stress testing across varying levels of speech interference. We conduct controlled experiments using synthetically generated mixtures of clean speech with both ambient noise and interfering speakers. The testing environment includes three progressively complex noise configurations: "1S+A", "2S+A", and "3S+A", where "S" denotes interfering speakers (excluding the target speaker), and "A" represents ambient noise. This setup enables a structured analysis of model robustness under increasing levels of speech interference in realistic acoustic scenarios. For each configuration, we construct 100 mixture samples, each 5 seconds in duration. After processing all samples through the pretrained model, we compute the mean values of three key objective metrics: MOSSIG (speech quality), MOSBAK (background noise intrusiveness), and SI-SNR (signal-to-interference ratio). Detailed metric definitions are provided in Sec. IV-B.

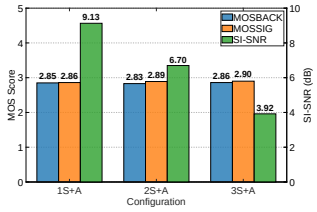

Fig. 2: Pretrained model performance vs. noise settings.

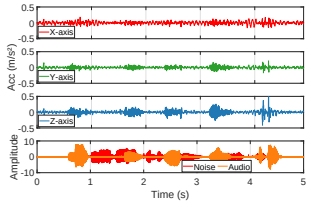

Fig. 3: Temporal alignment of accelerometer and speech.

Fig.2 shows that while remain stable across all interference levels, SI-SNR degrades significantly as more interfering speakers are introduced. This observation highlights a key limitation: **pretrained audio-only models excel at suppressing background noise but struggle to disentangle competing speakers in multi-speaker conditions.** This motivates our dual-modality design, which introduces auxiliary signals to enhance target speaker separation while preserving the denoising strength of the pretrained model.

B. Speech-Coupled Properties of Accelerometer Signals

During phone calls, smartphones typically maintain contact with the user's cheek and jaw. Movements of articulatory organs such as the jaw, lips, and tongue, along with vibrations from the vocal cords, are transmitted through bone and soft tissue to the built-in accelerometer. This provides a unique opportunity to capture speech-related motion passively and reliably during real-world phone call scenarios. To assess the suitability of accelerometer signals for identifying target speech, we examine their temporal alignment with acoustic speech and their spatial distribution across sensing axes. We conduct a case study in which a participant holds a Samsung Galaxy S10 in a typical phone call posture and speaks the phrase "nice to meet you today". We simultaneously record microphone audio and accelerometer signals. As shown in Fig.3, After applying a high-pass filter, the accelerometer signals exhibit strong temporal alignment with the target speaker's speech waveform, while showing no correlation with interfering speech, confirming that speech-related motion is uniquely and consistently encoded in the inertial data stream. Notably, the articulatory motion is distributed across multiple axes. In this example, the Y and Z axes carry the dominant components, while the X-axis contribution is relatively weak. This variation reflects natural differences in phone orientation during use. Consequently, a key challenge arises: **how to effectively fuse multi-axis accelerometer signals to extract stable, speaker-specific features across diverse phone-holding postures**.

III. LargeCall Design

A. LargeCall Architecture Design

We propose LargeCall, a dual-modality speech enhancement framework that leverages both audio and accelerometer inputs to generate a fused mask for isolating the target speaker while suppressing background noise. The overall architecture is illustrated in Fig. 4.

The input consists of a raw speech waveform and a 3-axis accelerometer signal. The speech signal is transformed into a complex-valued spectrogram $\mathbf{S}_\text{s} \in \mathbb{C}^{T\times F}$ via short-time Fourier transform (STFT). In parallel, the accelerometer signal is first denoised using a high-pass filter and then converted into its spectral representation $\mathbf{S}_\text{a}$ via STFT.

The speech branch fed $\mathbf{S}_\text{s}$ into a pretrained encoder (detailed in Sec. III-B) to extract speech features $\mathbf{F}_s$, which are passed to a mask generator to predict the denoising mask $\mathbf{M}_s \in [0,1]^{T\times F}$. In parallel, the accelerometer branch applies a multi-scale fusion module (detailed in Sec. III-C) on the inertial spectrogram and generates speaker-specific features $\mathbf{F}_a$, which, along with $\mathbf{F}_s$, are used to predict a speaker-specific mask $\mathbf{M}_a \in [0,1]^{T\times F}$.

To effectively integrate the complementary cues from both modalities, we introduce the Cross-modal Mask Fusion Module (Sec. III-C). This module adaptively merges $\mathbf{M}_s$ and $\mathbf{M}_a$ to produce the final fused mask $\mathbf{M}_\text{fuse}$, which captures both noise suppression and speaker specificity.

We adopt a residual enhancement structure for final output. The enhanced magnitude spectrogram is computed as:

$$\mathbf{S}_\text{out} = |\mathbf{S}_\text{s}| \circ \mathbf{M}_\text{fuse} + \text{Decoder}(\mathbf{F}_s), \tag{1}$$

where $\circ$ denotes element-wise multiplication. The decoder refines the residual components and reconstructs the enhanced complex spectrogram.

