Toward Flexible Phone Call Quality Enhancement via Contactless Bone-Conduction Sensing with mmWave

Lingyu Wang[†*], You Zuo[§*], Chenming He[†✉], Dequan Wang[†], Chengzhen Meng[†], Xiaoran Fan[‡], Yanyong Zhang[†◇✉]

[†]University of Science and Technology of China, [§]University of Melbourne, [‡]Independent Researcher, [◇]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

ABSTRACT

Phone calls are a fundamental form of voice communication, yet microphone audio remains highly vulnerable to airborne noise. To enable flexible phone-call enhancement, we present *BoneWave*, which uses contactless mmWave sensing in the near-ear setting to capture bone-conducted vibrations, complementing microphone audio. The primary challenge is the pronounced spectral artifacts in mmWave-extracted vibration signals under near-ear settings. We trace these artifacts to a previously underexplored mechanism: near-field multipath propagation and user micromotions jointly disrupt signal continuity. To mitigate such artifacts, we design a user-adaptive initialization scheme that tunes radar configurations to consolidate reflected signals and preserve signal continuity. Meanwhile, to mitigate interference from handheld activities, we detect motion-active periods and suppress the resulting signal distortions efficiently. Moreover, we optimize the speech enhancement pipeline through data-side refinement and a lightweight transformer-based model, enabling efficient fusion of detail-rich audio and noise-robust mmWave cues. Extensive evaluations with 54 users verify the effectiveness of *BoneWave*. Under severe noise, it improves STOI and PESQ by 28% and 74%, respectively, increases SI-SDR by 14.57 dB, and outperforms all five state-of-the-art baselines. It also remains robust across diverse daily-use conditions and generalizes well to real-world scenarios, showing potential for practical deployment. Our data, model, and demo are available on **GitHub**.

CCS CONCEPTS

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

KEYWORDS

Speech Enhancement, mmWave Bone-Conduction Sensing

ACM Reference Format:

Lingyu Wang, You Zuo, Chenming He, Dequan Wang, Chengzhen Meng, Xiaoran Fan, Yanyong Zhang. 2027. Toward Flexible Phone Call Quality Enhancement via Contactless Bone-Conduction Sensing with mmWave. In *ACM/IEEE International Conference on Embedded Artificial Intelligence and Sensing Systems (SenSys '27), May 17–20, 2027, Boulder, CO, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3862575.3862839

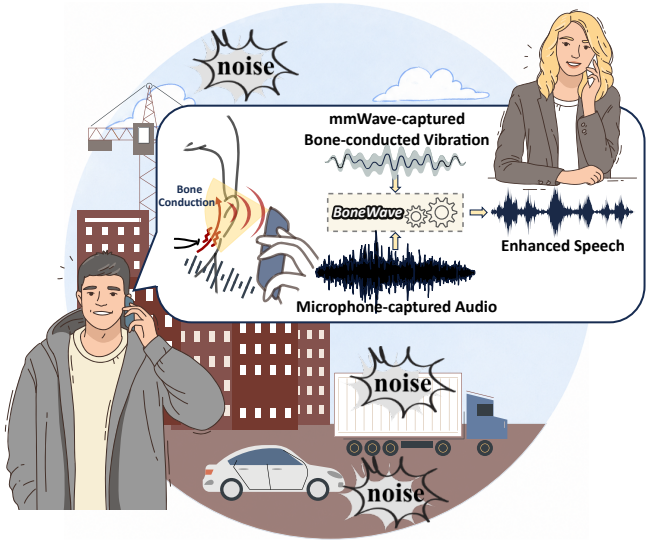


Figure 1: Applications of *BoneWave* for flexible noisy phone-call enhancement using microphone audio and mmWave-captured bone-conducted vibrations.

1 INTRODUCTION

Phone calls are among the most common forms of voice communication, spanning everyday conversations, work collaboration, and emergency situations. Their quality directly affects the efficiency and reliability of information exchange. In real-world use, however, phone calls rely primarily on microphones to capture airborne speech, making them highly vulnerable to acoustic noise, especially under low signal-to-noise ratio (SNR) conditions [1, 31], such as public crowds and urban traffic. This motivates the use of noise-robust complementary modalities for speech enhancement. Unlike face-to-device interactions, phone calls impose a constrained sensing geometry: the device is naturally handheld near the ear, requiring complementary modalities to remain reliable at the ear side and robust during handheld operation.

