mmMUSE: An mmWave-based Motion-resilient Universal Speech Enhancement System

LINGYU WANG, University of Science and Technology of China, China
KAI WANG, University of Science and Technology of China, China
DEQUAN WANG, University of Science and Technology of China, China
YOU ZUO, University of Science and Technology of China, China
CHENMING HE, University of Science and Technology of China, China
CHENGZHEN MENG, University of Science and Technology of China, China
XIAORAN FAN, Google, USA
HAOJIE REN, University of Science and Technology of China, China
YANYONG ZHANG*, University of Science and Technology of China, China and Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, China

Speech enhancement improves the user interaction experience in voice-based smart systems. While microphone-based speech perception is limited by airborne noise, mmWave is immune to such interference. However, user-device motion hinders mmWave-based vocal extraction, dispersing vocal signals and introducing distortions. In this paper, we propose *mmMUSE*, an mmWave-based motion-resilient universal speech enhancement system that integrates mmWave and audio. To mitigate motion interference, we propose a two-stage method for robust vocal vibration extraction. Moreover, by proposing the Vocal-Noise-Ratio metric to assess the prominence of the vocal vibration, we enable real-time voice activity detection. We also design a complex-valued network that includes an attention-based fusion network for cross-modal complementing and a time-frequency masking network for correcting amplitude and phase of speech to isolate noises. Using datasets from 46 participants, *mmMUSE* outperforms the state-of-the-art speech enhancement models by 26% in SISDR and 34% in STOI on average. It also achieves SISDR improvements of 16.72 dB, 17.93 dB, 14.93 dB, and 18.95 dB in controlled environments involving intense noise, extensive motion, multiple speakers, and various obstructive materials, respectively. Finally, in real-world scenarios, including running, public spaces, and driving, *mmMUSE* achieves WER below 10%.

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

Additional Key Words and Phrases: Speech Enhancement, Motion-resilient mmWave Sensing, Multi-modality Fusion

*Corresponding author.

Authors' Contact Information: Lingyu Wang, University of Science and Technology of China, Hefei, Anhui, China, lywang19@mail.ustc.edu.cn; Kai Wang, University of Science and Technology of China, Hefei, Anhui, China, wkzcml@mail.ustc.edu.cn; Dequan Wang, University of Science and Technology of China, Hefei, Anhui, China, wdq15588@mail.ustc.edu.cn; You Zuo, University of Science and Technology of China, Hefei, Anhui, China, leftright@mail.ustc.edu.cn; Chenming He, University of Science and Technology of China, Hefei, Anhui, China, hechenming@mail.ustc.edu.cn; Chengzhen Meng, University of Science and Technology of China, Hefei, Anhui, China, czmeng@mail.ustc.edu.cn; Xiaoran Fan, Google, Mountain View, California, USA, gunanjiluzhe@gmail.com; Haojie Ren, University of Science and Technology of China, Hefei, Anhui, China, rhj@mail.ustc.edu.cn; Yanyong Zhang, University of Science and Technology of China, Hefei, Anhui, China and Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, Anhui, China, yanyongz@ustc.edu.cn.

ACM 2474-9567/2025/12-ART218
https://doi.org/10.1145/3770672

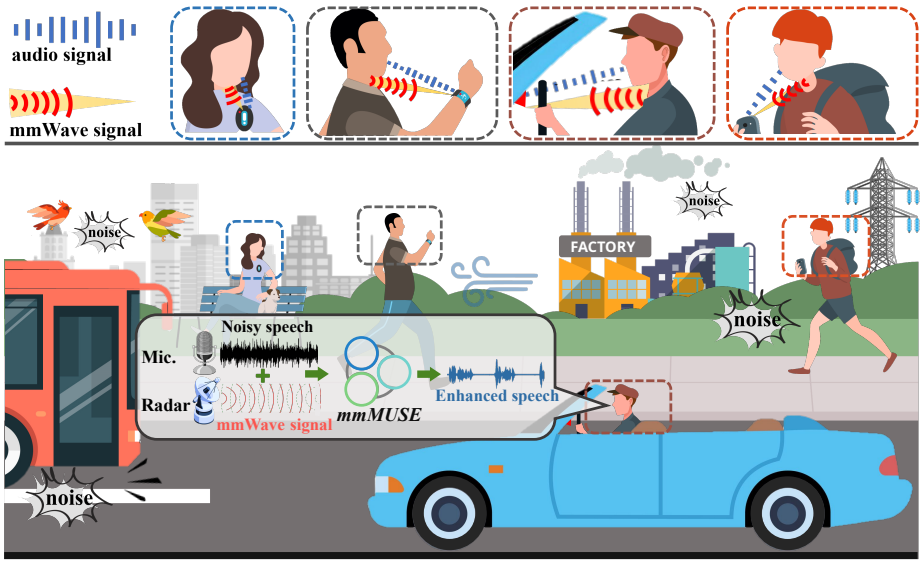

Fig. 1. Applications of *mmMUSE* in daily life. Users can interact with smart devices equipped with *mmMUSE* (e.g., smartphones, watches, and cockpits) for noise-resistant and motion-resilient voice interaction.

ACM Reference Format:
Lingyu Wang, Kai Wang, Dequan Wang, You Zuo, Chenming He, Chengzhen Meng, Xiaoran Fan, Haojie Ren, and Yanyong Zhang. 2025. *mmMUSE*: An mmWave-based Motion-resilient Universal Speech Enhancement System . *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 9, 4, Article 218 (December 2025), 33 pages. https://doi.org/10.1145/3770672

1 INTRODUCTION

Smart systems that leverage advanced voice perception have been increasingly adopted in Internet of Things (IoT) applications, from enhancing the voice quality of smartphone conversations to improving the driver's voice interaction with cars. Such voice-based smart systems can greatly enhance user experiences with their social and physical environments by allowing higher-quality interactions through better voice perception. For example, fixed devices such as in-car assistants (e.g., Li Auto's MindGPT [5] and NIO's NOMI [60]) enhance the driver experience by enabling voice-activated tasks like path selection and cabin temperature adjustments, allowing drivers to focus on driving. Similarly, wearable mobile devices (e.g., iPhone's Siri [3] and Humane's AI PIN [39]) support voice interactions for tasks such as making calls, sending texts, and managing smart home devices, enabling hands-free operation. Given the growing ubiquity of smart devices [9, 13, 44], advancing speech enhancement technology for practical applications is a critical area of research.

Imagine users carrying smart devices strolling through crowded streets or driving on uneven roads. The omnipresent noise can hinder these devices from accurately processing voice commands, leading to unsatisfactory user experiences. Therefore, it is imperative that smart devices incorporate advanced voice enhancement technologies to mitigate these disturbances effectively. However, the voice interaction systems embedded in current smart devices, which predominantly utilize cost-effective microphones or arrays, are inadequate at attenuating noise in environments with low signal-to-noise ratios (SNRs) [45, 67, 92, 97]. In contrast, millimeter-wave (mmWave) radar, operating as active sensors, can precisely detect vocal cord vibrations [93], thereby capturing the clean fundamental frequency components of voice signals. As such, a few recent studies [52, 102] have integrated microphones with mmWave radars to deal with ambient noises to enhance speech recognition.

However, these studies have not fully addressed the common issue of *motion interference* everywhere in everyday life. Specifically, they extract vocal cord vibration signals (vocal signals) based on the differences in mmWave phase signals. User-device motions, however, can shift vocal signals distributions across range bins. Additionally, phase signals within the same range bin can also suffer from motion-induced distortions, as shown in Fig. 4. Such distortion causes overlapping or loss of vocal frequency components (i.e., the fundamental frequency and harmonics) and spectral leakage, hindering voice activity detection [52, 64] and voice authentication [49].

To mitigate the impact of motion interference, we devise a two-stage method for robust vocal vibration extraction, first tracking the signals containing vocal information and then compensating for motion-induced distortions. *Our objective is to develop a system that enables users to engage in voice interaction naturally with wearable or fixed smart devices in open environments, without auxiliary accessories such as earphones.* Specifically, we propose *mmMUSE*, an **mm**Wave-based **m**otion-resilient **u**niversal **s**peech **e**nhancement system with universality across real-world conditions, such as intense noise, drastic user-device movement, clothing occlusion, and multiple speakers. As shown in Fig. 1, *mmMUSE* enhances interaction in noisy open-field environments when integrated into mobile devices (e.g., smartphones and smartwatches) or fixed platforms such as vehicle cabins. Leveraging the ranging capability of mmWave radar, *mmMUSE* tracks the user's reflected vocal signals in real-time. Furthermore, it reconstructs the time-frequency (T-F) spectrograms of these signals to isolate the pure voice fundamental frequencies. Simultaneously, a microphone captures comprehensive voice signals. *mmMUSE* then integrates the mmWave and audio signals for voice activity detection (VAD) while optimizing computational resource efficiency. To effectively enhance speech by fusing two modalities while considering both amplitude and phase, we design a complex-valued fusion network based on attention mechanisms, followed by a T-F mask network that estimates amplitude and phase information from the fused features. This multi-modal interaction allows *mmMUSE* to integrate both modalities, producing clear voice signals that significantly improve downstream tasks like speech recognition and translation. Specifically, the challenges we face during the system design process are as follows:

The first challenge is to obtain clean vocal vibration signals under extensive motion interference. The user-device distance varies during motion, shifting vocal signal distributions across range bins. Thus, the system needs to continuously track and determine which range bin contains the most vocal signals. Existing methods [52, 65, 102] for extracting vocal vibration are based on the phase difference of a single range bin. However, during the phase extraction, motion and direct current (DC) components of mmWave radar cause nonlinear modulation of the vibration phases, which can overlap the vocal vibration frequencies. These methods do not consider the movements across range bins, which are very common in real life, as we have pointed out. To resolve the issues, we first propose an intensity-based neighborhood search algorithm to continuously track potential signals containing the user's vocal vibration. Then, we propose the Vocal-Noise-Ratio (VNR) metric to quantify vocal quality and select the portion containing the most vocal signals. Finally, we design a motion compensation algorithm to reconstruct the T-F spectrogram of the vocal signals, effectively neutralizing motion-induced distortions.

The second challenge is to detect the voice activities of the user. In daily life, the user may be in noisy places with other people, resulting in a low SNR. The method relying solely on audio for voice activity detection can be exceedingly challenging [14, 16, 21] with low SNRs. Moreover, voices of others may also activate the device. To this end, we design an mmWave-based VAD algorithm utilizing the distance measurement capability of mmWave radar. Based on the vocal information in mmWave signals, we develop a CFAR-based spectrogram masking algorithm to eliminate radar noise and spectral leakage caused by motion interference. This approach efficiently isolates vocal frequencies, yielding a purified spectrogram of vocal signals. Additionally, by using VNR to assess the prominence of vocal signals, it precisely identifies the timing of the user's speech through the mask, filtering out non-speech activities outside of the speaking intervals to capture the user's vocal signal.

The third challenge is to fuse the modalities of mmWave and audio for speech enhancement. To maximize the utilization of data from both modalities, we employ a complex-valued network that considers both amplitude and phase. mmWave signals contain fundamental frequency information of voice and thus have an intrinsic

relationship with the audio signals. In cases where data from one of the modalities is missing, we use an attention mechanism to interact and fuse global and local features of both modalities. For the fused data, we design a T-F masking network that synthesizes a complex ratio mask (CRM), integrating both temporal and spectral information to simultaneously correct the amplitude and phase of noisy speech.

The contributions of this paper are as follows:

- We propose an mmWave-based motion-resilient speech enhancement system called *mmMUSE*. It fuses mmWave and audio signals to enhance speeches in cases of intense noises, macro motions, multiple speakers, and common clothing obstructions around the throat area under complex conditions.
- We perform an in-depth investigation of how the modulation process of mmWave signal phases is affected by motions, observing that vocal signals extracted from phase differences are distorted due to motion-induced overlapping, which is difficult to separate. Furthermore, we observe that the vocal vibration velocities are superimposed on body torso motion velocities without any overlapping. Based on it, we propose a motion compensation method via Doppler spectrogram to extract vocal signals resilient against motion interference.
- We design a pipeline for mmWave-based speech enhancement. Firstly, we design a CFAR-based spectrogram masking algorithm to extract vocal frequencies from mmWave. By introducing the Vocal-Noise-Ratio metric to assess the prominence of vocal signals, we achieve real-time voice activity detection. Moreover, we design a two-stage complex-valued network that includes an attention-based multi-modal fusion network and a T-F masking network for correcting the amplitude and phase of speech to isolate noises.
- To fully validate the effectiveness of *mmMUSE*, we recorded a paired dataset of mmWave and audio. We invited 46 participants and conducted the recordings (data duration approximately 18 hours) under various challenging conditions, including intense noises (with a minimum SNR of -15 dB), extensive motion, multiple speakers, and different obstructive materials. We validate that the radar effectively enhances speech within an angular range of (-60°, +60°) in azimuth and (-60°, +15°) in elevation relative to the user. It features a field of view (FOV) of 80° horizontally and 40° vertically, making it suitable for daily handheld and fixed device scenarios. Using the datasets, *mmMUSE* outperforms four state-of-the-art speech enhancement models by 26% in SISDR and 34% in STOI on average. Additionally, it achieves SISDR improvements of 16.72 dB, 17.93 dB, 14.93 dB, and 18.95 dB in controlled environments involving noise, motion, multiple speakers, and material obstructions, respectively. Finally, we evaluate *mmMUSE* in real-world scenarios including running, public spaces, and driving, maintaining WER below 10%. The demo video can be found at https://www.youtube.com/watch?v=HKQnhr0uwXI.

2 BACKGROUND

2.1 Principles of mmWave Radar Sensing

mmWave radar comprises transmitting antennas (TX) and receiving antennas (RX). The TX emits periodic FMCW chirps, which are reflected by the subject and captured by the RX. Subsequently, the transmitted and received signals are mixed to generate intermediate frequency (IF) signals:

$$IF(t) = A_t A_r \exp\left[j2\pi\left(\rho t + f_0\right)\tau\right], \tag{1}$$

where ρ represents the chirp rate, f_0 is the start frequency, and τ is the signal delay. Thus, the frequency of the IF signal is given by $f_{IF} = 2\pi\rho\tau$. These IF signals are then processed to estimate the subject's range and velocity. Furthermore, due to the millimeter-scale wavelength, the phase variations enable detection of micro-movements.

Range and Velocity Estimation. Due to the time of flight, a frequency shift occurs between transmitted and received signals. Thus, the frequency of the IF signal (f_{IF}) is proportional to the distance (d) between the radar

and the subject: $d = \frac{f_{IF}c}{2\rho}$, where c is the speed of light. For a moving subject, the phase of the IF signal changes with the distance between the radar and the subject: $\varphi \approx 2\pi f_0 \tau = \frac{4\pi d}{\lambda}$, where λ is the wavelength. Therefore, the phase difference between consecutive chirps, separated by a time interval T, is linearly proportional to the subject's Doppler velocity $v = \frac{\lambda \Delta\varphi}{4\pi T}$.

Vocal Vibration Detection. The vibration-induced tiny displacements ($\delta \ll \lambda$) can be expressed as:

$$\delta = \frac{c}{4\pi f_0} \Delta\varphi. \tag{2}$$

Both phase difference and IQ (In-phase and Quadrature) signals (see Equation 1) contain vibration-induced phase variations. As shown in Fig. 5(b), the upper and lower spectrograms correspond to the phase difference and IQ signals, respectively, both capturing vocal vibrations.

It is worth noting that mmWave sensing of vocal vibrations is limited to voiced segments, such as vowels. To address this limitation, *mmMUSE* adopts an mmWave–audio fusion strategy for full-sequence speech enhancement, motivated by two key observations. First, vowels typically exhibit higher energy and function as acoustic anchors, playing a dominant role in speech intelligibility and perceived clarity [18–20]. Second, speech is inherently sequential, with voiced and unvoiced segments alternating. Robust mmWave cues from voiced regions can thus support the enhancement of neighboring unvoiced segments through temporal modeling [28, 87]. In our speech enhancement network, mmWave cues enhance voiced segments, while unvoiced segments are inferred from audio and learned temporal dependencies.