Finally, the enhanced spectrogram $\mathbf{S}_\text{out}$ is passed through the inverse STFT to reconstruct the time-domain waveform. The model is trained using a joint loss function that combines spectrogram-domain mean squared error $\mathcal{L}_\text{mse}$, which ensure

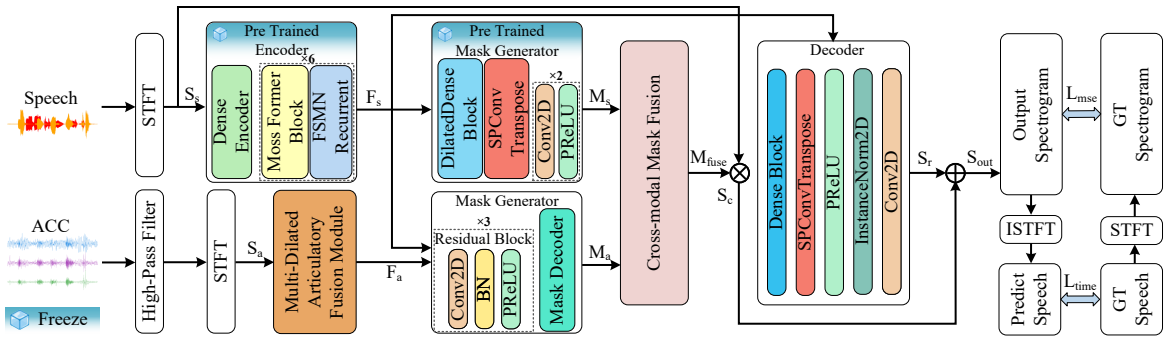

Fig. 4: Architecture of LargeCall.

accurate frequency-domain reconstruction, and time-domain loss $\mathcal{L}_{\text{time}}$ [16], which preserves temporal consistency.

B. Bootstrapping Dual-Modal Models with Pretrained Audio Backbone

Training a dual-modality speech enhancement model from scratch typically demands large-scale paired audio-accelerometer datasets [17], which are expensive to collect and often limited in diversity. For instance, building a well-generalized audio-only model such as our adopted pretrained model typically requires millions of training samples [15]. Collecting paired audio and accelerometer data at this scale is highly impractical in real-world settings . Moreover, such data must be precisely synchronized across modalities to ensure temporal alignment, adding further complexity to the data collection process.

To address this, we build on a key insight: audio-only speech enhancement is well studied, and many pretrained models already demonstrate strong denoising performance. Rather than training a dual-modality system from scratch, we ask: *why not build on these proven audio foundations to support multi-modal learning?* By using a pretrained audio encoder as the backbone of the audio stream, we inherit robust speech representations learned from large-scale audio data. This approach offers two key benefit. First, leveraging audio models pretrained on large-scale corpora yields robust and generalizable speech representations across diverse acoustic conditions. Second, freezing the audio encoder reduces the number of trainable parameters, thereby mitigating overfitting and enhancing training stability in limited data scenarios.

Building on our findings in the preliminary study (Sec. II), which confirmed the strong denoising capability of pretrained audio-only models, we adopt MossFormerGAN as the audio backbone in LargeCall. We integrate two key components from MossFormerGAN [16]: the encoder and the mask generator. The encoder consists of a Dense Encoder followed by six stacked layers, where each layer comprises a MossFormer block and a FSMN-based recurrent unit. Given the input speech spectrogram $\mathbf{S}_s$, the encoder extracts high-level audio features $\mathbf{F}_s = \text{Encoder}(\mathbf{S}_s)$.

The extracted features $\mathbf{F}_s$ are then fed into the pretrained mask generator, which is composed of a Dilated Dense Block and a Sub-Pixel Convolution Transpose module. The generator predicts a denoising mask as $\mathbf{M}_s = \text{MaskGenerator}(\mathbf{F}_s)$. Both $\mathbf{F}_s$ and $\mathbf{M}_s$ are passed downstream as the audio branch outputs, preserving the denoising capability of the pretrained model while enabling cross-modal integration with the accelerometer stream through the Mask Fusion Module.

C. Posture-Invariant Accelerometer Encoder

Posture-Invariant Transformation: In real-world usage, users hold their phones at various orientations during calls, introducing substantial variability in the distribution of accelerometer signals across the three axes. Such inconsistencies increase the burden on the model, requiring it to learn orientation-invariant representations from limited training data. To improve generalization and reduce data requirements, we explicitly normalize the sensor input through a posture-invariant transformation.

The key idea is to rotate the raw tri-axial accelerometer data into a unified global reference frame. This allows the model to consistently interpret the directional components of articulatory motion, regardless of how the phone is held. We achieve this by estimating the device's orientation in real-time using a quaternion-based algorithm and applying a corresponding rotation to the accelerometer vectors.

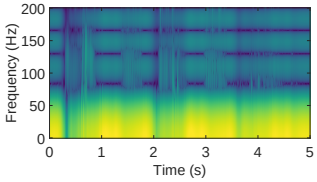

Fig. 5: Spectrogram without filtering.

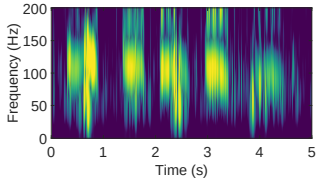

Fig. 6: Spectrogram with filtering.

To estimate the device's orientation, we adopt the Madgwick filter [18], a computationally efficient AHRS (Attitude and Heading Reference System) algorithm that fuses accelerometer and gyroscope data. Let $\mathbf{a}_t$ and $\mathbf{g}_t$ denote the accelerometer and gyroscope readings at time t, respectively. The orientation quaternion $\mathbf{q}_t$ is updated as:

$$\frac{d\mathbf{q}}{dt} = \frac{1}{2}\mathbf{q}_t \otimes \begin{bmatrix} 0 \\ \mathbf{g}_t \end{bmatrix} - \beta \nabla f(\mathbf{q}_t, \mathbf{a}_t), \tag{2}$$

where $\otimes$ denotes quaternion multiplication, $\nabla f(\cdot)$ is the gradient of an objective function enforcing alignment between

gravity and the estimated orientation, and β is a tunable filter gain that controls the trade-off between responsiveness and noise rejection. In practice, we empirically tune the parameter $\beta = 0.02$ to balance convergence speed and numerical stability.

The corrected global-frame accelerometer reading $\mathbf{a}_t^{\mathrm{g}}$ is obtained by applying the inverse quaternion rotation:

$$\mathbf{a}_t^{\mathrm{g}} = \mathbf{q}_t^{-1} \otimes \mathbf{a}_t \otimes \mathbf{q}_t. \tag{3}$$

This transformation significantly reduces variability across users and sessions, enabling the network to extract more consistent articulatory features without requiring the model to learn rotational invariance from scratch.