Compared with noise-sensitive microphone speech, articulatory gestures such as lip motions provide noise-robust auxiliary cues through cameras [4, 38] or ultrasound [1, 39]. Yet, these approaches rely on near-frontal sensing and thus do not fit the near-ear posture of handheld phone calls. In contrast, bone-conducted vibrations are concentrated around the ear and largely insensitive to airborne acoustic noise [52], making them well suited for this setting. Existing methods [14, 46] mainly capture such vibrations with inertial

*Equal contribution. ✉ Corresponding authors.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.
SenSys '27, May 17–20, 2027, Boulder, CO, USA
© 2027 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-3124-2/2027/05
https://doi.org/10.1145/3862575.3862839

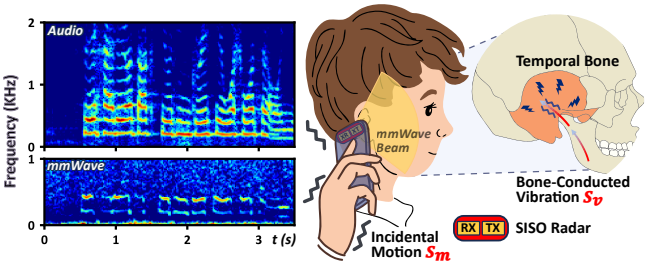


Figure 2: Contactless mmWave bone-conduction sensing during natural handheld phone calls.

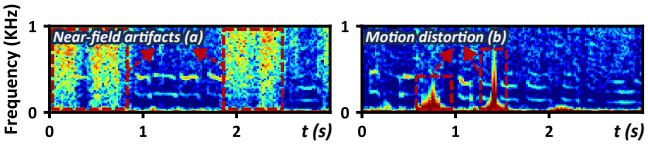


Figure 3: Near-field artifacts and motion distortion.

measurement units (IMUs), requiring stable device-skin contact with consistent tightness, placement, and orientation. In natural handheld calls, users often shift the phone, making contact intermittent or variable; user activities can also redistribute vibrations across axes, degrading signal quality. These requirements burden users and limit handheld-use flexibility. *An ideal solution should remain transparent to the user while fitting seamlessly into natural handheld calling.* To this end, we turn to millimeter-wave (mmWave) sensing, whose ability to capture minute displacements enables contactless bone-conduction sensing, as shown in Fig. 2. This opens the door to flexible and robust phone-call quality enhancement.

Although prior work has applied mmWave sensing to speech recognition [21, 53] and enhancement [30, 45], it primarily relies on frontal sensing of vocal cord vibrations and operates beyond 10 cm. In contrast, the constrained sensing geometry of phone calls requires ear-side sensing of bone-conducted vibrations at a much closer range of 1-3 cm. Such sensing is robust to airborne noise, yet near-field multipath and handheld activities can still interfere with it. We find that this near-field sensing introduces significant spectral artifacts into the mmWave phase variations that encode vibration information, as shown in Fig. 3(a). We trace these artifacts to a previously underexplored mechanism: *near-field multipath spatially spreads users' micromotion reflections, causing intermittent dispersion across adjacent range bins.* This severely disrupts phase continuity, introducing these artifacts. Since multipath is unavoidable in near-field use, this mechanism motivates a natural mitigation strategy: tuning radar configurations to increase the spatial coverage of each bin, thereby keeping dispersed reflections within one bin and preserving phase continuity. However, overly large bins may capture off-target reflections, degrading vibration quality. Moreover, due to differences in users' head geometries, the optimal configuration varies across users. Meanwhile, mmWave sensing is highly sensitive to motion in handheld use (Fig. 3(b)), and resource-constrained phones further require real-time and lightweight processing.