2.2 Network Preliminary

Attention Mechanism for Fusion. Attention mechanisms have been widely adopted in multi-modal fusion [8, 71, 101], achieving notable success in natural language processing and computer vision. Attention modules such as Squeeze-and-Excitation (SE) [37] and Spatial Attention (SA) [91] enhance multi-modal fusion by capturing both global and local features. Moreover, attention mechanisms can effectively resist the situation where there are losses in either modality [48, 52]. Additionally, residual connections [34] promote feature reuse and improve generalization. In this paper, we design an attention-based fusion network for cross-modal complementing, utilizing SE blocks, SA modules, and residual connections to enable efficient feature extraction.

Complex-valued T-F Masking Network. Most amplitude-only methods estimate the magnitude spectrogram while reusing the noisy phase, which degrades performance, especially at low SNRs [15, 66, 103]. However, phase is essential for natural and intelligible speech reconstruction [66, 103]. Although iterative methods improve phase estimation [29, 59], their slow convergence limits real-time applicability [103]. Recent advances [1, 53, 97] in T-F mask learning improve enhancement by integrating temporal and spectral cues. Compared to time-domain methods [54, 68, 69], T-F domain approaches are more robust to SNR variations and can achieve better performance, even with limited datasets [100]. Therefore, we adopt a complex-valued network to generate T-F masks that jointly model both the amplitude and phase components of speech.

3 SYSTEM DESIGN

We design an mmWave-based motion-resilient universal speech enhancement system named *mmMUSE*, which fuses mmWave and audio signals while considering both amplitude and phase to improve speech quality and intelligibility under intense noise, extensive motion, clothing occlusion, and multiple speakers. As shown in Fig. 2, we propose the following three modules.

Motion-resilient Vocal Vibration Extraction. We use an mmWave radar to transmit chirps and then extract vocal signals from IF signals obtained from reflections. Motion interference leads to vocal signals being distributed across multiple range bins, while the narrow width of each bin (just a few centimeters) still permits Doppler

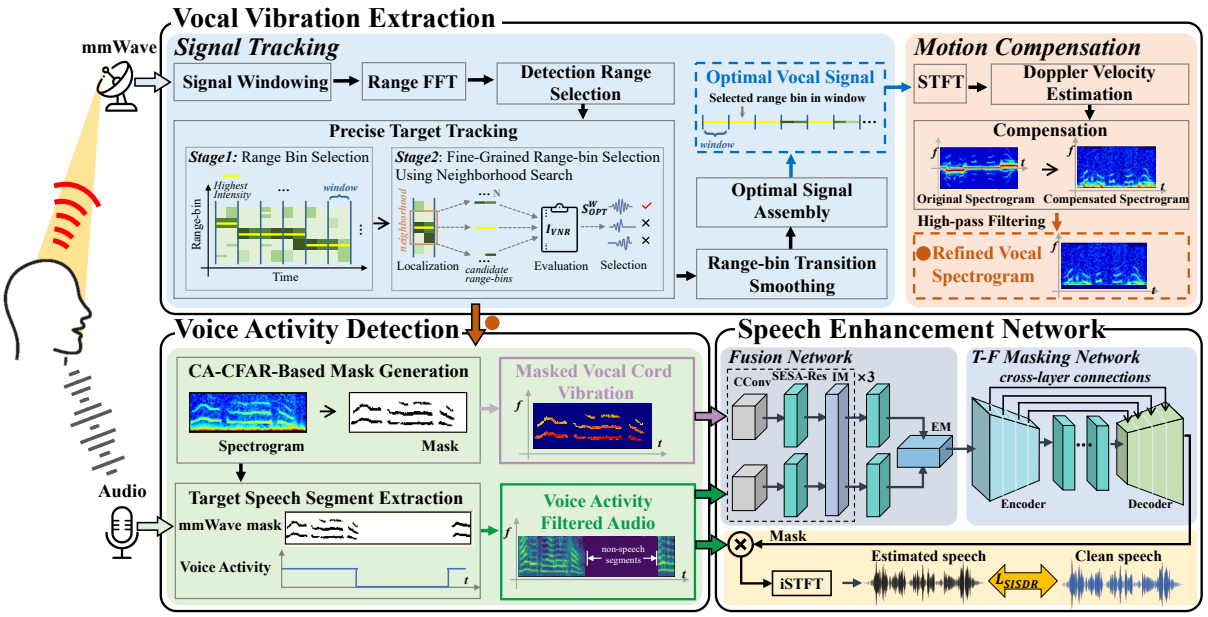

Fig. 2. ***mmMUSE*, an mmWave-based motion-resilient universal speech enhancement system that leverages an mmWave radar and a low-cost microphone to improve the resistance against noise, motion interference and multiple speakers in complex environments.**

shifts within a single bin. To obtain clean vocal signals under extensive motion interference, there are two issues to be addressed. (1) Considering the vocal signal distributions to shift across range bins caused by user-device motion, we propose an intensity-based neighborhood search algorithm to continuously trace signals containing the most vibration information. (2) To eliminate motion-induced distortions in vocal signals within a single range bin, we design a motion compensation algorithm that yields the refined vocal vibration spectrogram.

mmWave-based Voice Activity Detection. We design a CFAR-based spectrogram masking algorithm to extract vocal frequencies from mmWave without acoustic noise. By using the Vocal-Noise-Ratio metric to assess the prominence of vocal signals, we achieve real-time voice activity detection. This enables efficient elimination of non-user voice parts, conserving computational resources.

Two-stage Speech Enhancement Network. To maximize the utilization of data from both modalities, we design a two-stage complex-valued network that considers both amplitude and phase to enhance speech. The network includes an attention-based fusion network for cross-modal complementing and a T-F masking network for correcting amplitude and phase of speech to isolate noises.

3.1 Motion-resilient Vocal Vibration Extraction

In this section, we first provide a theoretical derivation of the interference caused by motion on mmWave phase difference signals and IQ signals, and validate it through simulation experiments. Based on these preliminary studies, we then devise a two-stage method for robust vocal vibration extraction, which first tracks signals containing vocal information and subsequently compensates for motion-induced distortions.

3.1.1 Preliminary Study. As shown in Fig. 3, the received mmWave signals (S) consist of body torso motion (S_b), sub-body motion (S_{sb}, e.g., head movements, facial expressions, and chest fluctuations), vocal vibration (S_v), and

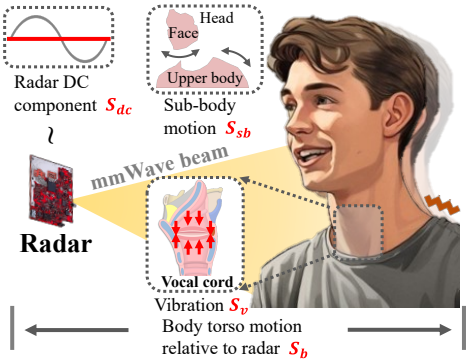

Fig. 3. The components of the radar's IF signal reflected from the user's vocal cord area.

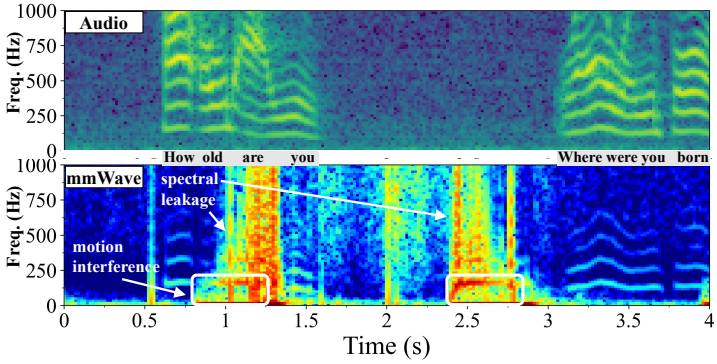

Fig. 4. Real-world motion interference on phase difference signals.

the internal DC component of the radar (S_{dc}), as expressed by:

$$S = \bar{S}_v + \bar{S}_{sb} + S_b + S_{dc}, \tag{3}$$

where both sub-body motion and vocal vibration are modulated by body torso motion:

$$\bar{S}_v = S_v \cdot S_b, \quad \bar{S}_{sb} = S_{sb} \cdot S_b. \tag{4}$$

The phase (Φ) can be calculated through the quadrature(Q) and in-phase(I) components of the signal: $\Phi = \arctan\left(\frac{Q}{I}\right) = \arctan\left(\frac{\bar{Q}_v+\bar{Q}_{sb}+Q_b+Q_{dc}}{\bar{I}_v+\bar{I}_{sb}+I_b+I_{dc}}\right)$. We represent the phases of $\bar{S}_v$, $\bar{S}_{sb}$, S_b, and S_{dc} as $\bar{\phi}_v$, $\bar{\phi}_{sb}$, ϕ_b, and ϕ_{dc}, respectively. Since arctan is nonlinear, Φ does not equal the sum of the aforementioned phases. The phase difference signal ($\Delta\Phi$) is calculated as $\Delta\Phi(n) = \Phi(n) - \Phi(n-1)$. Due to the nonlinear modulation, $\Delta\Phi$ in the frequency domain reveals the distortion of vocal signals caused by motion, as shown in Fig. 4. Motion interference hinders the extraction of vocal signals.

To avoid explicitly calculating Φ, we can use IQ signals (as shown in Equation 1), thereby bypassing nonlinear modulation. However, in the presence of motion, both the vocal vibration velocity and the sub-body motion velocity are affected by the superposition of body torso motion, as described in Equation 4. In this case, the IF signal can be expressed as:

$$IF(t) = A_t A_r \exp\left[j\left(2\pi f_0 \tau + 2\pi\rho\tau t - \pi\rho\tau^2\right)\right]. \tag{5}$$

Substituting $\tau = \frac{2(d+Vt)}{c}$ yields the frequency of IF signals as:

$$f_{IF} \approx \frac{2f_0 V}{c} + \frac{2\rho d}{c}. \tag{6}$$

Here, V represents the velocity at different body positions. As a result, the frequencies of vocal vibration and sub-body motion experience Doppler shifts [7]: $\bar{f} = \frac{2f_0(v+V_b)}{c}$, where V_b is the velocity of body torso motion. Therefore, we can obtain the body torso motion frequency f_b, vocal vibration frequency $\bar{f}_v$, and sub-body motion frequency $\bar{f}_{sb}$ using Doppler fast fourier transform (FFT), where motion causes a frequency shift rather than overlap, as shown in Fig. 6(b).

Simulation Experiment. To explore the interference of motion on the signal, we conducted a simulation experiment. Fig. 5 illustrates the negative impact of motion on the extraction of vocal signals. The phase difference spectrogram and Doppler spectrogram for the stationary vocal signal, shown in Fig. 5(b), serve as a reference, revealing the vocal frequencies without motion interference. The simulations involve two motion signals: $\mathbf{V_1}$ with a constant speed of 0.4 m/s and $\mathbf{V_2}$ with variable velocity up to 1 m/s, as illustrated in Fig. 5(a).

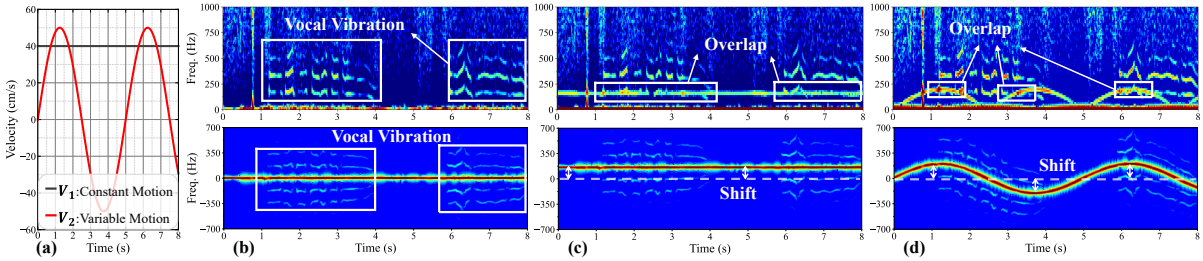

Fig. 5. Simulation of motion interference on phase difference and IQ signals at different velocities. (a) V_1: Constant velocity of 0.4 m/s, V_2: Variable velocity up to 1 m/s; (b) Phase difference spectrogram and Doppler spectrogram (from top to bottom) showing vocal fundamental frequencies under stationary conditions; (c) Motion interference under V_1; (d) Motion interference under V_2.

The phase difference and its spectrogram, along with the Doppler spectrogram derived from the IQ signals, modulated by motion, are shown in Fig. 5(c) and 5(d). The horizontal lines (Fig. 5(c)) and the arc-shaped lines (Fig. 5(d)) in the phase difference spectrogram represent frequency components modulated by motion, leading to the overlap of vocal frequencies. Compared to the phase difference spectrogram, the Doppler spectrograms in Fig. 5(c) and 5(d) show vocal frequencies affected by Doppler shift due to motion, but without overlap.

Therefore, motion interference in Doppler spectrograms can be mitigated by compensating for the frequency shift of the vocal signals. Due to the limitations of phase difference in removing the overlap-induced distortion, we design a motion compensation algorithm that yields a refined vocal vibration Doppler spectrogram.

3.1.2 Vocal Vibration Signal Tracking. Due to user-device motion, the vocal signals are distributed across different range bins, as shown in Fig. 6(a). We need to obtain the complete signals reflected from the vocal cord area to acquire vocal signals. We design an intensity-based neighborhood search algorithm to continuously track signals containing the most vocal information.

Considering that the signal with the highest intensity contains rich information about the user, we employ Range FFT to process the mmWave IF signals within a time window of size W to track it in real-time. To avoid interference from reflected signals outside the range of the user, we choose a detection range. The matrix of the range spectrogram in this range is denoted by R shaped as $M \times W$, where M represents the detection ranges. We can calculate the range bin ($r_{\max}$) that contains the signal with the highest intensity by:

$$r_{\max} = \text{argmax}_r \sum_{t=1}^{W} |R(r,t)|. \tag{7}$$

We consider that the signal at $r_{\max}$ may not contain the richest vocal signals. To capture the signal with the most vocal information, we perform a search within a neighborhood of size $N = 2n + 1$, encompassing $r_{\max}$ and its surroundings:

$$\mathcal{S}_{\text{neigh}} = \{R(r,t) \mid r \in [r_{\max} - n, r_{\max} + n]\}. \tag{8}$$

The $\mathcal{S}_{\text{neigh}}$ covers all potential reflection areas of the user. If $r_{\max}$ changes after the time W, resulting in discontinuities in the signal's phase and subsequent spectral leakage, we mitigate this effect by applying a window function, such as a Hamming window, to smooth the signals in $\mathcal{S}_{\text{neigh}}$ over W.

For the signals in $\mathcal{S}_{\text{neigh}}$, it is crucial to identify which segment within the neighborhood N contains the most vocal signals. We can generate the spectrum of each segment $S_i (i \in N)$ using the FFT and get its amplitude A_i. Subsequently, we employ CA-CFAR [46] to detect vocal frequencies in A_i by analyzing the internal noise levels of the radar, thereby generating a 0-1 mask M_i. In this mask, regions likely containing vocal frequencies are set to 1,

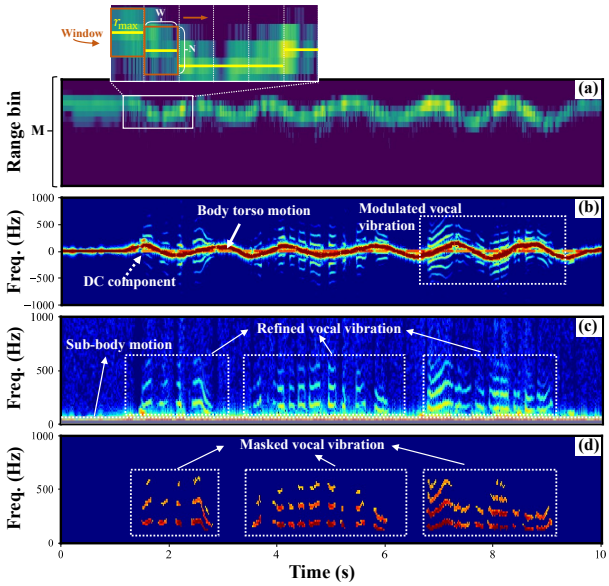

Fig. 6. Vocal vibration spectrogram extraction. (a) User-device motion shifts the range bin with the user's information; (b) Motion-induced modulated vocal vibration; (c) Spectrogram after the motion compensation; (d) Purified spectrogram by masking.

while other noise is set to 0. Similar to the SNR, we introduce the Vocal-Noise-Ratio (VNR) metric to quantify the vocal signals in S_i:

$$I_{\text{VNR}} = \frac{\sum(M_i \cdot A_i)}{\sum(\neg M_i \cdot A_i)}. \tag{9}$$

Specifically, signals within the neighborhood $\mathcal{S}_{\text{neigh}}$ contain nearly identical body reflections (from sub-body and torso regions) and varying vocal content, with body reflections dominating due to their larger reflective surface. Frequencies arising from both components may be marked as 1 in the mask. To mitigate errors in vocal content estimation caused by body motion, we exclude the strongest energy components when computing $M_i \cdot A_i$. As a result, a higher I_{VNR} indicates a greater presence of vocal content. Thus, we select S_{opt}, the segment with the highest I_{VNR}, as the signal containing the most vocal signals for the time W. By reassembling S_{opt} from non-overlap intervals W_i, we obtain the optimal vocal signal $\mathbf{S}_{\text{T}} = \text{Concat}(S_{\text{opt}}^{W_1}, S_{\text{opt}}^{W_2}, \dots)$ containing the user's vocal vibration. For temporal synchronization with the microphone-recorded voice spectrograms, we employ linear interpolation to upsample $\mathbf{S}_{\text{T}}$ to 16 kHz.