High-Pass Filter Denoising Accelerometers are highly sensitive to various body motions such as walking, nodding, or shifting the phone during a call. These low-frequency motion artifacts can obscure speech-induced vibrations, making it difficult to extract clean articulatory features in the spectrogram. As illustrated in Fig. 5, which shows the STFT spectrogram of a sample where speech and walking occur simultaneously, the motion-induced noise dominates the low-frequency bands and masks the speech-relevant patterns.

Prior studies have shown that the majority of body-induced motion energy is concentrated below 50Hz [19], while the fundamental frequency components of speech typically reside above this threshold. To suppress motion noise while retaining speech-relevant signals, we apply a high-pass filter to the accelerometer data. Rather than employing computationally intensive high-order filters such as FIR filters [20], Butterworth filter [21], and high-pass IIR filters [22], we adopt a low-order Biquad high-pass filter. This design uses coefficients and provides a sharp cutoff response with low computational overhead, making it suitable for mobile deployment.We set the cutoff frequency to 50Hz. The resulting spectrogram, shown in Fig. 6, reveals that speech-induced vibration patterns become substantially more prominent after filtering. This simple yet effective denoising step enhances the quality of inertial features and reduces the burden on accelerometer encoder.

Multi-Dilated Articulatory Fusion Module: To extract robust, posture-invariant vibration features from the accelerometer stream, we introduce the Multi-Dilated Articulatory Fusion Module, which comprises two key components: the Multi-Dilated Articulatory Motion Extractor and the Scale-Aware Fusion Module, as shown in Fig. 7.

The design is motivated by the observation that articulatory gestures vary significantly in temporal duration, ranging from 100 ms to 700 ms [23]. A fixed temporal window cannot adequately capture such diverse dynamics, resulting in either poor frequency resolution for slow gestures or loss of detail for fast transitions. To address this, the extractor applies parallel 2D convolutions with multiple dilation rates ($d = 1, 2, 4$) to model articulatory patterns at different temporal scales. Each path is followed by BatchNorm (BN) and Rectified Linear Unit (ReLU) activation, and enhanced with an Efficient Channel Attention (ECA) module that reweights channel activations based on

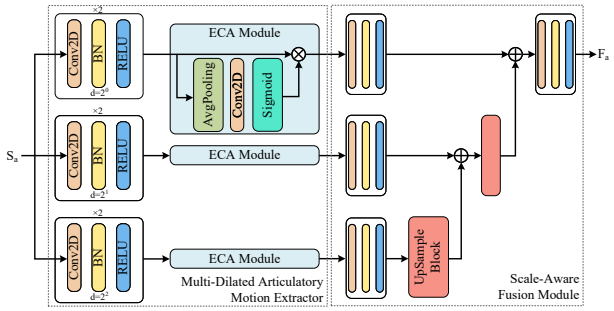

Fig. 7: Architecture of Acc-Encoder.

global average pooled statistics, promoting discriminative feature encoding across all three accelerometer axes.

The resulting multi-scale features are fed into the Scale-Aware Fusion Module, where each branch is projected to a common embedding via a convolution layer, then combined through residual summation. A final convolution block outputs the unified feature representation $\mathbf{F}_{\mathrm{a}}$, which encodes both fine-grained and coarse articulatory dynamics in a compact form.

This architecture enables the model to capture speaker-specific vibration cues across different motion durations and orientations, while maintaining high computational efficiency. It also improves downstream mask generation by providing temporally rich and noise-resilient accelerometer features.

D. Cross-Modal Mask Fusion

Most existing multi-modal speech enhancement approaches perform fusion at the feature level, where audio and auxiliary modality embeddings are concatenated or jointly encoded before mask generation [24], [25]. While this strategy allows early interaction between modalities, it often blurs modality-specific representations and introduces interference between distinct signal characteristics. This is particularly problematic in our setting, where we adopt a pretrained audio backbone whose internal structure has already been optimized for denoising.

To preserve modality-specific advantages, we propose a cross-modal mask fusion strategy in which each modality independently predicts a semantic mask aligned with its respective strengths: the audio stream produces a denoising mask $\mathbf{M}_s$, while the accelerometer stream generates a speaker-specific mask $\mathbf{M}_a$. Effectively combining these complementary cues is critical to overall system performance. As shown in the Fig. 4, the final fused mask is applied directly to the spectrogram $\mathbf{S}_{\mathrm{s}}$ to reconstruct the enhanced output. To this end, we design a lightweight Mask Fusion Module that adaptively weights the two masks based on contextual reliability, as illustrated in Fig. 8.

The two masks are first concatenated along the channel dimension and passed through a convolutional block, batch normalization, activation, and a sigmoid layer. This produces a pair of dynamic weights:

$$W_s, W_a = \mathrm{Sigmoid}\left(\mathrm{Conv2D}\left(\mathrm{Concat}(\mathbf{M}_s, \mathbf{M}_a)\right)\right), \tag{4}$$

where $W_s, W_a \in (0, 1)$ represent the learned confidence scores for each modality.

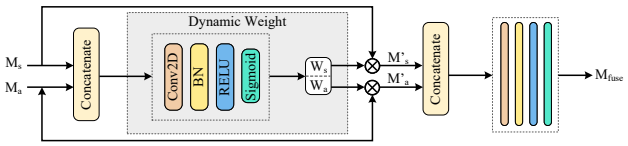

Fig. 8: Mask-fusion module.

These weights are then used to reweight the original masks:

$$\mathbf{M}'_s = W_s \cdot \mathbf{M}_s, \quad \mathbf{M}'_a = W_a \cdot \mathbf{M}_a, \tag{5}$$

and the weighted masks are concatenated and projected via another convolution to produce the final fused mask:

$$\mathbf{M}_{\text{fuse}} = \text{Conv}\left(\text{Concat}(\mathbf{M}'_s, \mathbf{M}'_a)\right). \tag{6}$$

This formulation allows the system to emphasize the modality that is more reliable under varying acoustic or motion conditions, thereby ensuring robust speech enhancement across diverse environments.