In this paper, we propose *BoneWave*, a flexible phone-call enhancement system that uses contactless mmWave bone-conduction sensing to enhance noisy speech, as shown in Fig. 1. To address the key challenges in this scenario, we introduce the following designs.

(I) Initializing an appropriate radar configuration under user-dependent near-field conditions. For artifacts caused by dispersed multipath-coupled micromotion reflections (Fig. 3(a)), appropriate bin coverage can mitigate them by consolidating reflected signals to preserve continuity while reducing off-target reflections. Yet, the optimal configuration varies across users due to differences in head geometry. To investigate this, we experimentally analyze users with substantially different head sizes and observe that individual optima shift within a common effective bandwidth range. Based on this, we design a user-adaptive sensing initialization procedure requiring only an 8-second speech sample before normal use. During initialization, the system automatically sweeps bandwidths to identify a suitable configuration within the effective range. To automate this, we design a speech-mask estimation method based on the frequency distribution of bone-conducted vibrations to extract speech-related components, and propose an artifact-sensitive quality metric to identify the optimal configuration.

(II) Suppressing motion interference introduced by handheld operation. During handheld phone calls, small but rapid incidental hand or head motions introduce distortion overlapping with bone-conducted vibrations (Fig. 3(b)). Prior work [47] attributes this distortion to coupling between motion and the radar's zero-frequency direct-current (DC) component. The key to suppression is to detect motion periods and remove DC only within them. We thus design an efficient motion detection method based on phase increments: motion induces pronounced phase variations, whereas static periods remain stable. Specifically, we quantify phase-increment magnitudes within each sliding window to detect motion-active periods and confine DC removal to them. This time-domain detection avoids the long-segment time-frequency transform required by prior work, reducing computational overhead.

(III) Supporting lightweight speech enhancement for real-time deployment. To enable efficient speech enhancement on resource-constrained phones, we optimize the pipeline from both the data and model sides. On the data side, radar noise floors in mmWave signals burden feature extraction, while long non-speech periods cause redundant computation. We thus separate out bone-conducted vibrations as a cleaner mmWave input and use them to gate subsequent processing to speech-active segments. On the model side, we aim to exploit the complementarity between detail-rich audio and noise-robust mmWave cues within a mobile-friendly size. We design a lightweight transformer-based model with only 1.7M parameters, using cross-attention to selectively extract complementary cross-modal cues and derive informative fused representations. This effective fusion allows the subsequent network to remain compact while achieving strong performance.

In conclusion, our contributions are as follows:

- To the best of our knowledge, *BoneWave* is the first system that reliably enables contactless near-field mmWave sensing of bone-conducted vibrations around the ear, complementing audio for flexible phone-call quality enhancement.
- We identify multipath-coupled micromotion reflection dispersion as the source of artifacts in near-field mmWave sensing. Accordingly, we enable effective sensing via user-adaptive initialization that tunes radar configurations to appropriately consolidate reflected signals, thereby mitigating these artifacts.

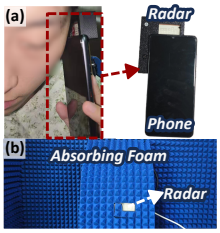


Figure 4: (a) Handheld use (1-3 cm); (b) multipath suppression with absorbing foam.

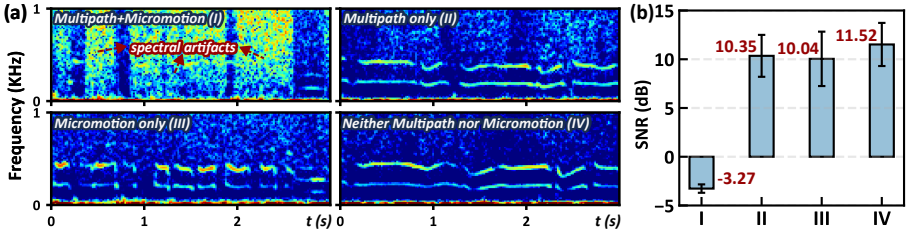


Figure 5: Controlled analysis of near-field spectral artifacts. (a) mmWave vibration spectrograms under four controlled conditions. Spectral artifacts appear only when multipath and micromotion coexist; (b) corresponding average-power SNR.