3.1.3 Motion Compensation. Since a single range bin covers several centimeters, motion interference still exists within the concatenated signal $\mathbf{S}_{\text{T}}$. As shown in Equation 3, $\mathbf{S}_{\text{T}}$ encompasses essential components, including the vocal vibration signal $\bar{\mathbf{S}}_v$ and the sub-body motion signal $\bar{\mathbf{S}}_{sb}$, both modulated by the user's body torso motion. Additionally, it contains the main Doppler velocity signal $\mathbf{S}_b$ generated by the user-device body torso motion, as well as the radar's internal DC component $\mathbf{S}_{dc}$. The spectrogram of $\mathbf{S}_{\text{T}}$, shown in Fig. 6(b), indicates that the frequency component with the highest intensity corresponds to the user-device body torso motion, which is associated with the largest reflected area. The DC frequency consistently remains near zero, and the vocal frequencies experience Doppler shifts, resulting in distortion.

To achieve the clean vocal frequencies of $\mathbf{S}_v$, it is imperative to eliminate the motion-induced shift and additional interference signals such as sub-body motion and DC components. Since $\mathbf{S}_{\text{T}}$ experiences varying degrees of

frequency shifts due to body torso motion over time, it is not feasible to effectively use filters in the time domain to isolate noises from $\bar{\mathbf{S}}_v$. Therefore, we perform denoising in the T-F domain of $\mathbf{S}_\mathrm{T}$.

As shown in Fig. 6(b), the vocal vibration is modulated by the body torso motion between the user and the device, resulting in frequency shifts. However, due to the oscillatory nature of the vocal vibration, these shifts are symmetric around the body torso motion frequency. Based on this, we propose a motion compensation algorithm to obtain the refined vocal vibration. We conduct short-time fourier transform (STFT) on $\mathbf{S}_\mathrm{T}$, extracting the frequencies at time t, denoted as $\mathbf{F}_\mathrm{S}(t)$. The amplitude of $\mathbf{S}_b$, induced by the body torso motion, is predominant. Hence, the body torso motion frequency f_b is determined using $f_b = \arg\max_f A(f,t)$, where $A(f,t)$ represents the amplitude at frequency f and time t. The shift at time t is compensated by adjusting the frequency components:

$$\widetilde{\mathbf{F}}_\mathrm{S}(f,t) = \mathrm{FreqShift}(\mathbf{F}_\mathrm{S}, f_b), \tag{10}$$

where the operator $\mathrm{FreqShift}(\cdot, f_b)$ shifts $\mathbf{F}_\mathrm{S}$ along the frequency axis by f_b towards zero-frequency, yielding motion-free $\widetilde{\mathbf{F}}_\mathrm{S}$, as shown in Fig. 6(c). Specifically, vacant regions caused by this shift are zero-padded to preserve the original spectrum dimensions.

Since most of the human speech is above 90 Hz [89], and sub-body motions like head movements, facial expressions, and chest fluctuations are generally lower, we establish a cutoff frequency f_cutoff to filter out the sub-body motion signal $\mathbf{S}_{sb}$. Thus, we can obtain the spectrogram of the refined vocal signal $\mathbf{F}_{\mathrm{S}_v}$:

$$\mathbf{F}_{\mathrm{S}_v}(f,t) = \begin{cases} \widetilde{\mathbf{F}}_\mathrm{S}(f,t) & \text{if } f \geq f_\mathrm{cutoff}, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

3.2 mmWave-based Voice Activity Detection

The radar can actively sense the vocal vibration of the user, undisturbed by external sound sources. Utilizing this capability, we extract vocal signals from the mmWave signals to detect the voice activities of the user. This allows for the localization of the speech timings recorded by the microphone, thereby isolating and removing voice components from non-user speakers to obtain truly effective voice signals. Without this detection mechanism, substantial computational resources would otherwise be expended processing irrelevant noise.

However, how can we determine when the signals reflected by the user will exhibit vocal information? To address this, we develop a CFAR-based algorithm that effectively identifies the occurrence times t of vocal vibration in the user and refines the frequencies of the vibration to obtain a clearer spectrogram of vocal vibration $\mathcal{F}_{\mathrm{S}_v}$. Additionally, using the identified vocal activity times t, we perform vocal activity localization on the spectrogram $\mathcal{F}_{\mathrm{S}_a}$ recorded by the microphone.

We apply the CA-CFAR at each moment t within $\mathcal{F}_{\mathrm{S}_v}$ to generate a frequency mask M_t, analyzing for vocal frequencies. For each moment t, we apply the M_t to $\mathcal{F}_{\mathrm{S}_v}$:

$$\widetilde{\mathcal{F}}_{\mathrm{S}_v}(t) = M_t \cdot \mathcal{F}_{\mathrm{S}_v}(t), \tag{12}$$

where $\widetilde{\mathcal{F}}_{\mathrm{S}_v}$ is a refined vocal spectrogram, shown in Fig. 6(d). Based on $\mathcal{F}_{\mathrm{S}_v}(t)$ and M_t, we can calculate I_VNR (see Equation 9) to determine the voice activity at time t:

$$\widetilde{\mathcal{F}}_{\mathrm{S}_a}(t) = \begin{cases} \mathcal{F}_{\mathrm{S}_a}(t) & \text{if } I_{VNR} \geq \alpha, \\ 0 & \text{otherwise,} \end{cases} \tag{13}$$

where $\widetilde{\mathcal{F}}_{\mathrm{S}_a}$ is the audio spectrogram after VAD and α is determined by the statistical values of I_VNR during non-speaking moments, obtained by recording mmWave data in silent conditions. Here, due to motion compensation, the α remains unaffected. As shown in Fig. 7, there is a significant difference in the I_VNR between speech segments and silent segments, thus allowing α to be determined. Specifically, since unvoiced consonants are undetectable

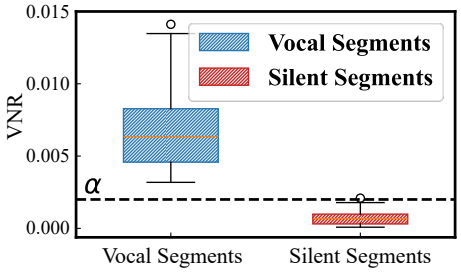

Fig. 7. I_{VNR} under vocal and silent segments.

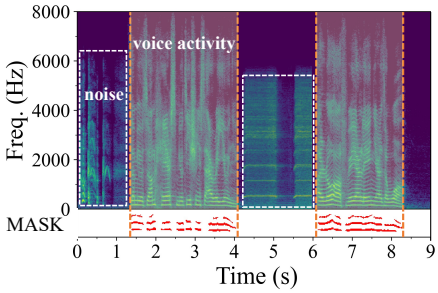

Fig. 8. mmWave-based voice activity detection.

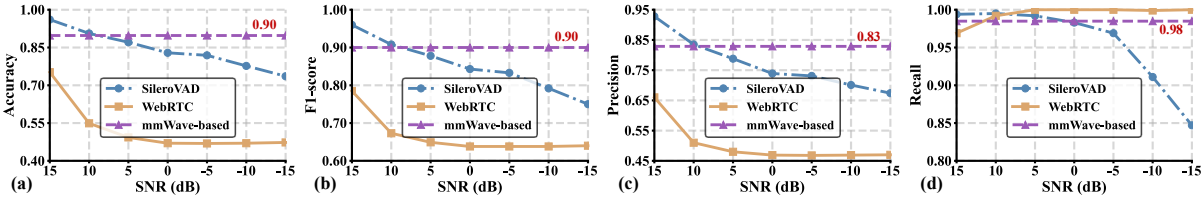

Fig. 9. Performance of VAD across SNRs. (a) Accuracy, (b) F1-score, (c) Precision, (d) Recall.

by mmWave, we use a state machine that merges speech segments separated by brief silence [64, 76]. Fig. 8 shows that the proposed method accurately identifies user presence while efficiently reducing computational overhead.

Performance under Varying SNRs. To evaluate the effectiveness of our mmWave-based VAD, we use paired speech and mmWave recordings, where speakers speak freely with natural body movements. After motion compensation, the mmWave data is processed for VAD analysis. We select two state-of-the-art audio-based VAD algorithms: SileroVAD [76] and WebRTC VAD [27], as baselines. Since precise manual annotation is challenging, we adopt frame-level outputs from SileroVAD under clean speech conditions as pseudo-ground-truth labels. Specifically, all methods use a state machine with a minimum silence duration of 60 ms. We then evaluate the effectiveness of different VAD algorithms under various SNRs, ranging from 15 dB to -15 dB, using accuracy, F1-score, precision, and recall as metrics. While the pseudo-ground-truth labels may contain inaccuracies, our focus is on performance trends across noise conditions rather than absolute metric values.

As shown in Fig. 9, the performance of SileroVAD and WebRTC degrades significantly under low-SNR conditions, with WebRTC generating substantial false positives. In contrast, our mmWave-based VAD remains robust against airborne noise, benefiting from mmWave's direct sensing of vocal vibrations during speech. Moreover, the motion compensation algorithm ensures stability and usability even during user movement.

3.3 Two-stage Speech Enhancement Network

We design a two-stage complex-valued speech enhancement network, as illustrated in Fig. 2. This network comprises an attention-based fusion network and a T-F masking network. The fusion network effectively extracts and integrates features from both mmWave and audio modalities. The T-F masking network effectively processes fused feature maps, generating the complex ratio mask (CRM) [90] while simultaneously estimating amplitude and phase to produce clean audio spectrograms. Finally, clean speech is synthesized using iSTFT. The network is trained using Scale-Invariant Signal-to-Distortion Ratio (SISDR) [47] loss.

3.3.1 Fusion Network. As shown in Fig. 2, the complex-valued fusion network, composed of SESA-based residual blocks (SESA-Res), Interactive Modules (IMs), and Enhancement Modules (EMs), extensively utilizes residual connections [34] for feature reuse. Specifically, the SESA-Res incorporates Squeeze-and-Excitation (SE) [37] and

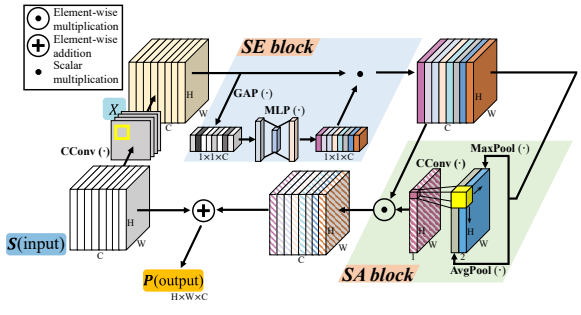

Fig. 10. SESA-based Residual Block.

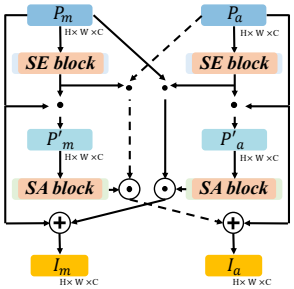

Fig. 11. Interactive Module.

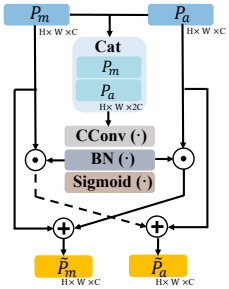

Fig. 12. Enhancement Module.

Spatial Attention (SA) [91] modules to capture both global and local features. This overall architecture enables the model to focus more on the critical aspects of data, thereby enhancing its generalization capabilities on unseen data. mmWave and audio spectrograms each enter two network branches for feature extraction. Each branch consists of a complex convolution (CConv) [86] layer, a batch normalization (BN) layer, and an SESA-Res layer. Subsequently, the outputs of the two branches are linked to a single IM for feature interaction, producing two feature map outputs. We employ three such blocks to increase their channel from 1 to 64. Afterwards, the outputs of both branches are fed into three similar blocks, where EMs replace the previously mentioned IMs, reducing the channel count from 64 back to 1. Eventually, the feature maps of the two modalities are fused together. Next, we detail the three modules: SESA-Res, IM, and EM.

SESA-based Residual Block. As shown in Fig. 10, we develop an SESA-Res to extract features from two modalities, comprising a residual block that incorporates a CConv layer, an SE block, and an SA module for enhanced feature extraction and localization. Input features $S \in \mathbb{R}^{H\times W\times C}$ first pass through a CConv layer, yielding output $X \in \mathbb{R}^{H\times W\times C}$. We apply an SE block [37] to X, performing channel-wise recalibration. The SE block first applies global average pooling to extract channel features, which are then passed through two fully connected layers with a ReLU and sigmoid activation to produce channel-wise weights $\mathbf{W} \in \mathbb{R}^{1\times 1\times C}$. These weights are applied to X via scalar multiplication, resulting in feature maps $X' \in \mathbb{R}^{H\times W\times C}$. The real part X_r and imaginary part X_i are separately input into an SE block to produce X'_r and X'_i. Subsequently, we utilize an SA module [91] to extract spatial features from X'. It uses average and max pooling operations across the channel dimension, concatenates the resulting feature maps, and passes them through a CConv layer followed by a sigmoid activation to generate a spatial attention map $\mathbf{A} \in \mathbb{R}^{H\times W\times 1}$. This attention map is applied to multiply X', highlighting important spatial regions, resulting in feature maps $X'' \in \mathbb{R}^{H\times W\times C}$. Finally, we concatenate a residual block with X'' to construct an SESA-Res as a fundamental module, expressed as:

$$P = \mathrm{SA}(\mathrm{SE}(\mathrm{CConv}(S))) + S, \quad P \in \mathbb{R}^{H\times W\times C}. \tag{14}$$

Interactive Module. We develop an IM for mmWave and audio features to interact by leveraging the intrinsic correlations between modalities. As shown in Fig. 11, the IM inputs the mmWave feature map (P_m) and the audio feature map (P_a) into separate SE blocks to generate weighted feature maps P'_m and P'_a, respectively:

$$P'_m = P_m \cdot \mathrm{SE_m}(P_m), \quad P'_a = P_a \cdot \mathrm{SE_a}(P_a). \tag{15}$$

Subsequently, the SA modules extract local features from P'_m and P'_a, respectively. Finally, P_m is supplemented by the global and local features of P_a, and similarly for P_m, producing final outputs $I_m \in \mathbb{R}^{H\times W\times C}$ and $I_a \in \mathbb{R}^{H\times W\times C}$:

$$\begin{aligned} I_m &= (P_m \cdot \mathrm{SE_a}(P_a)) \odot \mathrm{SA_a}(P'_a) + P_m, \\ I_a &= (P_a \cdot \mathrm{SE_m}(P_m)) \odot \mathrm{SA_m}(P'_m) + P_a, \end{aligned} \tag{16}$$

where $\odot$ denotes element-wise multiplication.