IV. Experiment Setup

A. Experimental Setup

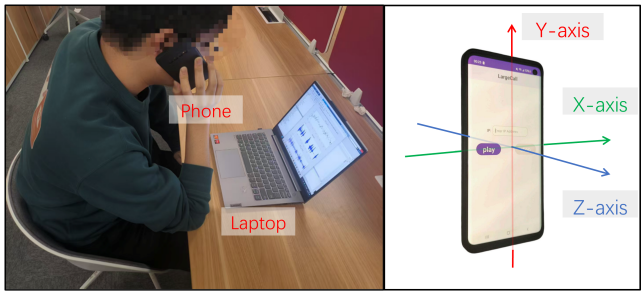

Fig. 9: Experiment scenario.

Data Collecting: We recruited 20 university students (4 females and 16 males) with an average age of 24. Among them, 3 are native English speakers of Caucasoid descent, while the remaining participants are non-native English speakers of Mongoloid descent. All participants provided informed consent, and we have obtained ethical approval from our institutional review board (IRB). Each participant was asked to read 200 sentences from the TIMIT corpus [26], with each sentence lasting approximately 5 seconds. Recordings were made in a quiet indoor environment using a Samsung Galaxy S10 held in phone-call position (see Fig. 9). To increase variability, participants were instructed to hold the phone at various angles during the recording sessions. To simulate real-world noise conditions, we synthesized noisy speech by mixing the clean recordings with interfering speech and ambient noise. Interfering speech was sampled from the TIMIT corpus [26], which includes 6,300 utterances from 630 speakers. Ambient noise was drawn from AudioSet [27], a large-scale dataset containing over 3.4 million 5-second clips spanning environmental sounds, machines, music, and human activities. Each clean utterance was mixed with 20 distinct noise conditions. For each condition, the number of interfering speakers was randomly selected between 1 and 4. Both interfering speech and ambient noise samples were randomly chosen for each mixture to avoid overfitting. The SNR was uniformly distributed between -5 dB and 10 dB, with an average of 2.5 dB.

Implementation: For data collection, we develop an app called AccRecord on the mobile phone to capture multi-modal data, including audio and accelerometer readings, with sampling rates of 16 kHz and 400 Hz, respectively, as shown in Fig. 9. The collected multi-modal data is then transmitted to a Lenovo ThinkBook 14 via WiFi for further processing. We implement our LargeCall model using PyTorch. We train LargeCall on an in-house server equipped with a single NVIDIA A100 GPU and 128 GB of system memory. After training, the model is deployed on the same server to run a real-time speech enhancement service. The system continuously receives noisy speech and accelerometer data as input, performs enhancement using LargeCall, and transmits the enhanced audio to the recipient. In our benchmarking, each 1-second input window requires approximately 7 ms for data transfer and 35 ms for model inference. These results demonstrate that LargeCall can be efficiently deployed on a server, enabling low-latency, real-time speech enhancement in practical applications.

Model Training/Testing: Each training and testing speech sample consists of three parts: clean speech, noisy speech, and the corresponding accelerometer data. All experiments are performed under a cross-user setting. Unless otherwise specified, every subject contributes to all experimental sessions, and system performance is assessed using 4-fold cross-validation. Specifically, given n subjects, we partition the data into four folds and repeat the evaluation four times. In each fold, data from $\frac{3n}{4}$ subjects are used for training, and data from the remaining $\frac{n}{4}$ subjects are used for testing.

B. Evaluation Metrics

We evaluate the performance of LargeCall using the following six commonly used metrics:

- SISNR [7]: SISNR is a scale-invariant version of SNR that removes the effect of amplitude differences by aligning the estimated signal with the reference. It focuses purely on signal quality, regardless of volume or scaling.
- STOI (Short-Time Objective Intelligibility) [28]: STOI measures speech intelligibility on a scale from 0 to 1, where 1 means perfectly clear speech and 0 means completely unintelligible.
- PESQ (Perceptual Evaluation of Speech Quality) [29]: PESQ estimates speech quality based on human hearing, with scores from 1 (poor) to 5 (excellent).
- DNSMOS (Deep Noise Suppression Mean Opinion Score) [30]: DNSMOS is a non-intrusive, deep learning-based metric that predicts subjective speech quality. It provides three scores: speech quality (SIG), noise intrusiveness (BAK), and overall quality (OVRL). We report the overall quality score (OVRL), also referred to as MOS, as the primary subjective metric.
- WER (Word Error Rate) [31]: We use the Whisper speech recognition model [32] to convert speech into text for evaluating

speech intelligibility. We compute the word error rate (WER) as: $WER = \frac{N_S+N_D+N_I}{N_R}$, where N_S, N_D, and N_I are the numbers of substitutions, deletions, and insertions, and N_R is the total number of words in the reference. A lower WER indicates better intelligibility. A lower WER indicates higher speech intelligibility.

C. Overall Performance

1) Speech Enhancement Model Comparison: We compare the performance of LargeCall against several SOTA speech enhancement models under diverse noisy conditions, using the same evaluation protocol described in Section IV-A. The test set includes four noise configurations: "1S+A", "2S+A", "3S+A", "4S+A", where "S" refers to interfering speakers (excluding the target speaker), and "A" denotes ambient noise. The SNR of the input is uniformly sampled from the range of $[-5, 10]$ dB. The average SNR for these configurations are 5, 2.5, 0, and -2.5 dB, respectively. We compare LargeCall with four SOTA baselines: VibVoice [33] and SEANet [24], which are multi-modal accelerometer-assisted models. SepFormer [13] and PHASEN [8], which are audio-only enhancement models. To ensure a fair comparison, all baselines are re-implemented and trained using the same settings as LargeCall. Since SepFormer and PHASEN are designed solely for denoising (not separation), they are trained only on the "1S+A" dataset. VibVoice and SEANet follow the full training protocol used for LargeCall.