- For handheld calls on resource-constrained phones, we develop a pipeline optimized for efficiency at each stage. It combines rapid motion-distortion suppression with an enhancement framework using data-side refinement and a lightweight transformer-based model to efficiently fuse complementary multimodal cues.
- We extensively evaluate *BoneWave* with 54 users. Under severe noise, it improves STOI and PESQ by 28% and 74%, respectively, increases SI-SDR by 14.57 dB, and outperforms five state-of-the-art baselines. It remains robust across unseen users and under daily-use conditions, including device variations, placement changes, handheld motion, and sensing obstructions from sweat, hair, and clothing. It also generalizes well to noisy real-world scenarios, including traffic, streets, and subways, reducing WER by 50.6 percentage points on average.

2 CONTACTLESS BONE-CONDUCTED VIBRATION SENSING

During speech, vocal vibrations propagate through the skull and concentrate around the peri-auricular region (Fig. 2). In this work, we use mmWave sensing to capture these bone-conducted vibrations contactlessly through phase variations, $\Delta\varphi_v = \frac{4\pi\delta}{\lambda}$, where δ denotes vibration displacement. Since $\Delta\varphi_v$ is linearly proportional to δ, it preserves vibration dynamics, including the fundamental frequency and harmonics.

2.1 Near-Field Spectral Artifacts

In near-ear calls, users typically hold the phone within 1-3 cm, where mmWave sensing exhibits pronounced spectral artifacts (Fig. 3(a)). To understand their origin, we model near-field reflections and conduct proof-of-concept experiments.

Modeling Near-Field Reflections. During handheld use, incidental micromotion arises from hand and head tremors, as well as facial movements (Fig. 2). At close range, reflections from both micromotion $S_m(t)$ and bone-conducted vibration $S_v(t)$ undergo near-field multipath. The received mmWave signal is modeled as $S(t) = \mathcal{M}\big(S_m(t), S_v(t)\big)$, where $\mathcal{M}(\cdot,\cdot)$ denotes the multipath operator. After Range-FFT, reflected components are resolved into discrete range bins, each corresponding to a radial distance interval from the radar. Under near-field multipath, signal components associated with a localized bin may be irregularly and intermittently redistributed across neighboring bins over time. Specifically, for the energy-dominant bin that carries rich user information, the signal is a superposition of continuous direct reflections and temporally

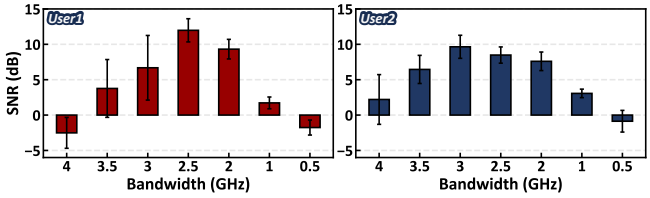


Figure 6: Results under different radar bandwidths.

intermittent multipath reflections:

$$S_{\text{bin}}(t) = \alpha_0^{(m)} e^{j\phi_0^{(m)}(t)} + \alpha_0^{(v)} e^{j\phi_0^{(v)}(t)} + \sum_{l=1}^{L_m} \gamma_l^{(m)}(t)\alpha_l^{(m)} e^{j\phi_l^{(m)}(t)} + \sum_{q=1}^{L_v} \gamma_q^{(v)}(t)\alpha_q^{(v)} e^{j\phi_q^{(v)}(t)}, \quad (1)$$

where the first two terms are the direct reflections of micromotion and bone-conducted vibration, respectively, and $\gamma_l^{(m)}(t), \gamma_q^{(v)}(t) \in \{0,1\}$ indicate the temporal presence of the multipath reflections. Since each bin spans a spatial region, it may capture reflections from multiple positions. Bone-conducted vibration follows a consistent pattern across the peri-auricular region, so vibration-related reflections from different positions in S_{bin} tend to show similar phase variations and preserve phase continuity even when some multipath reflections disappear. In contrast, micromotion varies across reflection positions and exhibits distinct phase variation patterns. Due to multipath-induced spatial spreading, dynamic micromotion reflections from different positions may intermittently disperse across adjacent bins or enter the current bin from neighboring bins, introducing phase variations distinct from the main direct reflection. These reflections disrupt intra-bin phase continuity, producing leakage-like spectral artifacts. This suggests that the interference source is not multipath alone, but multipath-coupled micromotion.