Enhancement Module. We develop an EM to further fuse audio and mmWave feature maps. As shown in Fig. 12, the mmWave feature map (P_m) and the audio feature map (P_a) are concatenated to form $G = [P_m; P_a]$. Subsequently, G passes through a CConv layer to yield the feature map G_{CConv}. A BN layer and a sigmoid function are applied to both the real and imaginary parts, resulting in the feature map G'. Next, G' is multiplied with P_a to obtain the feature E_m for enhancing mmWave features, which is simultaneously used to enhance the audio features. Similarly, E_a is obtained. Finally, the feature maps are updated:

$$\widetilde{P}_m = P_m + E_m, \quad \widetilde{P}_a = P_a + E_a, \quad \widetilde{P}_m, \widetilde{P}_a \in \mathbb{R}^{H \times W \times C}, \tag{17}$$

where $\widetilde{P}_m$ and $\widetilde{P}_a$ are the enhanced mmWave and audio feature maps, respectively. In the final EM of the fusion network, only $\widetilde{P}_a$ is used as the final fused feature, while the other EMs output both branches.

3.3.2 T-F Masking Network. The T-F masking network generates a CRM that adjusts amplitude and phase of noisy speeches, thereby producing a clear speech spectrogram. To enhance *mmMUSE*'s generalization capabilities and reduce the risk of overfitting, the CRM is applied to the original audio spectrogram to preserve the integrity of the data [90]. The network consists of an encoder and a decoder with cross-layer connections, as shown in Fig. 2. The encoder is used to extract features, and the decoder reconstructs the feature map. Between the encoder and the decoder, SESA-Res blocks are utilized to maintain signal integrity and minimize information loss. Additionally, cross-layer connections are employed to preserve details from earlier layers.

Encoder. The encoder has 4 levels of CConv layers, each performing incremental feature extraction from fused maps. Following each CConv layer, a complex BN layer and a Leaky ReLU function [56] are applied. The layers progress from using 16 kernels (3×3) at a stride of 2 for initial downsampling, to subsequent layers that increase the channel count from 32 to 128, producing increasingly refined feature maps m_1, m_2, m_3, and m_4.

Decoder. The decoder module employs 4 symmetrical complex deconvolution layers to reconstruct the spectrogram. 4 deconvolution layers are linked through cross-layer connections to the outputs of the encoder, m_4, m_3, m_2, and m_1, producing a single-channel feature map. Each deconvolution layer is followed by a complex BN layer and a Leaky ReLU function. Ultimately, a CConv layer adjusts the size and details to the final output d_{out}.

CRM Estimation. CRM is used for speech enhancement by isolating speech from noisy environments. Firstly, we obtain the magnitude spectrogram of d_{out}, denoted as $M_{\text{mag}} = |d_{\text{out}}|$. We apply the hyperbolic tangent function to constrain M_{mag} within [0,1), defined as $\hat{M}_{\text{mag}} = \tanh(M_{\text{mag}})$. The phase of d_{out} is derived as $\hat{M}_{\text{phase}} = d_{\text{out}}/M_{\text{mag}}$, obtaining the CRM $\hat{M}_{\text{CRM}} = \hat{M}_{\text{mag}} \cdot \hat{M}_{\text{phase}}$. This CRM filters the noise-mixed spectrogram F_{noise} to extract clean speech through:

$$\hat{F} = \hat{M}_{\text{CRM}} \cdot F_{\text{noise}} = |\hat{M}_{\text{CRM}}| \cdot |F_{\text{noise}}| \cdot e^{i(\theta_{\hat{M}_{\text{CRM}}} + \theta_N)}. \tag{18}$$

The clean signal $\hat{S}$ is recovered by applying iSTFT to $\hat{F}$.

3.3.3 Loss Function. Given the limitations of mean squared error in phase estimation for speech enhancement [15], we adopt SISDR as a time-domain loss function. SISDR evaluates speech quality independently of amplitude variations, focusing on signal fidelity:

$$L_{\text{SISDR}} = 10 \log_{10} \frac{|S_{\text{target}}|^2}{|S_{\text{error}}|^2}. \tag{19}$$

The components are defined by:

$$\begin{cases} S_{\text{error}} = \hat{S} - S_{\text{target}}, \\ S_{\text{target}} = \frac{\hat{S} \cdot S}{|S|^2} S, \end{cases} \tag{20}$$

where $\hat{S}$, S denotes the estimated speech and ground truth, respectively.

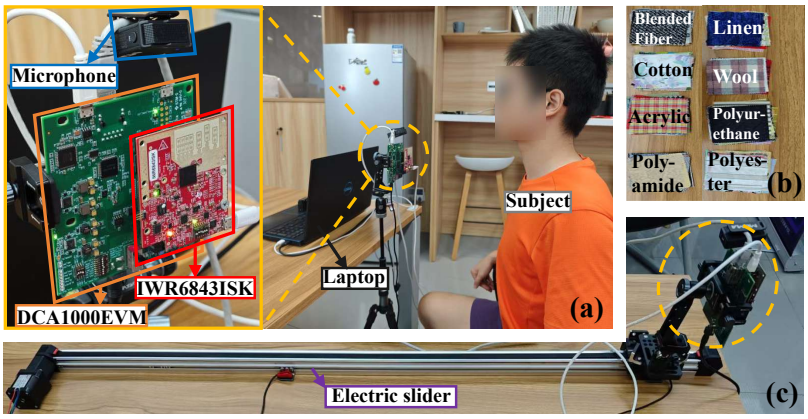

Fig. 13. Experimental setup: (a) An mmWave radar and a microphone receive signals from subjects; (b) An electric slider adjusts user-device motion velocities; (c) Different materials placed over the throat area.

4 EVALUATION THROUGH CONTROLLED EXPERIMENTS

4.1 Experiment Setup

Hardware. The platform is based on an mmWave radar, a laptop, and a single microphone, as shown in Fig. 13(a). We employ TI's IWR6843ISK [83] radar and Infineon's BGT60TR13C [41] radar to capture vocal signals. For TI's IWR6843ISK radar, we use DCA1000EVM [85] to collect raw data. IWR6843ISK and BGT60TR13C both transmit FMCW signals from 60 to 64 GHz. Specifically, *mmMUSE* uses an mmWave radar in a Single Input Single Output (SISO) configuration with one TX and one RX. For portability in outdoor scenarios, we opt for the BGT60TR13C, whereas the IWR6843ISK is utilized in indoor settings. The detailed configurations of radars are shown in Table 1. This configuration enables our radars with a 4.2 cm range resolution. For BGT60TR13C, the vocal vibration sampling rate is 10000 Hz, and for IWR6843ISK, it is 2500 Hz, ensuring the capture of all vocal vibration.

Software. We control the radar using TI's mmWaveStudio GUI [84] or Infineon's Radar Fusion GUI [40] on the laptop, configured with the above-described parameters. We write a Python script to simultaneously control the microphone and GUI, capturing both mmWave and audio signals.

Dataset. We invited 22 male and 24 female volunteers to participate in our experiment. Our studies were approved by the Institutional Review Board (IRB) of our institution. Each volunteer was required to select 15 voice commands from ok-google.io [26] and 13 sentences from the TIMIT Corpus [22], and read each sentence 10 times with each repetition lasting one minute. As shown in Fig. 13(a), participants were asked to maintain

Table 1. Configurations of the mmWave radars.

	IWR6843ISK	BGT60TR13C
		
Frame periodicity	0.4 ms	1 ms
Chirps per frame	1	10
Frequency slope	90.018 MHz/μs	60 MHz/μs
Chirps' idle time	60 μs	40 μs
Chirps' ramp end time	40 μs	60 μs

a distance of approximately 0.3-0.5 m from the microphone and the mmWave radar. We collected a total of 15333 paired data samples, totaling approximately 15 hours of data. Each sample contained both mmWave and audio signals, lasting 3.6 seconds. Specifically, we randomly selected 80% of the samples from both male and female groups for training and allocated the remaining 20% for testing to ensure gender balance. This dataset is referred to as *BasicDataset*. The *BasicDataset* is used exclusively for training and testing in Section 4.2 and does not include data from Section 4.3, which is used solely for testing. Therefore, our model is trained solely in static noise environments using *BasicDataset* and tested under various conditions, including different velocities, distances, angles, multiple speakers, and varying thicknesses of different obstructive materials.

Metrics. We evaluate *mmMUSE* using the following widely adopted metrics. Higher scores indicate better quality.

(1) Scale-Invariant Signal-to-Distortion Ratio (**SISDR**) [47] is a comprehensive assessment of speech clarity and distortion, releasing dependency on amplitude scaling.
(2) Short Time Objective Intelligibility Metric (**STOI**) [80] is designed to quantify intelligibility with scores ranging from 0 to 1, correlating with the word error rate.
(3) Perceptual Evaluation of Speech Quality (**PESQ**) [75] evaluates the quality of speech ranging from -0.5 to 4.5.

We prioritize speech clarity (SISDR) and intelligibility (STOI), which effectively measure improvements in human-machine interaction and call quality. PESQ aims to emulate the human perception of speech quality, focusing on auditory attributes, which have minimal impact on human-machine interaction.

4.2 Overall Performance

We generate the noisy speech by mixing the speech from *BaseDataset* with noises from the TUT dataset [36], which contains 10 noisy scenes (e.g., shopping centers, subways, and pedestrian streets). Each participant's voices are mixed with 30 different noises to generate noisy speeches with SNRs ranging from -15 dB to 15 dB. *The SISDR metrics for noisy speech at different SNRs closely correspond to their respective SNR.*

Baselines. We compare *mmMUSE* with current state-of-the-art speech enhancement networks such as MP-SENet [53], PHASEN [97] and the open-source tool Resemble Enhance [74]. Additionally, we make efforts to reproduce RADIOSES [65], referring to the GitHub repository [4]. MP-SENet and PHASEN are T-F domain methods using dual-stream networks to process amplitude and phase for speech enhancement. Resemble Enhance applies a complex-valued network for noise separation and subsequent speech enhancement. RADIOSES integrates time-domain mmWave and audio signals to enhance speech. Notably, all models serving as baselines are adequately and fully pre-trained on our collected dataset before being validated on the same test set.

Furthermore, we conduct an ablation study to quantify the effectiveness of our system for comparison:

- **Noisy speech after VAD** refers to the **Noisy speech** processed with VAD.
- **Speech-only**, where the network operates without mmWave signals, thus removing the mmWave subnetwork.
- **mmWave-only**, where the network operates without audio signals, thus removing the audio subnetwork.
- **W/O Phase**, where the network takes the amplitude spectrograms of both mmWave and audio as input, and reuses the phase of the noisy audio for speech reconstruction.
- **W/O Attention**, where the network operates without attention mechanisms. We remove the SESA-Res and IM from the fusion network and replace SESA-Res with classical residual blocks [34] in the T-F masking network.
- **W/O Masking**, where the T-F masking network does not generate a complex ratio mask (CRM) but directly estimates the enhanced speech spectrogram.

In this section, we conduct a detailed analysis of *mmMUSE*'s enhancement performance under varying SNRs ranging from -15 dB to 15 dB. Table 2 shows that the average metrics of SISDR, STOI, and PESQ reach 15.79

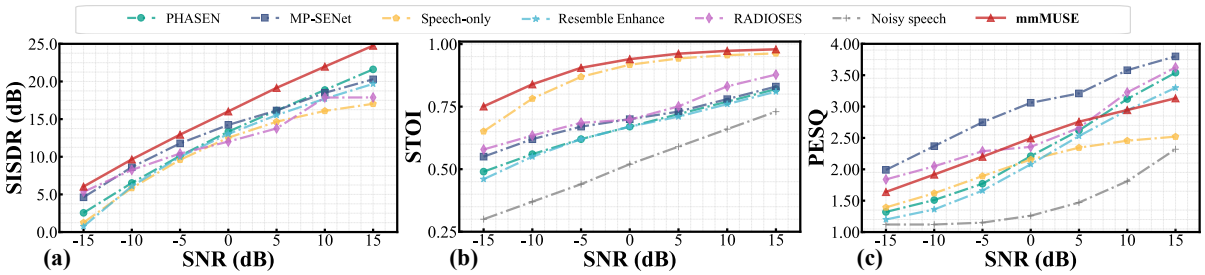

Fig. 14. Performance of *mmMUSE* across SNRs.

dB, 0.91, and 2.44, respectively. Correspondingly, the average improvements in these metrics are 16.72 dB, 0.37, and 0.94, respectively. These results demonstrate *mmMUSE*'s robustness across varying noise levels. *mmMUSE* achieves the best performance in SISDR and STOI, demonstrating its effectiveness in enhancing speech clarity and intelligibility. Due to the time-domain transformation operation of iSTFT, there may be slight changes in timbre, which can lower PESQ scores but do not impact the application of *mmMUSE*. Compared to audio-only methods like MP-SENet, PHASEN, Resemble Enhance, and Speech-only, *mmMUSE* achieves superior performance in SISDR and STOI due to the clean vocal information provided by mmWave. As Fig. 14(a) and 14(b) illustrate, *mmMUSE* surpasses audio-only methods across SNRs, achieving the average SISDR improvements of 3.54 dB and over 0.19 STOI improvement. However, Fig. 14(c) shows *mmMUSE*'s PESQ outperforms PHASEN and Resemble Enhance at lower SNRs but is slightly lower at higher SNRs. The trade-off is attributed to mmWave integration that marginally impacts timbre while maintaining intelligibility. Unlike RADIOSES, which relies only on time-domain features of mmWave and audio while ignoring frequency information, *mmMUSE* exploits both time and frequency domains for superior performance. Additionally, the VAD module effectively detects vocal segments in noisy speeches, enhancing SISDR, STOI, and PESQ by 4.02 dB, 0.30, and 0.15, respectively. Overall, *mmMUSE* surpasses these methods, achieving average improvements of 3.23 dB in SISDR and 0.22 in STOI compared to the baselines, effectively leveraging the intermodal correlations.

Furthermore, Table 2 also presents the results of the ablation study. mmWave-only cannot fully enhance speech, with an average SISDR reduction of 6.18 dB due to incomplete capture of the vocal signal. Removing

Table 2. Performance comparison among speech enhancement methods under different conditions.

	SISDR	STOI	PESQ
Noisy speech	-0.93	0.54	1.50
Noisy speech after VAD	3.09	0.84	1.65
MP-SENet [53]	13.45	0.70	**2.96**
PHASEN [97]	12.74	0.66	2.30
Resemble Enhance [74]	11.83	0.65	2.15
RADIOSES [65]	12.21	0.72	2.57
Speech-only	11.00	0.87	2.05
mmWave-only	9.61	0.84	1.90
W/O Phase	8.65	0.80	1.74
W/O Attention	9.03	0.81	1.82
W/O Masking	9.52	0.82	2.00
mmMUSE	**15.79**	**0.91**	2.44

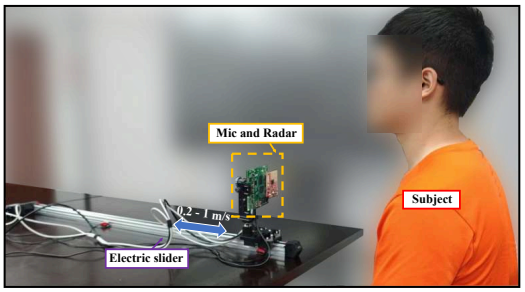

Fig. 15. Experimental setup for user-device relative motion: The participant faces the radar and microphone, with the electric slider moving at velocities from 0.2 m/s to 1 m/s.

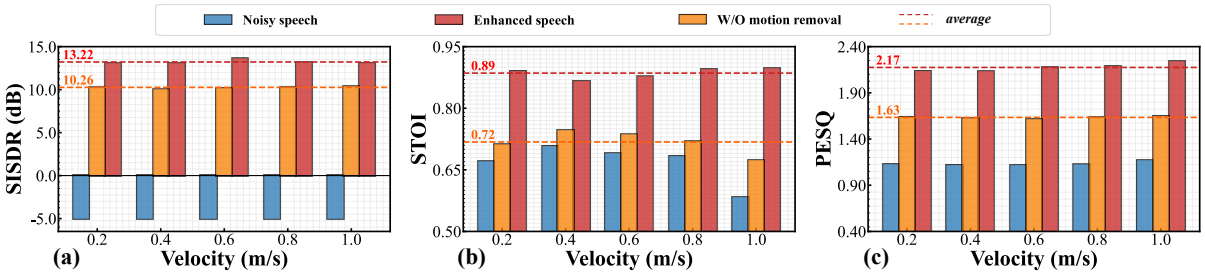

Fig. 16. Ablation study for motion interference removal.

attention mechanisms, the W/O Attention setup fails to utilize both modalities' information, resulting in worse performance than audio-only methods. Moreover, the performance of the W/O Phase drops significantly compared to *mmMUSE*, where SISDR, STOI, and PESQ decrease notably by 7.14 dB, 0.11, and 0.70, respectively. This highlights the importance of phase information for speech enhancement under high noise levels. Removing the learned masking forces the model to directly reconstruct the full spectrogram, which reduces generalization ability and results in an SISDR drop of 6.27 dB. In contrast, T-F masking explicitly modulates the noisy spectrogram, suppressing the distribution of noise and thereby improving enhancement stability. Therefore, the complex-valued design, attention mechanism, and T-F masking in *mmMUSE* 's network architecture enable it to fully exploit the complementary strengths of mmWave and speech modalities. By jointly leveraging amplitude and phase information, the masking operation modulates the noisy spectrogram to achieve optimal performance.