The results are summarized in Table II. Across all four conditions, LargeCall consistently achieves superior performance. For instance, in the "1S+A" scenario, LargeCall obtains 13.61 dB SISNR, 0.87 STOI, 3.11 PESQ, and 2.00 MOS, outperforming SepFormer by a substantial margin. As the number of interfering speakers increases, the advantage of using inertial signals becomes more evident. In these low-SNR conditions, LargeCall leverages the posture-invariant and speaker-specific accelerometer stream to retain intelligibility where audio-only models degrade significantly. Furthermore, LargeCall outperforms both prior accelerometer-assisted methods (VibVoice and SEANet) and audio-only baselines across all objective and subjective metrics. On average, LargeCall achieves 10.96 dB SISNR and 1.93 MOS, compared to 7.34 dB and 1.79 from SepFormer, and 5.31 dB and 1.64 from PHASEN. This performance gain is primarily attributed to the system's architectural design, which includes a pretrained audio backbone, a multi-dilated articulatory encoder, and a dedicated cross-modal mask fusion module. Lastly, we evaluate the pretrained model MossFormerGAN without any fine-tuning. While it performs well in terms of MOS, its SISNR is significantly lower than other audio-only baselines. This discrepancy is expected, as the model was primarily optimized for subjective quality rather than objective metrics.

2) Performance of LargeCall: We further evaluate the overall enhancement capability of LargeCall under "2S+A" setting using the same configuration, described in Section IV-A.

Taking Fig. 10 as an example, it visualizes the model's performance using a scatter plot, where each point corresponds to a test utterance. The x-axis denotes the input SISNR before enhancement, while the y-axis indicates the SISNR after applying LargeCall. The diagonal black line represents the identity line, i.e., no change in SISNR. The red line shows the average SISNR across input bins, serving as a visual guide for performance trends. The distance between the red and black lines reflects the SISNR improvement. As shown in Fig. 10, when the average input SISNR is 3.16 dB, the corresponding output improves to 11.17 dB, yielding an enhancement of 8.01 dB ΔSISNR. This demonstrates LargeCall's effectiveness in low-SNR scenarios, where clean speech is difficult to recover using audio-only models. Similar trends are observed in subjective speech quality. Fig. 11 shows that the average MOS of the noisy input is 1.53 , which increases to 1.96 after applying LargeCall. These consistent gains across both objective and subjective metrics validate the effectiveness of LargeCall. Its strong performance stems from the integration of three key modules: a pretrained speech encoder, a multi-dilated accelerometer encoder, and a cross-modal mask fusion module.

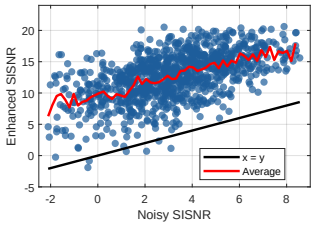

Fig. 10: SISNR enhancement.

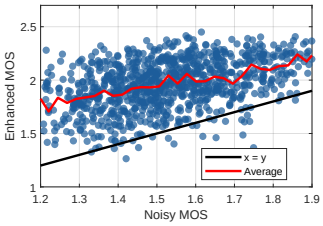

Fig. 11: MOS enhancement.

D. Ablation Study

We perform ablation studies to validate the effectiveness of the proposed components in LargeCall under "2S+A" setting. The corresponding results are shown in Tab. III.

"No Pretrained Encoder Weights" disables weight initialization from MossFormerGAN, training the speech encoder from scratch with random initialization. This leads to notable performance degradation across all metrics: a drop of 3.91 dB in SISNR, 0.07 in STOI, 0.64 in PESQ, and 0.33 in MOS. These results underscore the importance of pretrained speech encoders in capturing robust and generalizable representations, particularly under noisy conditions. This also demonstrates the impracticality of training dual-modal models from scratch.

"No Acc Fusion Module" replaces the Multi-Dilated Articulatory Fusion module in the accelerometer stream with a simple convolutional block. This module is designed to capture multi-scale vibration patterns and integrate signals across the three accelerometer axes. Its removal results in consistent performance degradation, including a 3.46 dB drop in SISNR, 0.03 in STOI, 0.10 in PESQ, and 0.12 in MOS. These findings underscore the importance of the proposed module in extracting stable, speaker-relevant inertial features across diverse phone orientations.

"No Mask Fusion" replaces the proposed weighted mask fusion module with a simple convolutional layer, removing the dynamic weighting mechanism between the audio and accelerometer masks. This leads to performance degradation,

	1S+A				2S+A				3S+A				4S+A				Average			
Method	SISNR	STOI	PESQ	MOS	SISNR	STOI	PESQ	MOS	SISNR	STOI	PESQ	MOS	SISNR	STOI	PESQ	MOS	SISNR	STOI	PESQ	MOS
VibVoice [33]	9.60	0.76	2.22	1.83	8.21	0.74	2.10	1.77	6.69	0.73	1.98	1.75	5.21	0.66	1.86	1.72	7.43	0.72	2.04	1.77
SEANet [24]	10.34	0.78	2.37	1.81	7.21	0.73	2.15	1.76	6.88	0.70	2.01	1.72	3.44	0.69	1.93	1.70	6.97	0.73	2.12	1.75
SepFormer [9]	9.46	0.76	2.22	1.83	7.83	0.75	2.08	1.80	6.62	0.69	1.93	1.78	5.45	0.66	1.85	1.75	7.34	0.72	2.02	1.79
PHASEN [8]	8.45	0.76	2.14	1.70	6.63	0.73	2.03	1.65	4.34	0.69	1.88	1.62	1.82	0.64	1.69	1.60	5.31	0.71	1.94	1.64
MossformerGAN	9.13	0.86	2.72	2.18	6.70	0.81	2.42	2.18	3.92	0.75	2.20	2.18	2.82	0.70	2.08	2.14	5.64	0.78	2.35	2.17
LargeCall	**13.61**	**0.87**	**3.11**	**2.00**	**11.17**	**0.83**	**2.78**	**1.96**	**10.21**	**0.82**	**2.70**	**1.89**	**8.87**	**0.80**	**2.46**	**1.85**	**10.96**	**0.83**	**2.76**	**1.93**

TABLE II: Model comparison.