Understanding Multipath-Coupled Micromotion. To disentangle the effects of multipath, micromotion, and their coupling in near-field sensing, we use a two-factor controlled design with four conditions: ***(I)*** normal handheld use, with both present; ***(II)*** multipath only, with micromotion suppressed; ***(III)*** micromotion only, with multipath suppressed; and ***(IV)*** both suppressed. Participants are synchronously recorded with audio and mmWave signals at a face-to-device distance within 3 cm (Fig. 4(a)). To mitigate micromotion, we fix the device and ask participants to remain still during closed-mouth phonation, preserving vibration while minimizing motion. We suppress multipath by covering the device surface and surroundings with absorbing foam (Fig. 4(b)). To quantify the quality of mmWave-captured bone-conducted vibrations,

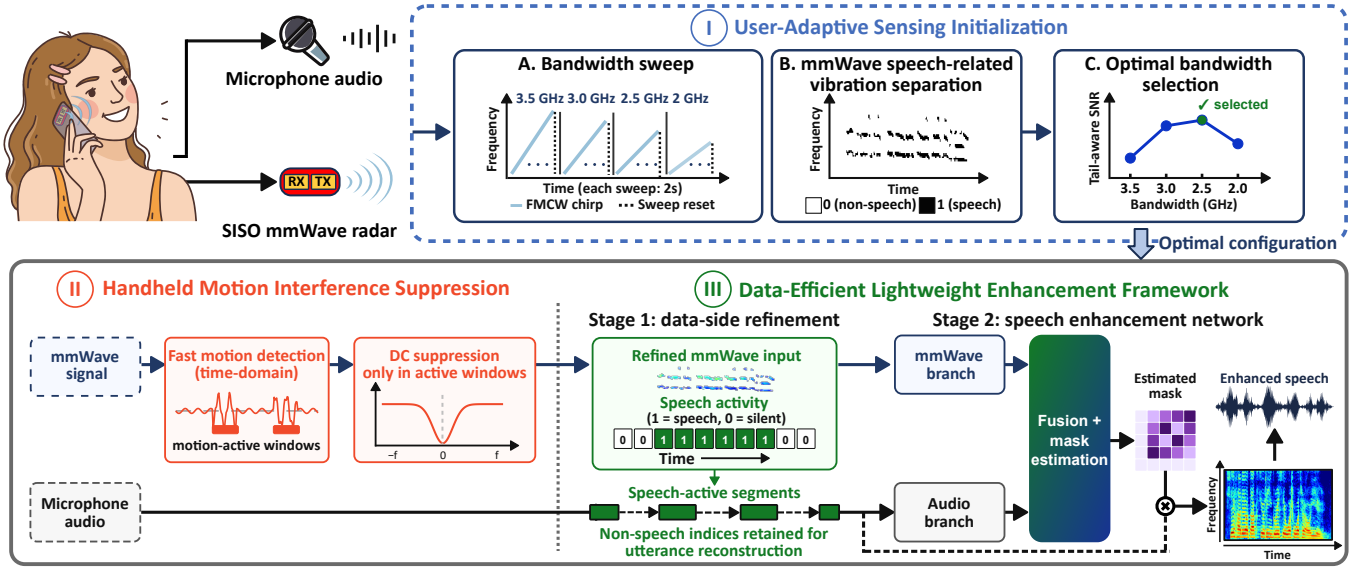


Figure 7: System overview of *BoneWave*. The system uses microphone audio and mmWave sensing to support near-ear phone-call enhancement through three modules: user-adaptive sensing initialization, handheld motion interference suppression, and data-efficient lightweight speech enhancement.