4.2.1 Ablation Study for Motion Interference Removal. We place the device on an electric slider and set the user-device velocity from 0.2 m/s to 1.0 m/s, positioning it directly in front of the participant's head to assess the effectiveness of *mmMUSE*'s motion interference removal at -5 dB SNR. The experimental setup is shown in Fig. 15. For comparison, we remove the motion-resilient vocal vibration extraction method from *mmMUSE* marked as W/O motion removal. The model used for testing is trained solely on static data (*BasicDataset*). As shown in Fig. 16, compared to the complete *mmMUSE*, the average reductions in SISDR, STOI, and PESQ are 2.96 dB, 0.17, and 0.54, respectively, across various velocities. Motion interference significantly reduces the intelligibility and clarity of the enhanced speech. Furthermore, the average SISDR for W/O motion removal is comparable to the audio-only method, but STOI and PESQ are reduced by 0.14 and 0.26, respectively, due to additional noise introduced by motion interference in the mmWave. These results indicate that the motion-resilient vocal vibration extraction method in *mmMUSE* effectively removes motion interference, thereby enhancing performance.

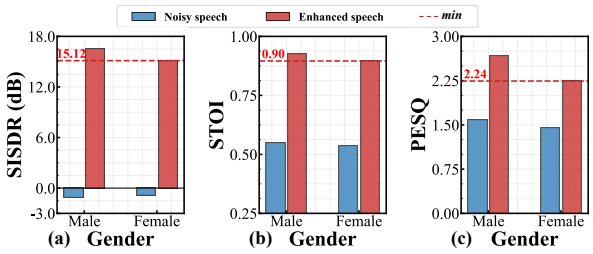

Fig. 17. Impact of gender.

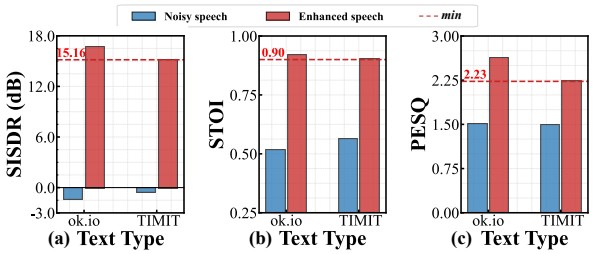

Fig. 18. Impact of text types.

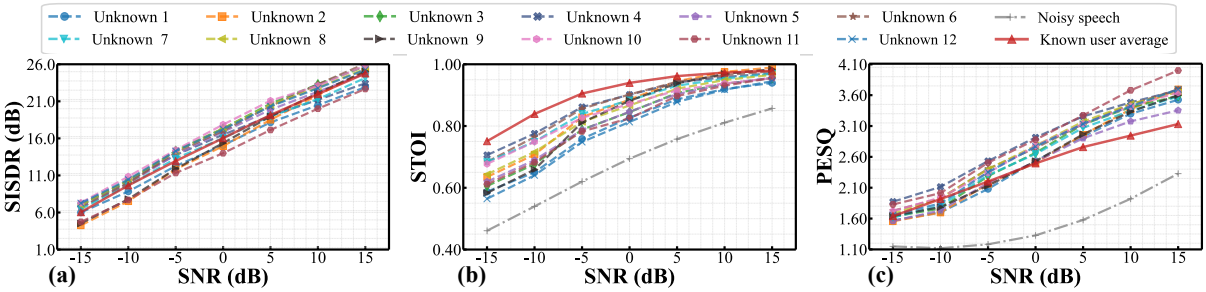

Fig. 19. Generalization for unknown users across SNRs.

4.2.2 Analysis of Data Composition. To analyze the impact of data composition on system performance, we explore the impact of gender and text types on *mmMUSE*. Googleok.io [26] consists of brief commands, while the TIMIT Corpus [22] contains more complex, longer sentences.

Impact of gender. As shown in Fig. 17, both male and female participants achieve SISDR over 15 dB, with STOI and PESQ scores of at least 0.90 and 2.24, respectively, across SNRs from -15 to 15 dB, illustrating *mmMUSE*'s effectiveness for all genders. Males typically have a lower fundamental voice frequency, usually below 155 Hz [89]. This characteristic allows the radar to capture more fundamental frequency information from males than females.

Impact of text types. Fig. 18 shows that *mmMUSE* consistently achieves an SISDR gain of over 15 dB, with STOI and PESQ scores reaching at least 0.90 and 2.23, respectively, across SNRs from -15 to 15 dB for both text types. Performance on the TIMIT Corpus is slightly lower than on Googleok.io, as the TIMIT Corpus contains longer and more complex sentences, leading to unintentional swallowing of voices by participants.

4.3 Robustness Analysis

To validate the performance of *mmMUSE* across various scenarios, we carry out robustness experiments with unknown users, different velocities, distances, angles (azimuths and elevations), multiple speakers, as well as varying thicknesses of different obstructive materials, encompassing 3 hours of paired mmWave and audio data. Notably, the data from the controlled experiments remains unseen to *mmMUSE*, which is trained solely in static noise environments with *BasicDataset*.

4.3.1 Generalization for Unknown Users. We collect mmWave and audio data using unseen texts from 12 new volunteers, both of which are not included in the *BasicDataset*, to evaluate generalization performance. Additionally, we introduce unseen noises from the TUT dataset [36] to generate noisy speeches, with SNRs ranging from -15 dB to 15 dB. As shown in Fig. 19, the Known user average represents the results for users included in the test set of the *BasicDataset*. For unknown users, *mmMUSE* maintains an average SISDR, STOI,

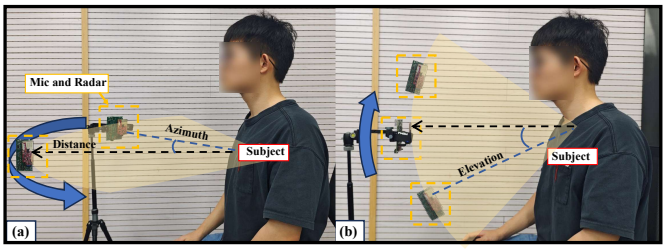

Fig. 20. Experimental setup for distances and angles: (a) The device's azimuth relative to the participant ranges from 0° to 90°, with distances set from 0.2m to 3m; (b) The device's elevation relative to the participant ranges from -90° to 90°, with distances between 0.5m and 1m.

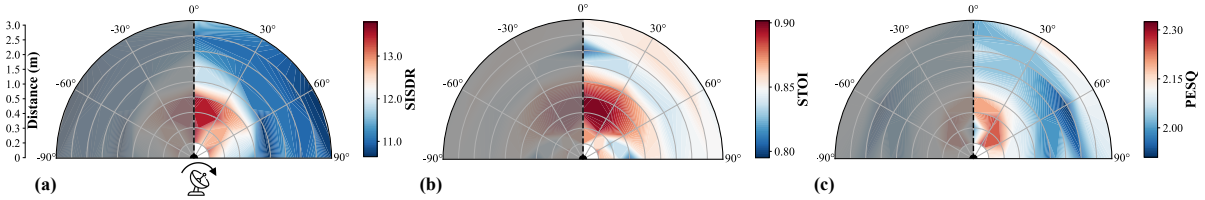

Fig. 21. Impact of distances and azimuths (the shaded area on the left is a duplicate of the data on the right).

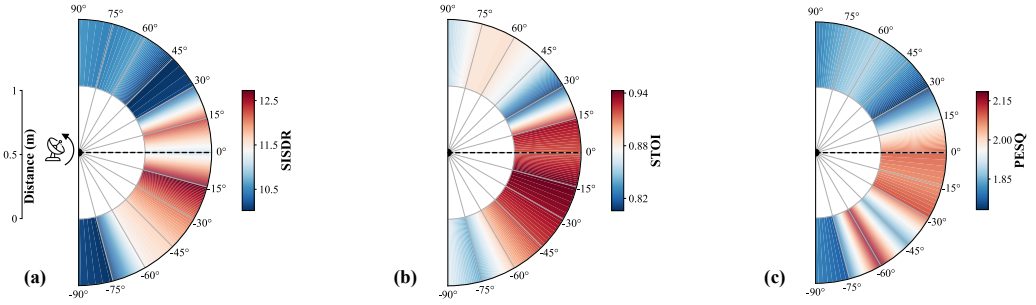

Fig. 22. Impact of elevations.

and PESQ of 15.93 dB, 0.83, and 2.66, respectively, comparable to the Known user average (15.79 dB, 0.91, and 2.44). These results demonstrate that *mmMUSE* can generalize to unknown users. The fusion network within *mmMUSE* utilizes SESA-based residual blocks for feature reuse, and the T-F masking network estimates the CRM that is applied to the original audio spectrogram without compromising its integrity. This endows *mmMUSE* with robust generalization capabilities.

4.3.2 Distance and Angle. We systematically investigate the impact of the user-device distance, as well as the device's azimuth and elevation relative to the participant, on *mmMUSE*. The experimental setup for distance and azimuth is shown in Fig. 20(a), where the azimuth of the microphone and radar relative to the participant varies from 0° to 90°, and the distance ranges from 0.2 m to 3.0 m. Since the vocal vibration is symmetrical, we only test the azimuth range from 0° to 90°. Additionally, the elevation of the microphone and radar relative to the participant varies from -90° to 90°, with distances fixed between 0.5 m and 1.0 m, as depicted in Fig. 20(b).

Impact of distances and azimuths. As shown in Fig. 21, *mmMUSE*'s performance decreases with increasing distances and angles. However, it remains comparable to PHASEN, MP-SENet, and Resemble Enhance. These

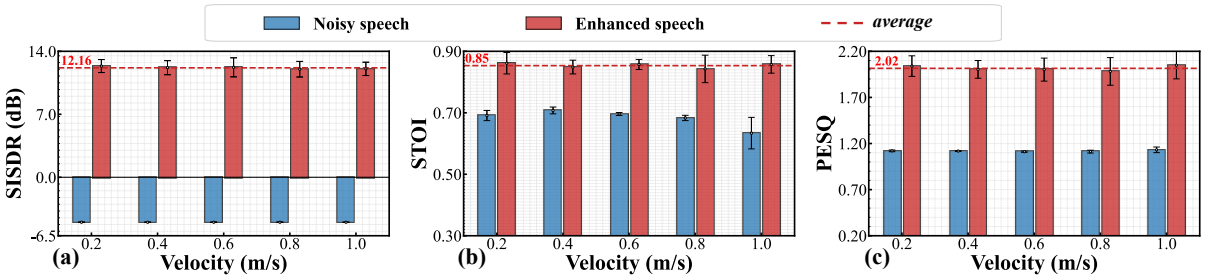

Fig. 23. Impact of velocities: the values represent the average performance of the device relative to the participant, moving at velocities from 0.2 m/s to 1 m/s within an azimuth range of 0° to 90°.

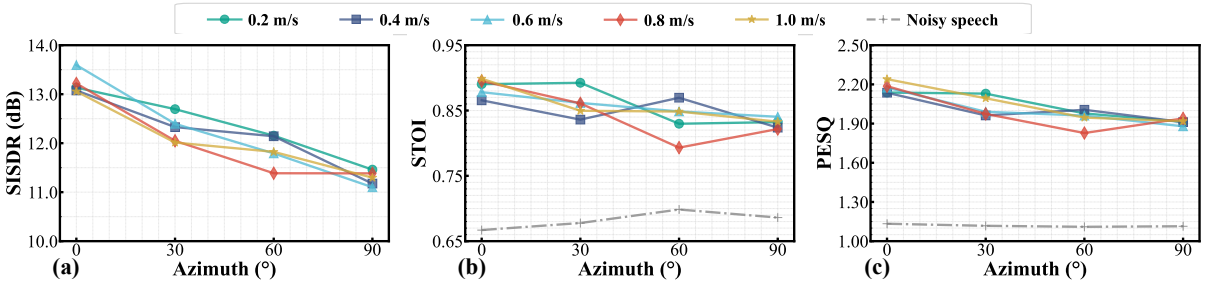

Fig. 24. Impact of velocities and azimuths.

models gain average 10.63 dB of SISDR. Specifically, Fig. 21(a) illustrates that *mmMUSE*'s SISDR gradually declines from 0.2 m to 3.0 m but remains stable above 11 dB, still higher than the aforementioned baselines. Similarly, Fig. 21(b) and 21(c) demonstrate that the STOI and PESQ consistently maintain performance levels of at least 0.80 and 1.91, respectively. Regarding angle, Fig. 21 indicates a gradual decline in performance from 0° to 90°, yet *mmMUSE* still effectively enhances speech. Overall, although mmWave's ability to capture the vocal vibration decreases at longer distances (> 1.5 m) and extreme angles (> 60°), the integration of audio signals in *mmMUSE* effectively counteracts the impacts of distance and angle, thus maintaining considerable performance.

Impact of elevations. As shown in Fig. 22, the primary benefits of mmWave for *mmMUSE* are concentrated within the elevation range of 15° to -60°. This occurs because, at higher elevations relative to the participant, mmWave signals are obstructed by other body parts, preventing the capture of the vocal vibration. However, the elevation range of 15° to -60° is suitable for most practical applications. Furthermore, although *mmMUSE*'s performance decreases at higher elevations (reaching a minimum SISDR of 10.02 dB), it still maintains performance comparable to audio-based speech enhancement models. This demonstrates that our strategy of fusing audio with mmWave enables the two modalities to complement each other, accommodating different spatial positions.

4.3.3 User-device Relative Motion. As shown in Fig. 15, we place the device on an electric slider to explore the impact of user-device velocity (0.2 m/s to 1.0 m/s) and the device's azimuth relative to the participant (0° to 90°) on *mmMUSE* 's performance under -5 dB SNR, with the slider traveling a distance between 0.5 m and 1.0 m.

Impact of velocities. As shown in Fig. 23, compared to the static performance at a distance between 0.5 m and 1.0 m, with an SISDR of 12.76 dB, *mmMUSE* maintains stable average performance across all azimuths, with fluctuations not exceeding 5%. Fig. 23(a) shows that average SISDR across all azimuths exhibits minor fluctuations across velocities, ranging from a low of 12.01 dB to a high of 12.36 dB, with an average of 12.16 dB. This indicates that *mmMUSE* effectively maintains the performance under dynamic conditions, demonstrating

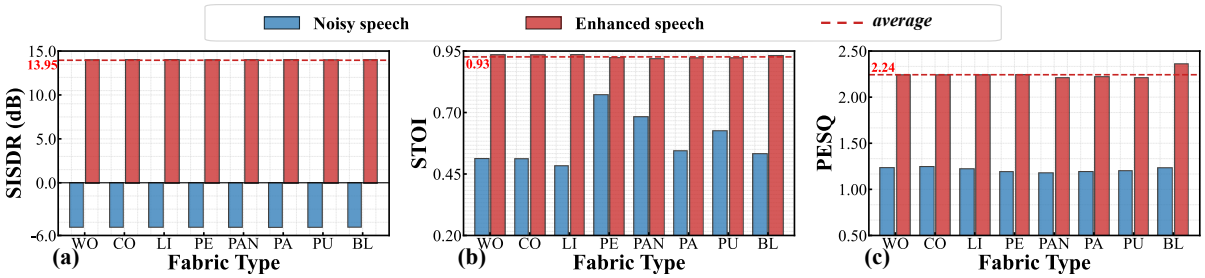

Fig. 25. Impact of obstructive materials.

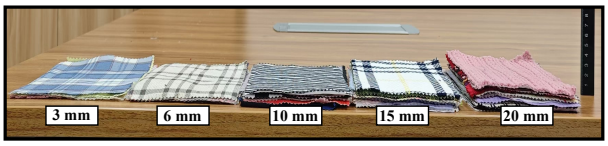

Fig. 26. Experimental setup for obstructive material thicknesses.

excellent robustness. Fig. 23(b) and 23(c) show that fluctuations in STOI and PESQ across all azimuths both do not exceed 2% of the average values, indicating that speech quality and intelligibility are well preserved across various motion velocities.