	SISNR	STOI	PESQ	MOS
LargeCall	**11.17**	**0.83**	**2.78**	**1.96**
No Pretrained Weights	7.26	0.76	2.14	1.63
No Acc Fusion Module	7.71	0.80	2.68	1.84
No Mask Fusion	8.54	0.79	2.47	1.74

TABLE III: Ablation study.

including a 2.33 dB drop in SISNR, and reductions of 0.06 in STOI, 0.31 in PESQ, and 0.18 in MOS. These results confirm that our mask-level fusion design is critical for maintaining modality complementarity and enhancing speech quality.

The ablation results clearly demonstrate that each component of our proposed framework plays a critical role in achieving robust and intelligible speech enhancement.

E. Evaluation on Impact Factors

In this section, we investigate how various factors influence the performance of LargeCall. Unless otherwise stated, the model is trained following the default setup described in Section IV-A. To assess generalization across users and conditions, we additionally recruit 10 volunteers who each participate in all experiments with the "2S1A" noise setting, Each participant speaks 100 sentences selected from the TIMIT corpus, with each utterance lasting approximately 5 seconds and recorded under diverse usage scenarios.

1) Phone Type: To evaluate the cross-device generalizability of LargeCall, we deploy the system on three different smartphones: Samsung Galaxy S10, Xiaomi 13, and OnePlus Ace3 Pro. Note that the model is trained exclusively on data collected from the Galaxy S10. For this experiment, new testing data are collected on all three devices for comparative evaluation. We configure the accelerometer sampling rate to 400 Hz across all devices to ensure consistency. As shown in Fig. 12 and 16, LargeCall achieves average ΔSISNR gains of 8.31 dB, 8.13 dB, and 8.26 dB, and ΔMOS improvements of 0.420, 0.401, and 0.362 on the S10, Xiaomi13, and OnePlus Ace3 Pro, respectively. These results demonstrate that speech enhancement performance remains comparable across all devices, despite training on a single phone type. The findings validate the feasibility of deploying LargeCall on a wide range of consumer devices.

2) Holding Angle: To evaluate the robustness of LargeCall under different phone orientations, we instruct volunteers to hold the phone at various angles during speech. The reference angle of 0° corresponds to the posture in which the phone's Y-axis (in Fig. 9) is aligned parallel to the user's face. Clockwise and counterclockwise rotations are defined as positive and negative angles, respectively. As illustrated in Fig. 13 and Fig. 17 , the enhancement performance remains stable across a typical range of holding angles from -30° to 30°. Specifically, LargeCall achieves ΔSISDR gains of 7.74 dB, 8.21 dB, and 8.06 dB at -30°, 0°, and 30°, respectively. The corresponding improvements in ΔMOS are 0.42, 0.38, and 0.36. These results demonstrate that moderate angular deviations have minimal impact on system performance. This robustness can be attributed to the effectiveness of the proposed Multi-Dilated Articulatory Fusion Module, which enables consistent extraction of speaker-relevant inertial features regardless of posture variation during phone call.

3) Motion Type: To assess the robustness of LargeCall under everyday user activity, we evaluate its performance across four representative motion types: remaining still, nodding, shaking the head, and walking. These scenarios capture a range of motion intensities commonly encountered during phone calls in real-world settings. As illustrated in Fig. 14 and Fig. 18 , LargeCall achieves the highest enhancement performance under the static condition. Despite increased movement, the system maintains stable improvements. Notably, even in the most challenging case of walking, the model achieves an average SISNR gain of 7.76 dB, which is only 0.60 dB lower than the still condition. Likewise, subjective MOS remains consistent. These results highlight the system's ability to suppress motion-induced artifacts while preserving speech-relevant features. This robustness is primarily attributed to the integration of a lightweight high-pass Biquad filter, which effectively attenuates low-frequency body motion noise.

4) Contact Tightness: We evaluate the robustness of LargeCall under different levels of contact tightness between the smartphone and the user's cheek. Volunteers are instructed to hold the phone in three distinct conditions: "Loose", "Normal", and "Tight". In the "Normal" condition, the contact pressure reflects typical phone call behavior. The "Tight" condition involves slightly increased pressure, while the "Loose" condition refers to a light touch between the phone and the cheek. All three conditions maintain physical contact between the phone and the cheek. The "Loose" setting corresponds to minimal skin contact, with increasing pressure applied to reach the "Normal" and "Tight" levels. As shown in Fig. 15 and Fig. 19, the results reveal a modest performance increase as contact tightness improves. For example, the system achieves an average SISNR gain of 7.85 dB in the "Loose" condition, 8.03 dB in "Normal", and 8.10 dB in "Tight". While the

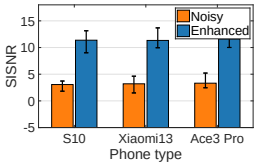

Fig. 12: SISNR vs. phone types.

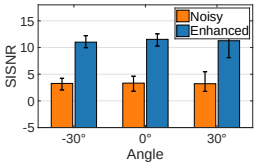

Fig. 13: SISNR vs. angles.

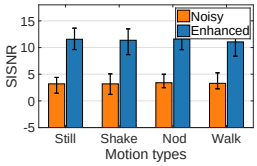

Fig. 14: SISNR vs. motions.

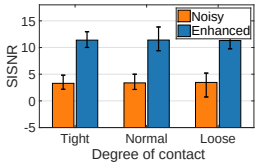

Fig. 15: SISNR vs. contact tightness.

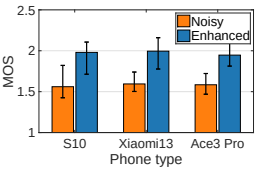

Fig. 16: MOS vs. phone types.