we apply cell-averaging constant false alarm rate (CA-CFAR) [36] to the speech spectrogram to identify speech-dominant regions and generate a frequency mask, as detailed in Appendix A. Since bone-conducted vibrations concentrate at low frequencies, we apply the mask below 0.5 kHz to the mmWave spectrogram to extract speech-related components. We then compute an average-power SNR as the ratio between the average power inside and outside the masked regions. The spectrograms in Fig. 5(a) show severe spectral degradation only when multipath and micromotion coexist. Fig. 5(b) shows that, when both are present, the SNR drops sharply to -3.27 dB, whereas suppressing either factor restores it to around 10 dB, and suppressing both further improves it to 11.52 dB. We also examine intra-bin reflection and phase stability through energy fluctuations and second-order phase differences. In condition ***(I)***, their standard deviations are 0.08 dB and 0.01 rad, respectively, 5.8-8.4× and 7.0-8.0× those in conditions ***(II)-(IV)***, revealing substantial reflection dispersion and phase instability, consistent with our theoretical analysis. Together, these results show that the artifacts primarily stem from multipath-coupled micromotion: *in the near field, micromotion reflections from different positions are spatially spread by multipath and intermittently dispersed across adjacent bins, breaking intra-bin phase continuity and introducing spectral artifacts.*

2.2 Sensing Configuration for Close-Range Phone Calling

We next analyze how to suppress these spectral artifacts. Since multipath is unavoidable in near-field use, the key is to preserve the temporal continuity of multipath-coupled micromotion reflections within each range bin. A natural mitigation is to increase each bin's spatial coverage, making dispersed reflections more likely to remain within the same bin and thus improving phase continuity. Yet, this introduces a trade-off: overly large bins may also admit off-target reflections outside the peri-auricular region, including interference (e.g., mouth motion), thereby degrading vibration fidelity. The goal, therefore, is to identify suitable bin coverage that suppresses artifacts while preserving bone-conducted vibration fidelity. In FMCW radar, this coverage is determined by the range resolution, $d_{\text{res}} = \frac{c}{2B}$, where B is the radar bandwidth. We experimentally study this trade-off by analyzing how bandwidth configurations affect mmWave bone-conduction sensing during near-ear use.

Evaluating Bandwidth Configurations. Under handheld near-ear use (Fig. 4(a)), we evaluate radar bandwidths of 4, 3.5, 3, 2.5, 2, 1, and 0.5 GHz, corresponding to per-bin spatial coverages of 3.75, 4.29, 5.00, 6.00, 7.50, 15.00, and 30.00 cm, respectively. We then analyze the average-power SNR of the resulting mmWave-captured vibration signals. Fig. 6 shows a consistent non-monotonic trend: as bandwidth decreases from 4 GHz, the SNR first rises, peaks at intermediate bandwidths, and then declines at lower bandwidths. Large bandwidths produce overly fine bins that cannot effectively consolidate multipath-coupled micromotion reflections, while small bandwidths produce overly coarse bins that admit off-target interference. Notably, the optimum varies slightly across users. User 1, with a larger build (185 cm tall), achieves the best result at 2.5 GHz (11.98 dB), whereas User 2, with a smaller build (approximately 150 cm tall), performs best at 3 GHz (9.65 dB). These two users are selected from 54 volunteers to represent clear body-size differences observed in our study (Sec. 4.1). This slight variation is due to differences in users' head geometry, which shift the suitable bin coverage. This motivates user-adaptive configuration for high-quality mmWave bone-conducted vibration sensing.

3 *BONEWAVE* DESIGN

Fig. 7 shows the architecture of *BoneWave*, a flexible phone-call enhancement system via contactless mmWave bone-conduction sensing, which consists of three modules:

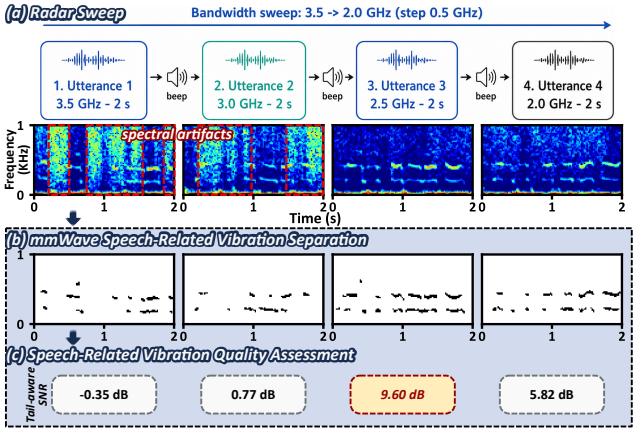


Figure 8: User-adaptive initialization for optimal radar configuration selection.