Impact of azimuths with motion. As shown in Fig. 24, *mmMUSE* is more sensitive to azimuth variations. Fig. 24(a) reveals that SISDR performance typically declines with increasing azimuths across velocities, decreasing from an average of 13.22 dB at 0° to 11.28 dB at 90°, a 15% reduction. Similarly, Fig. 24(b) and 24(c) demonstrate that STOI and PESQ decrease by 6% and 12%, respectively, from 0° to 90°. Overall, *mmMUSE*'s performance is impacted by the azimuth because the structure of the vocal cords inhibits mmWave from capturing comprehensive vocal vibration from the side. Nonetheless, *mmMUSE* still effectively enhances speech by integrating audio signals.

4.3.4 Obstructive Materials. As shown in Fig. 13(c), we simulate daily clothing obstructions by placing eight different materials over the throat area: blended fiber (BL), synthetic fibers (polyester [PE], acrylic [PAN], polyamide [PA], polyurethane [PU]), and natural fibers (wool [WO], cotton [CO], linen [LI]), each with a thickness of 5 mm. Additionally, we explore the impact of varying material thicknesses (3 mm to 20 mm) on *mmMUSE*, as shown in Fig. 26. All experiments are at -5 dB SNR.

Impact of material types. Fig. 25(a) shows that SISDR performance across materials is stable, ranging from 13.90 dB to 13.99 dB with an average of 13.95 dB. Similarly, Fig. 25(b) and 25(c) illustrate that fluctuations in STOI and PESQ among the materials do not exceed 2% and 5%, respectively. The types of obstructive materials used nearly cover all types of clothing materials encountered in daily life. Therefore, these results emphasize *mmMUSE*'s adaptability and stability in the presence of material obstructions, highlighting its robust performance under various physical obstruction conditions.

Impact of material thicknesses. Fig. 27(a) shows that the average SISDR fluctuations for Blended Fiber, Synthetic Fiber, and Natural Fiber do not exceed 2%, 2%, and 3% across various thicknesses, respectively, indicating *mmMUSE*'s stable performance. Fig. 27(b) and 27(c) demonstrate high stability in STOI and PESQ improvements compared to the original noisy speech across various material thicknesses. Specifically, STOI and PESQ metrics maintain comparable performance at both the maximum thickness of 20 mm and the minimum of 3 mm. These results highlight *mmMUSE*'s robustness in facing physical obstructions of various material types and thicknesses, attributable to the penetrating capabilities of mmWave, effectively handling everyday clothing obstructions.

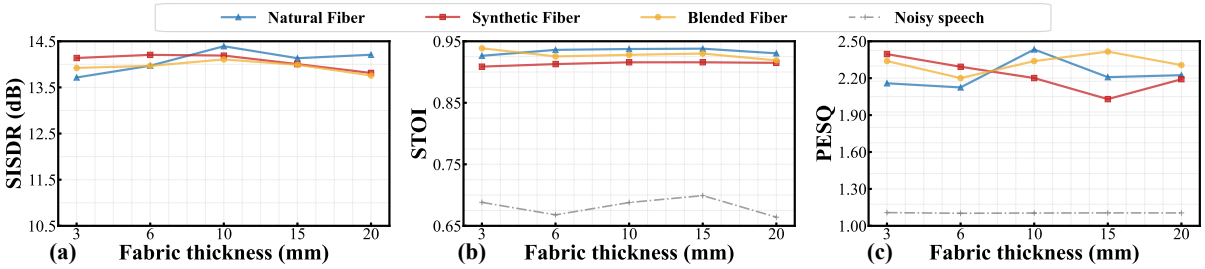

Fig. 27. Impact of obstructive material thicknesses.

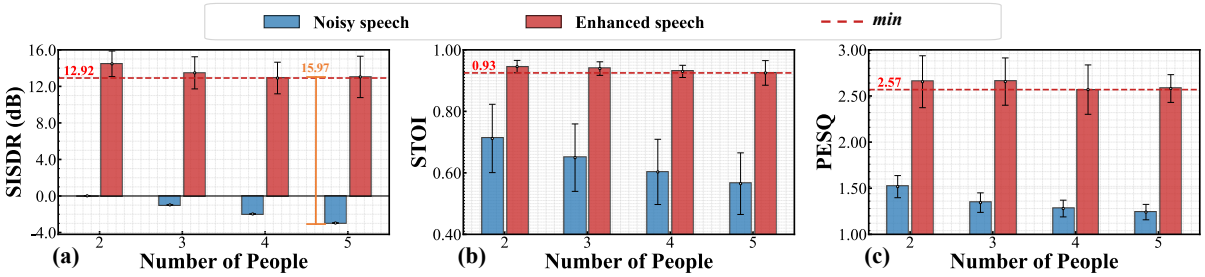

Fig. 28. Impact of multiple speakers.

4.3.5 Multiple Speakers. *mmMUSE* is designed to focus on the target user, while treating other speakers as background noise. Voice interference significantly reduces the intelligibility of the target user speech, thereby hindering human–machine interaction. However, mmWave radar can leverage its range resolution capability to define a spatial detection zone, thereby isolating interference from non-target users. To test the performance of *mmMUSE* in multiple speakers, we randomly select participants with varying numbers of speakers ranging from 2 to 5 to mix voices. In each group of mixed voices, we alternately enhance the speech of all participants, treating each one in turn as the subject. As shown in Fig. 28, SISDR, STOI, and PESQ remain at or above 12.92, 0.93, and 2.57, respectively, maintaining excellent performance. Fig. 28(a) shows that the improvement of SISDR demonstrates *mmMUSE*'s adaptability in multi-person conditions. Although the performance decreases with more participants, the improvement increases from 14.45 dB for two people to 15.97 dB for five people, indicating the system's enhanced ability to separate speech from noise in more complex environments. Fig. 28(b) and 28(c) show that although speech intelligibility and quality decrease with an increasing number of mixed voices, the fluctuations in STOI and PESQ remain less than 2% below the average. Overall, *mmMUSE* demonstrates enhanced capability in more complex environments from a relative improvement perspective. These results underscore *mmMUSE*'s robustness and high adaptability to multi-speaker scenarios.

5 REAL-WORLD APPLICATIONS

5.1 Preliminary Evaluation

5.1.1 Multi-position Evaluation. When users hold a device or interact with a fixed one, the device is often not directly aligned with the vocal cord region. Although Section 4.3.2 explores various user-device angles and distances, the FOV of a One-TX and One-RX radar remains limited. Therefore, we further evaluate common device placement configurations during real-world usage under -5 dB SNR, as illustrated in Fig. 29, covering a horizontal FOV of 80° and a vertical FOV of 40°. As shown in Fig. 30, across 12 different positions, *mmMUSE* maintains an average SISDR of 12.93 dB, STOI of 0.86, and PESQ of 1.81. At position 12, where the radar's

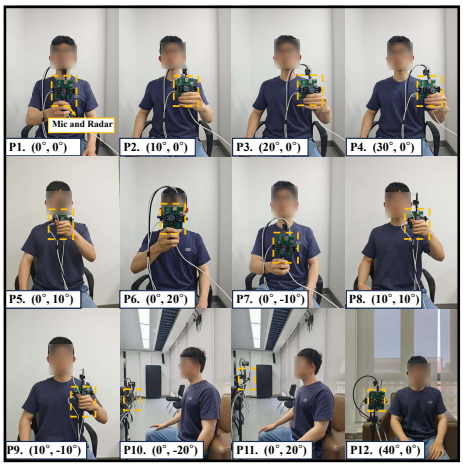

Fig. 29. Device placement configurations in real-world usage scenarios. The angles in parentheses indicate the horizontal and vertical AoA.

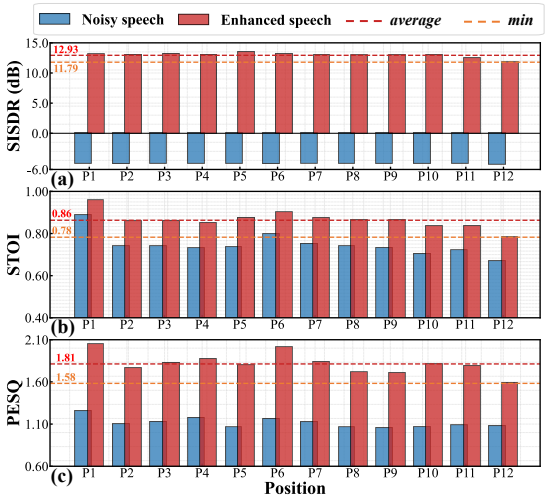

Fig. 30. System performance across various device placement settings.

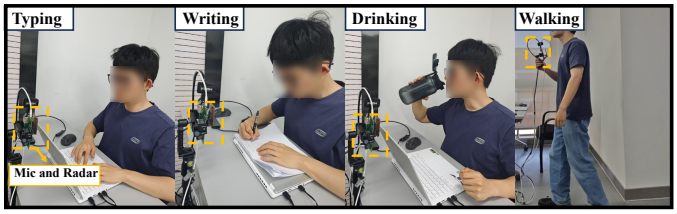

Fig. 31. Experimental setup for different types of natural user motions.

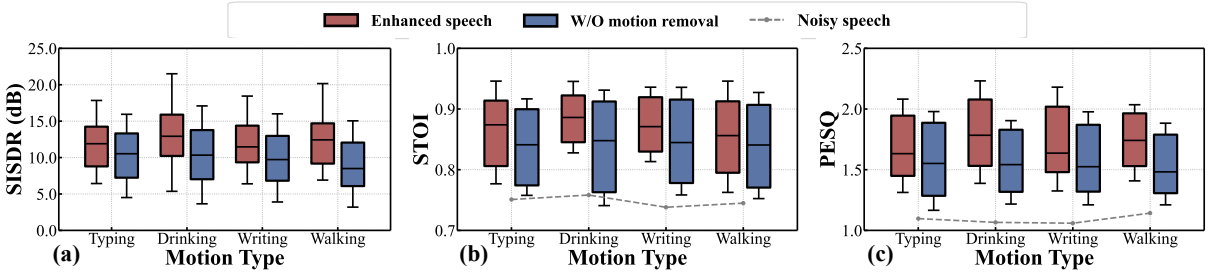

Fig. 32. System performance under different natural user motions.

horizontal angle of arrival (AOA) reaches approximately 40°, *mmMUSE*'s radar approaches its sensing limit, resulting in the lowest observed performance.

5.1.2 Multi-motion Evaluation. In Section 4.3.3, we evaluated the impact of different user-device motion velocities on the performance of *mmMUSE*. Considering the diversity of user movements during real-world device interaction, we further assess how various natural user motions affect system performance. As shown in Fig. 31, the motion types include typing, writing, drinking, and walking. These actions are performed naturally and

involve not only torso movement but also head motions, facial expressions, and limb activity, all under -5 dB SNR. As shown in Fig. 32, the SISDR during typing and writing is slightly lower, likely due to inevitable radar occlusion caused by head tilting. In addition, when the radar is hand-held, substantial swinging can cause the sensing region to shift away from the vocal cord area, thereby degrading enhancement performance. However, across all four natural motion types, *mmMUSE* still maintains an SISDR above 12.0 dB, achieving improvements of 0.12 in STOI and 0.70 in PESQ compared to the average noisy input scores of 0.74 and 1.09, respectively. Furthermore, relative to the baseline without motion interference removal, *mmMUSE* yields an average gain of 2.5 dB. These results further demonstrate the robustness of *mmMUSE* against common body movements.

5.2 Deployment Feasibility

Through the above experimental analysis, we evaluated *mmMUSE*'s performance with respect to sensing range, angular coverage, relative motion, and clothing occlusion. The results are as follows:

- **Sensing range.** Due to power limitations, *mmMUSE* achieves a range coverage of 3.0 m, with mmWave failing beyond approximately 1.5 m.
- **Angular coverage.** The SISO radar effectively enhances speech within an azimuth range of -60° to +60° and an elevation range of -60° to +15°, supported by a horizontal FOV of 80° and a vertical FOV of 40°.
- **Relative motion.** *mmMUSE* remains robust to natural user–device relative motion, maintaining stable performance across velocities from 0.2 to 1.0 m/s.
- **Clothing occlusion.** Leveraging the penetrability of mmWave, *mmMUSE* is resilient to everyday clothing occlusion, operating reliably through fabric up to 2.0 cm thick.

Within the sensing coverage described above, mmWave sensing in *mmMUSE* effectively enhances speech and consistently outperforms audio-based models. Beyond this coverage, *mmMUSE* remains capable of leveraging audio alone to achieve performance comparable to the state-of-the-art audio-based models. Therefore, the mmWave–audio fusion design of *mmMUSE* proves both practical and effective in a variety of *open-field scenarios*.

Mobile Devices. One primary scenario is mobile devices such as smartphones and smartwatches. For example, users may engage in video calls on smartphones or voice interactions on smartwatches integrated with *mmMUSE*. In noisy environments (e.g., subways, shopping malls, or busy streets), conventional microphones are easily overwhelmed by background noise, and speech quality deteriorates further when users wear masks. mmWave can sense vocal vibrations and effectively penetrate obstructions such as face masks. At typical user–device distances (< 0.5 m) with approximately ±30° of angular variation, it provides a reliable complementary cue that enhances call clarity and speech recognition accuracy.

Fixed Devices. A second scenario involves fixed devices, notably in-car assistants. A radar mounted on the dashboard or display (usually within 1 m of the speaker) remains robust to engine, wind, and cabin reverberation noise that degrades purely audio-based methods. Alternatively, during meetings in conference rooms, the radar can be integrated into fixed microphone devices. It is worth noting that at longer deployment distances (> 1.5 m) and extreme angles (60° or more), the radar becomes less effective in capturing vocal vibrations due to power constraints. However, the fusion of audio allows *mmMUSE* to maintain enhancement performance comparable to advanced audio-based baselines.

As detailed below, we conduct real-world evaluations in public spaces and vehicle cabins, confirming consistent speech enhancement across these settings.

5.3 Real-World Evaluation

Speech recognition is extensively used in IoT applications, from smartphone interactions to driver-vehicle voice communications. These applications process speech using automatic speech recognition (ASR) models to

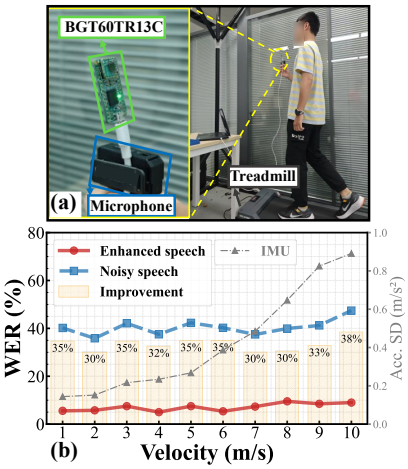

Fig. 33. Running performance.

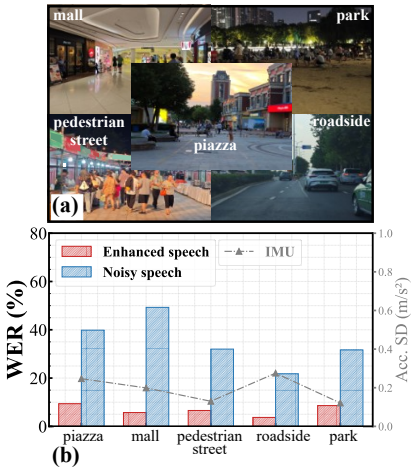

Fig. 34. Public-area performance.

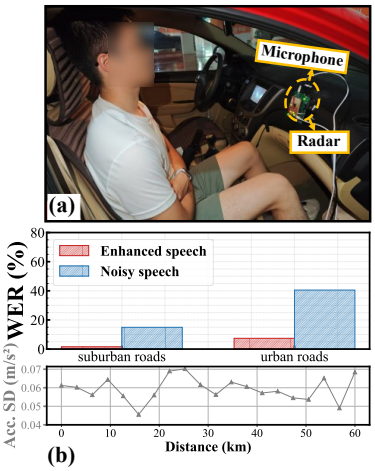

Fig. 35. In-vehicle performance.

transcribe it into text. In this study, we evaluate the enhanced speech with the ASR tool Whisper [72], using word error rate (WER) to gauge enhancement effectiveness. Notably, the Whisper model is pre-trained and was not specifically trained on our dataset. Firstly, we evaluate *mmMUSE* in real-world scenarios involving participants running on a treadmill and walking in noisy public spaces. Additionally, we explore *mmMUSE*'s performance in driving, which involves fine-grained vibration and dynamic noises on different roads. We demonstrate motion extent using the standard deviation of acceleration data recorded by the Inertial Measurement Unit (IMU).