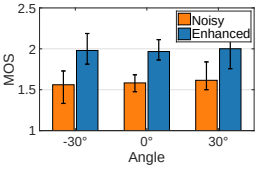

Fig. 17: MOS vs. angles.

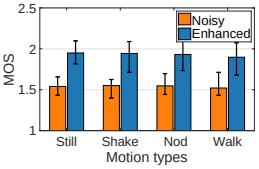

Fig. 18: MOS vs. motions.

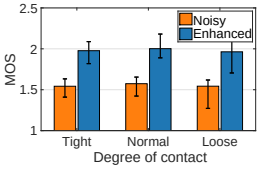

Fig. 19: MOS vs. contact tightness.

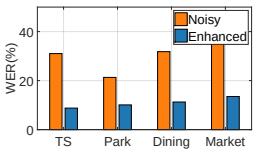

Fig. 20: WER vs. different environments.

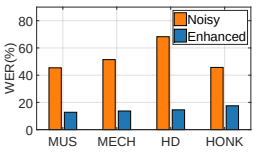

Fig. 21: WER vs. different noise types.

differences are relatively small, they indicate that stronger contact slightly improves vibration sensing by increasing the area of contact and the transmission of articulatory motion.

F. Evaluation in Real-World Scenarios

In this section, we assess the performance of LargeCall in real-world noisy environments where clean reference signals are unavailable. For all experiments, we employ the same pretrained LargeCall model described in Section IV-A. An additional group of 10 volunteers is recruited, each tasked with reading 100 previously unseen sentences, with each utterance lasting 5 seconds. Recordings are conducted across diverse real-world acoustic settings.

1) Different Environments: To evaluate the practical effectiveness of LargeCall, we measure WER performance across four representative noisy environments: a train station(TS), a park, a dining hall(Dining), and a supermarket(Market). As illustrated in Fig. 20, LargeCall consistently reduces WER in all scenarios, achieving an average relative improvement of 18.92%. The most notable gain is observed in the train station environment, which exhibits the highest noise level, where WER drops from 31.03% to 8.79%, corresponding to a 22.24% reduction. In quieter settings such as the park, WER is reduced from 21.31% to 10.09%. These results confirm the robustness of LargeCall in enhancing speech intelligibility under diverse real-world noise conditions.

2) Different Noise Sources: To evaluate the robustness of LargeCall under various acoustic interference types, we test its performance in four representative noise conditions: music(MUS), mechanical noise(MECH), hair dryer noise(HD), and car honking(HONK). As shown in Fig. 21, LargeCall consistently enhances speech intelligibility across all scenarios, achieving an average WER reduction of 38.07%. The largest improvement is observed under the hair dryer condition, where the WER decreases from 68.25% to 14.53%, representing a relative reduction of 53.72%. In other high-noise settings such as mechanical noise and music, WER is reduced by 37.73% and 32.65%, respectively. Even in the presence of transient and non-stationary noise such as car honking, LargeCall still achieves a 28.18% relative improvement. These findings demonstrate the system's ability to generalize across a wide range of noise types.

V. Related Work

Audio-Only Speech Enhancement: Traditional single-modality speech enhancement methods, such as Kalman [34], [35], Wiener [36], [37], and adaptive filtering [38], [39], rely on prior knowledge of clean speech or noise characteristics, limiting their use in real-world scenarios. Recent deep learning models improve performance by mapping noisy to clean speech [40], [41], but remain limited by their reliance on acoustic input, difficulty in separating target speakers, and degraded performance under low-SNR or multi-speaker conditions [11], [25]. Some even require the number of speakers to be known in advance [42], which is impractical for phone calls.

Multi-Modality Speech Enhancement: To improve robustness, multi-modality speech enhancement methods incorporate visual [43], ultrasonic [11], [25], or radio-frequency modalities [44]. Existing multi-modality approaches using video, ultrasound, or RF suffer from practical limitations such as privacy concerns, environmental sensitivity, high energy consumption, and lack of portability, making them unsuitable for real-world phone call scenarios.

VI. Conclusion

In this paper, we present LargeCall, a dual-modality speech enhancement framework that integrates microphone audio with inertial signals from a smartphone's built-in accelerometer. To address key challenges in cross-modal modeling, including limited paired training data, variability in phone placement and user motion, and modality mismatch, we introduce several key components: a bootstrapped pretrained speech encoder, a posture-invariant accelerometer encoder, and a cross-modal mask fusion module. Our results highlight the effectiveness of combining audio and accelerometer signals for real-world speech enhancement, paving the way for more resilient, on-device communication systems in challenging acoustic settings.

REFERENCES

[1] Statista. https://www.deadzones.com/2011/05/how-many-cell-phone-calls-are-made-day.html, 2021.

[2] Matthew B Hoy. Alexa, siri, cortana, and more: an introduction to voice assistants. *Medical reference services quarterly*, 37(1):81–88, 2018.

[3] Ronald E Rice and Douglas E Shook. Voice messaging, coordination, and communication. In *Intellectual teamwork*, pages 341–364. Psychology Press, 2014.

[4] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. Lip-interact: Improving mobile device interaction with silent speech commands. In *Proceedings of ACM UIST*, pages 581–593, 2018.

[5] Shang Zeng, Haoran Wan, Shuyu Shi, and Wei Wang. msilent: Towards general corpus silent speech recognition using cots mmwave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 7(1):1–28, 2023.

[6] Santiago Pascual, Antonio Bonafonte, and Joan Serra. Segan: Speech enhancement generative adversarial network. *arXiv preprint arXiv:1703.09452*, 2017.

[7] Yi Luo and Nima Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *Proceedings of IEEE ICASSP*, 2018.

[8] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. Phasen: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of AAAI*, 2020.

[9] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *Proceedings of IEEE ICASSP*, 2021.

[10] Zirun Zhu, Hemin Yang, Min Tang, Ziyi Yang, Sefik Emre Eskimez, and Huaming Wang. Real-time audio-visual end-to-end speech enhancement. In *Proceedings of IEEE ICASSP*, pages 1–5. IEEE, 2023.