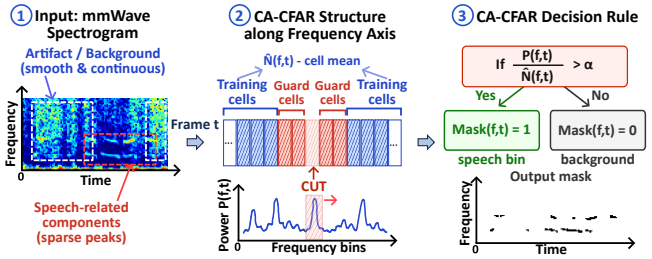


Figure 9: CA-CFAR workflow for mask estimation.

- **User-Adaptive Sensing Initialization.** This module adaptively selects user-specific radar configurations to appropriately consolidate reflected signals, thereby suppressing near-field artifacts and reducing off-target reflections.
- **Motion Interference Suppression.** This module localizes motion periods from phase fluctuations to avoid unnecessary static-period processing, while suppressing motion distortion to refine bone-conducted vibration signals.
- **Efficient Speech Enhancement Framework.** This module improves efficiency through data-side refinement and a lightweight transformer-based model, enabling efficient fusion of detail-rich audio and noise-robust mmWave cues.

3.1 User-Adaptive Sensing Initialization

We use a single-input single-output (SISO) mmWave radar to extract bone-conducted vibrations from phase differences in the energy-dominant range bin that carries rich user information. In near-ear handheld phone calls, mmWave bone-conducted vibration sensing suffers from spectral artifacts (Sec. 2.1). To address this, we need to identify a suitable radar bandwidth that provides proper range-bin coverage to consolidate reflections, preserve temporal continuity, and suppress the resulting artifacts. Although differences in users' head geometry lead to variations in the optimal bandwidth, our extensive user experiments show that the optima mainly fall within an intermediate range, i.e., 2-3.5 GHz (4.29-7.50 cm per bin). We thus design a user-adaptive sensing initialization procedure that sweeps this range to identify a suitable bandwidth before normal use. Using only radar, users can complete initialization without concern for airborne acoustic noise.

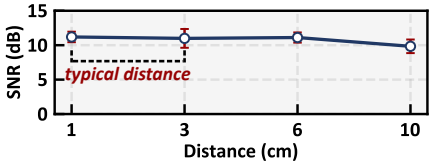


Figure 10: Performance under different distances.

3.1.1 mmWave Speech-Related Vibration Separation. During initialization, the user repeats the utterance four times while holding the device naturally near the ear (1-3 cm), as the radar sweeps the bandwidth from 3.5 to 2 GHz in 0.5 GHz steps and records 2 s of bone-conducted vibration signals at each setting (Fig. 8(a)). Under unsuitable bandwidths, spectral artifacts arise, raising the background energy and masking bone-conducted vibrations. To evaluate the quality of captured vibrations, we need to separate speech-related vibrations from the background and measure their relative energy prominence. We observe that artifacts form column-like patterns continuous along the frequency axis, while speech-related vibrations appear as sparse local peaks. CA-CFAR [36] is well suited for this separation: speech-dominant frequencies show strong local prominence relative to neighboring bins, while the more continuous energy of artifacts is less likely to trigger detection. We thus apply CA-CFAR along the frequency axis of the mmWave spectrogram to separate speech-active regions from background regions.