5.3.1 Running with Handheld Device. To test *mmMUSE*'s resilience to motion in real-world scenarios, participants exercise on a treadmill at speeds ranging from 1 km/h to 10 km/h, with the experimental scene shown in Fig. 33(a). Fig. 33(b) shows that as the movement speed increases, there is a slight decline in *mmMUSE*'s performance. This is because an increase in speed results in greater body movement, as indicated by a rise in the acceleration standard deviation (Acc. SD). However, enhanced speech still maintains a WER below 10%, representing improvements of over 30% compared to noisy speech. This substantiates *mmMUSE*'s adaptability to common movements.

5.3.2 Public Spaces. We evaluate *mmMUSE* in real-world environments across five settings: piazzas, malls, pedestrian streets, roadsides, and parks. Participants freely speak while walking through these scenarios, as illustrated in Fig. 34(a). Fig. 34(b) shows that average WER improvements in the five scenarios are 30.47%, 43.57%, 25.42%, 18.06%, and 23.05%, respectively. *mmMUSE* effectively enhances speech in various real-world noisy environments, with the improvement more pronounced as the level of ambient noise increases. Even with considerable scene differences, there are only slight variations in speech recognition accuracy for the enhanced speech, demonstrating the universality of *mmMUSE* across diverse real-world scenarios.

5.3.3 In-Vehicle. To assess *mmMUSE*'s performance in enclosed environments with fine-grained vibration and varying noise levels, we conduct an in-vehicle experiment. Participants freely speak while driving, with the mmWave radar and microphone placed on the dashboard, as illustrated in Fig. 35(a). The vehicle travels roughly 60 km through parkways, tunnels, freeways, overpass bridges, and noisy urban areas, reaching a maximum speed of approximately 70 km/h. These scenarios featured varying degrees of motion and noise due to different speeds and road types shown as Acc. SD. As shown in Fig. 35(b), urban roads feature higher noise levels compared to suburban roads, resulting in a higher WER. Nevertheless, compared to noisy speech, the results from both roads show improvements of 13.41% and 33.13%, respectively, demonstrating *mmMUSE*'s effective noise separation capabilities under varying noise conditions.

6 EVALUATION OF COMPUTATIONAL REQUIREMENTS

Model Parameters. As shown in Table 3, *mmMUSE* is lightweight, with approximately 2.1 M parameters. Compared to audio-based models such as MP-SENet, PHASEN, and Resemble Enhance, *mmMUSE* has fewer parameters. This is partly due to the complementary role of mmWave signals in enhancing speech, enabling *mmMUSE* to achieve better performance with fewer parameters. Compared to cross-modal models like UltraSpeech, UltraSE, and RADIOSES, *mmMUSE* also has fewer parameters, highlighting the efficiency of our two-stage speech enhancement network. With the increasing computational power and memory capacity of modern smart devices [98], models with millions of parameters have become widely applicable.

Time Consumption. We perform latency tests on *mmMUSE*, including preprocessing (motion interference removal and VAD) time and model inference time. The preprocessing phase of *mmMUSE* is tested on two different hardware setups. One setup, released in 2019, features an Intel(R) Core(TM) i7-9750H CPU [12], and the other, released in 2024, uses an AMD Ryzen AI 9 HX 370 [2]. Our preprocessing stage can handle streaming data using time windows, thus we only need to test the latency within a single time window, empirically set to the duration of 600 chirps (60 ms). We run 100 iterations of mmWave and audio signals within the time window on each hardware setup, yielding latencies of 3.51 ms (± 0.51 ms) and 2.12 ms (± 0.41 ms), respectively. Additionally, we deploy the *mmMUSE* model on a GPU (NVIDIA GeForce RTX 3090) to test inference time on 3.6 s mmWave–audio pairs. As shown in Table 3, the inference time of *mmMUSE* is comparable to that of both audio-based and cross-modal models, with a latency of 43.72 ms (± 0.26 ms). We further evaluate the inference latency of *mmMUSE* on the Intel i7-9750H CPU and the AMD Ryzen AI 9 HX 370, obtaining results of 1.05 s and 0.53 s, respectively. Notably, recent edge AI chips (e.g., NVIDIA Jetson AGX Orin, Google Coral) have demonstrated superior inference performance compared to traditional desktop processors [25, 62]. Therefore, *mmMUSE* is suitable for on-device deployment while meeting real-time requirements, with latency remaining below the duration of input data clips.

7 RELATED WORK

7.1 mmWave-based Sensing

In recent years, the development of mmWave radar has made it possible to sense environment [33, 57, 70, 96] and human [88, 94, 95]. Yang et al. [96] proposed a non-intrusive approach for multi-point vibration monitoring using mmWave radar. The mmWave technology is also utilized for place recognition [57], vehicle detection [33], and imaging [70]. Moreover, mmWave radar supports various human perceptions, such as gait recognition [88], face authentication [94], and pose estimation [95].

Table 3. Parameters and time consumption.

	Parameter	Inference	Input Modality[2]
MP-SENet [53]	2.2 M	98.59 ms (± 0.18 ms)	A
PHASEN [97]	7.7 M	19.40 ms (± 0.44 ms)	A
Resemble Enhance [74]	10.6M	67.56 ms (± 3.29 ms)	A
UltraSpeech [15]	4.9 M	/	A+U
UltraSE [79]	18.6 M	/	A+U
RADIOSES [65]	2.1 M [1]	21.67 ms (± 0.83 ms)	A+M
mmMUSE	2.1 M	43.72 ms (± 0.26 ms)	A+M

[1] The number of parameters in the RADIOSES model that we reproduced is 4.0 M.
[2] The abbreviations are as follows: A represents audio, U represents ultrasound, and M represents mmWave.

Despite significant progress in mmWave sensing, its application in fine-grained perception remains fundamentally limited by the requirement for stable conditions. For tasks such as respiration and heartbeat detection [10], motion interference can cause mmWave sensing failure. A recent study [99] proposed using static objects, such as walls in the surrounding environment, as references to mitigate device motion. While effective, this method is only applicable to periodic target motion and cannot handle body motion, as the reflected signals from static objects lack body movement information. Additionally, Chang et al. [11] proposed using reflected signals from non-target areas to counteract motion interference, adapting to non-periodic motion. However, this approach is limited by the need for similar motion patterns in both target and non-target areas, risking failure. Both methods also require Multi-Input Multi-Output (MIMO) mmWave radar, which increases cost and latency. Our approach does not require external references. Instead, it utilizes the prominent Doppler shifts in reflected signals to compensate for user-device motion, enabling the use of SISO mmWave radar.

7.2 Audio-based Speech Enhancement

Traditional statistical methods [45, 92] rely on prior noise knowledge, limiting effectiveness, while microphone array techniques [51, 82] isolate noises using spatial information but introduce distortions and have spatial constraints. Recently, learning-based speech enhancement methods have demonstrated improved performance. Applications of deep learning in speech enhancement can be broadly categorized into time-domain and T-F domain approaches. Time-domain methods [54, 68, 69] map noisy speech to clean speech using encoder-decoder architectures but lack frequency-domain details. In contrast, T-F domain methods [38, 55] estimate T-F masks from spectrograms, effectively integrating both temporal and spectral information for superior performance. Moreover, MP-SENet [53] and PHASEN [97] use dual-stream networks to process both amplitude and phase for speech enhancement. Compared to time-domain methods, T-F domain approaches are more sensitive to SNR variations and can achieve better performance, even with limited datasets [100]. In our work, we employ a complex-valued network to generate T-F masks capturing both phase and amplitude components of speech.

7.3 Cross-modal Voice Perception

With the development of multimodal fusion, modalities such as vision, ultrasound, and mmWave are increasingly being utilized for speech enhancement due to their immunity to acoustic noise. Additionally, some studies [17, 30] have used in-ear acoustic signals, which are less sensitive to environmental noise, as complementary inputs to out-ear microphones for speech enhancement. Additionally, accelerometers have been used to capture bone-conducted vibration signals that are immune to acoustic noise for speech enhancement [35, 81]. However, these methods require contact-based devices (e.g., headphones), which impose additional burdens on users and are not suitable for touchless scenarios.

Vision, like lip movements [23, 78], has been used as a complementary modality to enhance speech. However, these methods pose privacy concerns and face challenges in synchronizing visual and audio elements. Alternative modalities, such as ultrasound [15, 79], enhance speech by detecting articulatory gestures, overcoming the limitations of visual methods. However, they are constrained by spatial and angular limitations, requiring the user's mouth to be close to the microphone [15, 79]. Compared to the aforementioned complementary modalities, mmWave offers a larger detection range and the ability to capture finer-grained sound information. mmWave radar enables the detection of minute displacements and remote capture of the vocal vibration [93] or other passive surfaces [50], offering touchless sound recovery. mmWave radar applications extend beyond speech recognition [52, 102] and enhancement [65] to include speaker recognition [49] and voice activity detection [64], highlighting its versatility in source identification and noise rejection. Despite their ability to denoise, these methods have not fully addressed the common issue of motion interference everywhere in everyday life. Compared

to these methods, *mmMUSE* eliminates motion interference in the mmWave modality and fuses audio signals to enhance speech in complex scenarios.

8 DISCUSSION

8.1 Future Applications

Cost-Effective. With advances in mmWave technology, radar costs have gradually decreased. For example, the ICLEGEND MICRO radar is priced at around $2 [58], the Infineon radar at $12 [42], and the TI radar at $15 [43]. Thus, *mmMUSE* is theoretically cost-effective.

Low-Power. *mmMUSE* incorporates a SISO mmWave radar alongside a low-power microphone. The MEMS microphone commonly used in smartphones consumes only 1.8 mW [63]. The TI IWR6843ISK radar, used in our implementation, transmits at 12 dBm (around 16 mW), with radar parameters configured to a 7.3% duty cycle (around 1.2 mW). Additionally, we also use the Infineon BGT60TR13C radar, which consumes less than 5 mW.

Integration Trends. Miniature mmWave radars are increasingly integrated into smart mobile devices for interaction [24, 31, 32] and communication [77]. Smartphones, such as Pixel 4 [24], already integrate mmWave radar for human–machine interaction [73], suitable for mobile power and size constraints. In addition, portable interactive devices such as PieX Pendant [32] and AIBI Pocket Pet [31] have adopted low-power radars integrated with edge AI chips. mmWave radar has also been deployed in vehicle cabins for vital sign monitoring [6] and user interaction [61]. These examples reflect a broader industry trend toward integrating radar sensors with edge AI processors for cost-effective, low-power, and on-device deployment. With this growing convergence of radar sensing and edge intelligence, we believe *mmMUSE* is well-aligned with such hardware developments and can be feasibly deployed in future consumer-grade platforms.

8.2 Limitations

Sensing Coverage. In Section 5.2, we summarize the sensing coverage of *mmMUSE*, demonstrating its suitability for typical handheld and fixed device scenarios. However, when the user is at an extreme distance from the radar (> 1.5 m), outside the FOV, or at a relative angle exceeding 60°, mmWave sensing fails to fully capture vocal vibrations. In addition, when the relative velocity exceeds around 2.0 m/s, substantial signal discontinuities arise during the mmWave tracking process, resulting in spectral leakage and incomplete capture of vocal vibration patterns. Moreover, when the vocal cord region is obstructed by metallic materials, other body parts, or clothing thicker than 6.0 cm, mmWave signals fail to penetrate, rendering vocal vibration sensing infeasible. Although beyond this coverage range, *mmMUSE* remains capable of leveraging audio alone to achieve performance comparable to the state-of-the-art audio-based methods.

Multi-user Speech Separation. *mmMUSE* is currently designed to focus on a single target user and has demonstrated effectiveness in multi-speaker scenarios, as shown in Section 4.3.5 and Section 5.3.2. Since SISO mmWave radar lacks angular resolution, speech separation is only feasible when speakers occupy distinct range bins. In extreme cases where a nearby speaker is positioned so close to the target user that their radial distance to the radar falls within the resolution (around 4.2 cm), the system may fail to separate their vocal signals. This issue is further exacerbated by natural body movement, which can effectively reduce spatial separability. To address this limitation, we plan to incorporate Multi-Input Multi-Output (MIMO) mmWave radar in future work, which would enable the distinction of different users based on angular resolution. Therefore, multiple users can share the device, as their respective vocal vibration signals can be distinguished and individually enhanced. Additionally, the network architecture of *mmMUSE* can be extended or parallelized to process multiple differentiated mmWave inputs, enabling speech separation in more complex scenarios.

9 CONCLUSION

We have proposed *mmMUSE*, an mmWave-based motion-resilient universal speech enhancement system. It eliminates motion interference in the mmWave modality and fuses audio signals to enhance speeches. Using datasets from 46 participants, *mmMUSE* outperforms the state-of-the-art speech enhancement models by 26% in SISDR and 34% in STOI on average. It also achieves SISDR improvements of 16.72 dB, 17.93 dB, 14.93 dB, and 18.95 dB in controlled environments involving intense noise, extensive motion, multiple speakers, and various obstructive materials, respectively. Moreover, in real-world scenarios, including running, public spaces, and driving, *mmMUSE* achieves WER below 10%.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62332016) and the Key Research Program of Frontier Sciences, CAS (No. ZDBS-LY-JSC001).

References

[1] Sherif Abdulatif, Ruizhe Cao, and Bin Yang. 2024. CMGAN: Conformer-Based Metric-GAN for Monaural Speech Enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 32 (2024), 2477–2493.

[2] AMD. 2024. AMD Ryzen AI 9 HX 370 Processor. [Online]. https://www.amd.com/en/products/processors/laptop/ryzen/300-series/amd-ryzen-ai-9-hx-370.html.

[3] Apple. 2024. iPhone's Siri. [Online]. https://www.apple.com/siri/.

[4] asteroid team. 2020. asteroid. [Online]. https://github.com/asteroid-team/asteroid/blob/master/asteroid/models/dprnn_tasnet.py.

[5] Li Auto. 2024. Li Auto Introduces Fully Self-Developed MindGPT. https://genaigazette.com/li-auto-introduces-fully-self-developed-mind-gpt/.

[6] Azcom CabinGuard. 2025. A Vehicle In-cabin Monitoring Solution. [Online]. https://www.azcomtech.com/markets/mmwave-radar-sensors/automotive/.

[7] AE Blaugrund. 1966. Notes on Doppler-shift lifetime measurements. *Nuclear Physics* 88, 3 (1966), 501–512.

[8] Ali Braytee, Andy Shueh-Chih Yang, Ali Anaissi, Kunal Chaturvedi, and Mukesh Prasad. 2024. A Novel Dual-Pipeline based Attention Mechanism for Multimodal Social Sentiment Analysis. In *Companion Proceedings of the ACM on Web Conference 2024 (WWW)*. 1816–1822.

[9] Canalys. 2024. The rise of Chinese automakers in 2024. [Online]. https://canalys.com/reports/china-automotive-ev-market.

[10] Zhaoxin Chang, Fusang Zhang, Xujun Ma, Pei Wang, Weiyan Chen, Duo Zhang, Badii Jouaber, and Daqing Zhang. 2024. MmECare: Enabling Fine-grained Vital Sign Monitoring for Emergency Care with Handheld MmWave Radars. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 8, 4 (2024), 1–24.

[11] Zhaoxin Chang, Fusang Zhang, Jie Xiong, Weiyan Chen, and Daqing Zhang. 2024. MSense: Boosting Wireless Sensing Capability Under Motion Interference. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking(MobiCom)*. 108–123.

[12] Intel Corporation. 2019. Intel Core i7-9750H Processor (12M Cache, up to 4.50 GHz). [Online]. https://www.intel.com/content/www/us/en/products/sku/191045/intel-core-i79750h-processor-12m-cache-up-to-4-50-ghz/specifications.html.

[13] DATAREPORTAL. 2024. Digital Around the World. [Online]. https://datareportal.com/global-digital-overview.