[11] Ke Sun and Xinyu Zhang. Ultrase: single-channel speech enhancement using ultrasound. In *Proceedings of ACM MobiCom*, 2021.

[12] Mubina Toa and Akeem Whitehead. Ultrasonic sensing basics. *Dallas: Texas Instruments*, pages 53–75, 2020.

[13] Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *Proceedings of IEEE ICASSP*, 2021.

[14] Shengkui Zhao, Bin Ma, Karn N. Watcharasupat, and Woon-Seng Gan. Frcrn: Boosting feature representation using frequency recurrence for monaural speech enhancement. In *Proceedings of IEEE ICASSP*, pages 9281–9285, 2022.

[15] Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In *Proceedings of IEEE ICASSP*, pages 10356–10360. IEEE, 2024.

[16] Alibaba. ClearerVoice-Studio: Real-Time Multilingual Speech Enhancement and Separation. https://github.com/modelscope/ClearerVoice-Studio, 2024.

[17] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *Proceedings of ICML*, pages 9226–9259. PMLR, 2022.

[18] Sebastian OH Madgwick et al. An efficient orientation filter for inertial and inertial/magnetic sensor arrays. 2010.

[19] Alipah Pawi. *Modelling and extraction of fundamental frequency in speech signals*. PhD thesis, Brunel University School of Engineering and Design PhD Theses, 2014.

[20] Sangeeta Mandal, Sakti Prasad Ghoshal, Rajib Kar, and Durbadal Mandal. Design of optimal linear phase fir high pass filter using craziness based particle swarm optimization technique. *Journal of King Saud University-Computer and Information Sciences*, 24(1):83–92, 2012.

[21] David Kubanek, Todd Freeborn, Jaroslav Koton, and Norbert Herencsar. Evaluation of (1+ α) fractional-order approximated butterworth high-pass and band-pass filter transfer functions. *Elektronika ir Elektrotechnika*, 24(2):37–41, 2018.

[22] Shapna Rani Sutradhar, Nazmus Sayadat, Ashfiqur Rahman, Sirajum Munira, AKM Fazlul Haque, and Syed Nazmus Sakib. Iir based digital filter design and performance analysis. In *Proceedings of IEEE EL-NET)*, 2017.

[23] Linghan Zhang, Sheng Tan, and Jie Yang. Hearing your voice is not enough: An articulatory gesture based liveness detection for voice authentication. In *Proceedings of ACM CCS*, pages 57–71, 2017.

[24] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. Seanet: A multi-modal speech enhancement network. *arXiv preprint arXiv:2009.02095*, 2020.

[25] Han Ding, Yizhan Wang, Hao Li, Cui Zhao, Ge Wang, Wei Xi, and Jizhong Zhao. Ultraspeech: Speech enhancement by interaction between ultrasound and speech. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 6(3):1–25, 2022.

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

[27] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proceedings of IEEE ICASSP*, pages 776–780. IEEE, 2017.

[28] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proceedings of IEEE ICASSP*, 2010.

[29] Antony W Rix, John G Beerends, Michael P Hollier, and Andries P Hekstra. Perceptual evaluation of speech quality (pesq)-a new method for speech quality assessment of telephone networks and codecs. In *Proceedings of IEEE ICASSP*, 2001.

[30] Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *Proceedings of IEEE ICASSP*, pages 6493–6497. IEEE, 2021.

[31] Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28, 2002.

[32] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, 2003.

[33] Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. Towards bone-conducted vibration speech enhancement on head-mounted wearables. In *Proceedings of ACM MobiSys*, 2023.

[34] K Paliwal and Anjan Basu. A speech enhancement method based on kalman filtering. In *Proceedings of IEEE ICASSP*, volume 12, pages 177–180, 1987.

[35] Stephen So and Kuldip K Paliwal. Modulation-domain kalman filtering for single-channel speech enhancement. *Speech Communication*, 53(6):818–829, 2011.

[36] M Abd El-Fattah, Moawad Ibrahim Dessouky, Salah Diab, and Fathi Abd El-Samie. Speech enhancement using an adaptive wiener filtering approach. *Progress In Electromagnetics Research M*, 4:167–184, 2008.

[37] Navneet Upadhyay and Rahul Kumar Jaiswal. Single channel speech enhancement: using wiener filtering with recursive noise estimation. *Procedia Computer Science*, 84:22–30, 2016.

[38] Pogula Rakesh and T Kishore Kumar. A novel rls based adaptive filtering method for speech enhancement. *World Academy of Science, Engineering and Technology, International Journal of Electrical, Computer, Electronics and Communication Engineering*, 9(2):624–628, 2015.

[39] Lalit P Patil, Amiteshwar Bhalavi, Rupesh Dubey, and Mukesh Patidar. Efficient algorithm for speech enhancement using adaptive filter. *International Journal of Electrical, Electronics and Computer Engineering*, 3(1):98, 2014.

[40] Xugang Lu, Yu Tsao, Shigeki Matsuda, and Chiori Hori. Speech enhancement based on deep denoising autoencoder. In *Proceedings of Interspeech*, volume 2013, pages 436–440, 2013.

[41] Yong Xu, Jun Du, Li-Rong Dai, and Chin-Hui Lee. A regression approach to speech enhancement based on deep neural networks. *IEEE/ACM transactions on audio, speech, and language processing*, 23(1):7–19, 2014.

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

[43] Ariel Ephrat, Inbar Mosseri, Oran Lang, Tali Dekel, Kevin Wilson, Avinatan Hassidim, William T Freeman, and Michael Rubinstein. Looking to listen at the cocktail party: A speaker-independent audio-visual model for speech separation. *arXiv preprint arXiv:1804.03619*, 2018.

[44] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, Min Wu, and KJ Ray Liu. Radio ses: mmwave-based audioradio speech enhancement and separation system. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:1333–1347, 2023.