[14] Delcroix, Marc and Zmolikova, Katerina and Kinoshita, Keisuke and Ogawa, Atsunori and Nakatani, Tomohiro. 2018. Single channel target speaker extraction and recognition with speaker beam. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 5554–5558.

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

[16] Shaojin Ding, Quan Wang, Shuo-Yiin Chang, Li Wan, and Ignacio Lopez Moreno. 2020. Personal VAD: Speaker-Conditioned Voice Activity Detection. In *The Speaker and Language Recognition Workshop (Odyssey)*. 433–439.

[17] 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 (IMWUT)* 7, 4 (2024), 1–33.

[18] Harvey Fletcher. 1953. Speech and hearing in communication. (1953).

[19] Daniel Fogerty and Diane Kewley-Port. 2009. Perceptual contributions of the consonant-vowel boundary to sentence intelligibility. *The Journal of the Acoustical Society of America* 126, 2 (2009), 847–857.

[20] Daniel Fogerty, Diane Kewley-Port, and Larry E Humes. 2012. The relative importance of consonant and vowel segments to the recognition of words and sentences: Effects of age and hearing loss. *The Journal of the Acoustical Society of America* 132, 3 (2012), 1667–1678.
[21] Ming Gao, Feng Lin, Weiye Xu, Muertikepu Nuermaimaiti, Jinsong Han, Wenyao Xu, and Kui Ren. 2020. Deaf-aid: Mobile IoT communication exploiting stealthy speaker-to-gyroscope channel. In *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 1–13.
[22] John S Garofolo et al. 1988. DARPA TIMIT acoustic-phonetic speech database. *National Institute of Standards and Technology (NIST)* 15 (1988), 29–50.
[23] Mandar Gogate, Kia Dashtipour, and Amir Hussain. 2024. Robust Real-time Audio-Visual Speech Enhancement based on DNN and GAN. *IEEE Transactions on Artificial Intelligence (TAI)* (2024).
[24] Google. 2019. Here's how the Pixel 4's Soli radar works and why Motion Sense has so much potential. [Online]. https://www.androidcentral.com/how-does-googles-soli-chip-work.
[25] Google. 2021. Edge TPU: System-on-Module Performance Benchmarks. https://coral.ai/docs/edgetpu/benchmarks/.
[26] Google. 2021. ok-google.io. [Online]. https://ok-google.io.
[27] Google. 2023. Google WebRTC. [Online]. https://webrtc.org/.
[28] Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. 6645–6649.
[29] Daniel Griffin and Jae Lim. 1984. Signal estimation from modified short-time Fourier transform. *IEEE Transactions on acoustics, speech, and signal processing* 32, 2 (1984), 236–243.
[30] 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 (IMWUT)* 8, 3 (2024), 1–30.
[31] Hardso. 2025. AIBI Pocket. [Online]. https://www.hardso.com/product/2fea3b01-c2a1-4050-8e7e-f9d62f88de88.
[32] Hardso. 2025. PieX Pendant. [Online]. https://www.hardso.com/product/ea05912b-a7fc-4bd5-aee4-f43f8ac4d200.
[33] Chenming He, Chengzhen Meng, Chunwang He, Xiaoran Fan, Beibei Wang, Yubo Yan, and Yanyong Zhang. 2024. See Through Vehicles: Fully Occluded Vehicle Detection with Millimeter Wave Radar. In *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 740–754.
[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 770–778.
[35] Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. 2023. Towards Bone-Conducted Vibration Speech Enhancement on Head-Mounted Wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services (MobiSys)*. 14–27.
[36] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. 2018. TUT urban acoustic scenes 2018, development dataset. (2018).
[37] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*. 7132–7141.
[38] 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. In *Interspeech 2020*. 2472–2476.
[39] Humane. 2024. Humane's AI PIN. [Online]. https://humane.com/aipin.
[40] Infineon. 2020. Radar Fusion GUI. [Online]. https://softwaretools.infineon.com/tools/com.ifx.tb.tool.radarfusiongui.
[41] Infineon. 2023. bgt60tr13c. [Online]. https://www.infineon.com/cms/en/product/sensor/radar-sensors/radar-sensors-for-iot/60ghz-radar/bgt60tr13c.
[42] Infineon. 2023. bgt60utr11aip. [Online]. https://www.infineon.com/cms/en/product/sensor/radar-sensors/radar-sensors-for-iot/60ghz-radar/bgt60utr11aip/.
[43] Texas Instruments. 2024. IWR6843. [Online]. https://www.ti.com/product/IWR6843#order-quality.
[44] Jack Leathem. 2023. Generative AI integrated voice assistants are the lifeline smart speakers need. [Online]. https://www.canalys.com/insights/ai-in-smart-speakers.
[45] Sunil Kamath and Philipos Loizou. 2002. A multi-band spectral subtraction method for enhancing speech corrupted by colored noise. In *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Vol. 4.
[46] Matthias Kronauge and Hermann Rohling. 2013. Fast two-dimensional CFAR procedure. *IEEE Trans. Aerospace Electron. Systems* 49, 3 (2013), 1817–1823.
[47] Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R Hershey. 2019. SDR–half-baked or well done?. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 626–630.
[48] Hui Li and Xiao-Jun Wu. 2024. CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion* 103 (2024), 102147.

[49] Huining Li, Chenhan Xu, Aditya Singh Rathore, Zhengxiong Li, Hanbin Zhang, Chen Song, Kun Wang, Lu Su, Feng Lin, Kui Ren, et al. 2020. Vocalprint: exploring a resilient and secure voice authentication via mmwave biometric interrogation. In *Proceedings of the 18th Conference on Embedded Networked Sensor Systems (SenSys)*. 312–325.

[50] Songxu Li, Yuyong Xiong, Peng Zhou, Zesheng Ren, and Zhike Peng. 2022. MmPhone: Sound recovery using millimeter-wave radios with adaptive fusion enhanced vibration sensing. *IEEE Transactions on Microwave Theory and Techniques* 70, 8 (2022), 4045–4055.

[51] Qing-Guang Liu, Benoît Champagne, and Peter Kabal. 1996. A microphone array processing technique for speech enhancement in a reverberant space. *Speech Communication* 18, 4 (1996), 317–334.

[52] 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 (SenSys)*. 97–110.

[53] Ye-Xin Lu, Yang Ai, and Zhen-Hua Ling. 2023. MP-SENet: A Speech Enhancement Model with Parallel Denoising of Magnitude and Phase Spectra. In *Interspeech 2023*. 3834–3838.

[54] Yi Luo and Nima Mesgarani. 2018. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 696–700.

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

[56] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. 2013. Rectifier nonlinearities improve neural network acoustic models. In *International conference on machine learning (ICML)*, Vol. 30. 3.

[57] Chengzhen Meng, Yifan Duan, Chenming He, Dequan Wang, Xiaoran Fan, and Yanyong Zhang. 2024. mmPlace: Robust Place Recognition With Intermediate Frequency Signal of Low-Cost Single-Chip Millimeter Wave Radar. *IEEE Robotics and Automation Letters (RAL)* 9, 6 (2024), 4878–4885.

[58] ICLEGEND MICRO. 2024. 24GHz mmWave Sensor SoC. [Online]. https://www.iclegend.com/zh-hans/product/category/Sensor.

[59] Rossen Nenov, Dang-Khoa Nguyen, and Peter Balazs. 2023. Faster than fast: Accelerating the Griffin-Lim algorithm. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[60] NIO. 2020. Artificial Intelligence with Warmth: Empathy, Sympathy, and the Human Elements. [Online]. https://www.nio.com/blog/artificial-intelligence-warmth-empathy-sympathy-and-human-elements.

[61] NOVELIC. 2025. A Whole-Cabin Solution for Vehicle Safety and Comfort. [Online]. https://www.novelic.com/acam-automotive-in-cabin-monitoring-radar/.

[62] NVIDIA Corporation. 2022. NVIDIA Jetson AGX Orin Series Product Design Guide. https://developer.nvidia.com/embedded/jetson-agx-orin.

[63] Geoffrey Ottoy, Bart Thoen, and Lieven De Strycker. 2016. A low-power MEMS microphone array for wireless acoustic sensors. In *IEEE Sensors Applications Symposium (SAS)*. 1–6.

[64] Muhammed Zahid Ozturk, Chenshu Wu, Beibei Wang, Min Wu, and K. J. Ray Liu. 2024. RadioVAD: mmWave-Based Noise and Interference-Resilient Voice Activity Detection. *IEEE Internet of Things Journal (IOT)* 11, 15 (2024), 26005–26019.

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

[66] Kuldip Paliwal, Kamil Wójcicki, and Benjamin Shannon. 2011. The importance of phase in speech enhancement. *Speech Commun.* 53, 4 (2011), 465–494.

[67] Ashutosh Pandey and DeLiang Wang. 2019. A new framework for CNN-based speech enhancement in the time domain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 27, 7 (2019), 1179–1188.

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

[69] Santiago Pascual, Antonio Bonafonte, and Joan Serrà. 2017. SEGAN: Speech Enhancement Generative Adversarial Network. In *Interspeech 2017*. 3642–3646.

[70] Akarsh Prabhakara, Tao Jin, Arnav Das, Gantavya Bhatt, Lilly Kumari, Elahe Soltanaghai, Jeff Bilmes, Swarun Kumar, and Anthony Rowe. 2023. RadarHD: Demonstrating Lidar-like Point Clouds from mmWave Radar. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 1–3.

[71] R Gnana Praveen and Jahangir Alam. 2024. Recursive Joint Cross-Modal Attention for Multimodal Fusion in Dimensional Emotion Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4803–4813.

[72] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning (ICML)*. 28492–28518.

[73] Google Research. 2020. Soli Radar-based Perception and Interaction in Pixel 4. [Online]. https://research.google/blog/soli-radar-based-perception-and-interaction-in-pixel-4/.

[74] RESEMBLE.AI. 2023. Introducing Resemble Enhance: Open Source Speech Super Resolution AI Model. [Online]. https://www.resemble.ai/introducing-resemble-enhance/.

[75] Antony W Rix, Michael P Hollier, Andries P Hekstra, and John G Beerends. 2002. Perceptual Evaluation of Speech Quality (PESQ) The New ITU Standard for End-to-End Speech Quality Assessment Part I–Time-Delay Compensation. *Journal of the Audio Engineering Society* 50, 10 (2002), 755–764.
[76] S. Team. 2021. Silero VAD: pre-trained enterprise-grade voice activity detector (VAD), number detector and language classifier. [Online]. https://github.com/snakers4/silero-vad.
[77] Samsung. 2021. mmWave 5G: Past, Present and Future. [Online]. https://www.samsung.com/global/business/networks/insights/blog/0218-mmwave-5g-past-present-and-future/.
[78] Ke Sun, Chun Yu, Weinan Shi, Lan Liu, and Yuanchun Shi. 2018. Lip-interact: Improving mobile device interaction with silent speech commands. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology (UIST)*. 581–593.
[79] Ke Sun and Xinyu Zhang. 2021. UltraSE: single-channel speech enhancement using ultrasound. In *Proceedings of the 27th annual international conference on mobile computing and networking (MobiCom)*. 160–173.
[80] Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. 2010. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 4214–4217.
[81] Marco Tagliasacchi, Yunpeng Li, Karolis Misiunas, and Dominik Roblek. 2020. SEANet: A multi-modal speech enhancement network. In *Interspeech 2020*.
[82] Vincent Mohammad Tavakoli, Jesper Rindom Jensen, Mads Græsbøll Christensen, and Jacob Benesty. 2016. A framework for speech enhancement with ad hoc microphone arrays. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)* 24, 6 (2016), 1038–1051.
[83] TI. 2020. IWR6843 intelligent mmWave sensor standard antenna plug-in module. [Online]. https://www.ti.com/tool/IWR6843ISK.
[84] TI. 2020. mmWave Studio. [Online]. https://www.ti.com/tool/MMWAVE-STUDIO.
[85] TI. 2020. Real-time data-capture adapter for radar sensing evaluation module. [Online]. https://www.ti.com/tool/DCA1000EVM.
[86] Chiheb Trabelsi, Olexa Bilaniuk, Ying Zhang, Dmitriy Serdyuk, Sandeep Subramanian, Joao Felipe Santos, Soroush Mehri, Negar Rostamzadeh, Yoshua Bengio, and Christopher J Pal. 2018. Deep Complex Networks. In *International Conference on Learning Representations (ICLR)*.
[87] DeLiang Wang and Jitong Chen. 2018. Supervised speech separation based on deep learning: An overview. *IEEE/ACM transactions on audio, speech, and language processing (TASLP)* 26, 10 (2018), 1702–1726.
[88] Dequan Wang, Xinran Zhang, Kai Wang, Lingyu Wang, Xiaoran Fan, and Yanyong Zhang. 2024. RDGait: A mmWave Based Gait User Recognition System for Complex Indoor Environments Using Single-chip Radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 8, 3 (2024), 1–31.
[89] Wikipedia. 2024. Voice frequency - Wikipedia. [Online]. https://en.wikipedia.org/wiki/Voice_frequency.
[90] Donald S Williamson, Yuxuan Wang, and DeLiang Wang. 2015. Complex ratio masking for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing (TASLP)* 24, 3 (2015), 483–492.
[91] 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.
[92] Mingyang Wu and DeLiang Wang. 2006. A two-stage algorithm for one-microphone reverberant speech enhancement. *IEEE Transactions on Audio, Speech, and Language Processing (TASLP)* 14, 3 (2006), 774–784.
[93] Chenhan Xu, Zhengxiong Li, Hanbin Zhang, Aditya Singh Rathore, Huining Li, Chen Song, Kun Wang, and Wenyao Xu. 2019. Waveear: Exploring a mmwave-based noise-resistant speech sensing for voice-user interface. In *Proceedings of the 17th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 14–26.
[94] Weiye Xu, Wenfan Song, Jianwei Liu, Yajie Liu, Xin Cui, Yuanqing Zheng, Jinsong Han, Xinhuai Wang, and Kui Ren. 2022. Mask does not matter: Anti-spoofing face authentication using mmWave without on-site registration. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 310–323.
[95] Hongfei Xue, Qiming Cao, Chenglin Miao, Yan Ju, Haochen Hu, Aidong Zhang, and Lu Su. 2023. Towards generalized mmwave-based human pose estimation through signal augmentation. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 1–15.
[96] Yanni Yang, Huafeng Xu, Qianyi Chen, Jiannong Cao, and Yanwen Wang. 2023. Multi-Vib: Precise multi-point vibration monitoring using mmWave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 6, 4 (2023), 1–26.
[97] Dacheng Yin, Chong Luo, Zhiwei Xiong, and Wenjun Zeng. 2020. Phasen: A phase-and-harmonics-aware speech enhancement network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. 9458–9465.
[98] yolegroup. 2024. Smartphone memory: Gen AI upgrades to drive spike in DRAM demand. [Online]. https://www.yolegroup.com/technology-outlook/smartphone-memory-gen-ai-upgrades-to-drive-spike-in-dram-demand/.
[99] Fusang Zhang, Jie Xiong, Zhaoxin Chang, Junqi Ma, and Daqing Zhang. 2022. Mobi2Sense: empowering wireless sensing with mobility. In *Proceedings of the 28th Annual International Conference on Mobile Computing and Networking (MobiCom)*. 268–281.
[100] Xiao-Lei Zhang and DeLiang Wang. 2016. A deep ensemble learning method for monaural speech separation. *IEEE/ACM transactions on audio, speech, and language processing (TASLP)* 24, 5 (2016), 967–977.

[101] Yongxian Zhang, Chaozhen Lan, Haiming Zhang, Guorui Ma, and Heng Li. 2024. Multimodal Remote Sensing Image Matching via Learning Features and Attention Mechanism. *IEEE Transactions on Geoscience and Remote Sensing (TGARS)* 62 (2024), 1–20.

[102] Running Zhao, Jiangtao Yu, Hang Zhao, and Edith CH Ngai. 2023. Radio2Text: Streaming Speech Recognition Using mmWave Radio Signals. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 7, 3 (2023), 1–28.

[103] Chengshi Zheng, Huiyong Zhang, Wenzhe Liu, Xiaoxue Luo, Andong Li, Xiaodong Li, and Brian CJ Moore. 2023. Sixty years of frequency-domain monaural speech enhancement: From traditional to deep learning methods. *Trends in Hearing* 27 (2023).