As illustrated in Fig. 9, CA-CFAR excludes nearby guard cells to avoid target leakage and uses the surrounding training cells to estimate the local background. Specifically, we compute the local background estimate as $\hat{N}(f,t) = \frac{1}{|\mathcal{T}(f)|}\sum_{u\in\mathcal{T}(f)} P(u,t)$, where $P(u,t)$ is the spectral energy at frequency bin u and time frame t, and $\mathcal{T}(f)$ is the training-cell set around f. The cell under test is then compared against the thresholded background estimate to determine whether it is speech-dominant. The resulting binary speech mask is:

$$M(f,t) = \begin{cases} 1, & \dfrac{P(f,t)}{\hat{N}(f,t)} > \alpha, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where α is the threshold coefficient, and $M(f,t) = 1$ indicates a speech bin. We set the CA-CFAR parameters to match the spectral structure of vocal vibration: *vocal-peak width determines the guard cells, while harmonic spacing determines the training span.* Guard cells exclude the tested vocal peak from the background estimate. Since the apparent width of a vocal peak under STFT is mainly determined by analysis-window mainlobe broadening [9], we set the guard-cell width according to the window mainlobe width in frequency bins, ensuring that the guard region covers the peak's main energy spread. Training cells estimate the local background without including neighboring vocal harmonics. Since adjacent harmonics are separated by one fundamental-frequency interval, and the human fundamental frequency is typically above 90 Hz [48], we constrain the training span within this spacing to avoid treating adjacent harmonics as background. The resulting mask effectively separates speech-related vibration regions from the background, as shown in Fig. 8(b). Detailed derivations are provided in Appendix A.

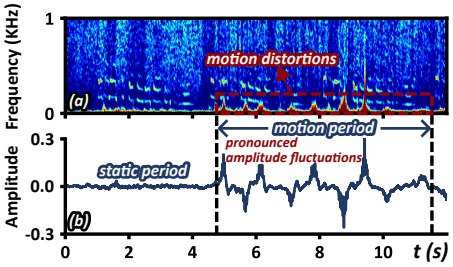


Figure 11: (a) Motion-induced distortion; (b) phase differences showing stable static periods and motion-induced fluctuations.

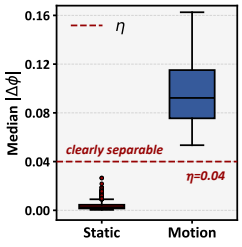


Figure 12: Distribution of median phase variation across static and motion periods.

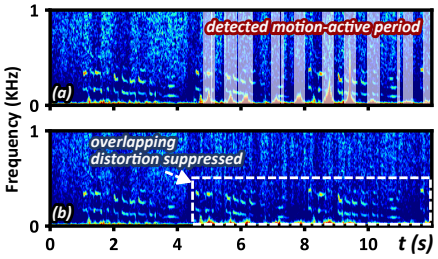


Figure 13: (a) Detected motion-active periods; (b) motion distortion suppressed by motion-gated DC removal.

3.1.2 Speech-Related Vibration Quality Assessment. After obtaining the mask that separates speech-related vibrations from the background, we assess signal quality by measuring the prominence of these vibrations over the surrounding background. However, artifacts in background regions occur irregularly over time and can be brief but high-intensity, making them more harmful than the low-level radar noise floor. To capture this degradation, we design a tail-aware SNR metric that uses upper-tail background statistics to emphasize high-energy artifacts during noise estimation.

Given an mmWave power spectrogram $P(f,t)$ and the corresponding speech-related mask $M(f,t)$, we restrict the analysis to frequencies above 50 Hz to exclude dominant body reflections concentrated in the low-frequency band. We obtain the speech and background regions as $\Omega_s = \{(f,t) \mid M(f,t) = 1\}$ and $\Omega_n = \{(f,t) \mid M(f,t) = 0\}$, respectively. The average speech-related vibration power is $S = \frac{1}{|\Omega_s|}\sum_{(f,t)\in\Omega_s} P(f,t)$. Over the background region, we form the non-speech energy set $\mathcal{E} = \{P(f,t) \mid (f,t) \in \Omega_n\}$. Rather than using its global mean, which can be dominated by low-energy background bins, we emphasize strong artifacts by computing the upper-quartile threshold $\tau = \text{Quantile}_{0.75}(\mathcal{E})$. We then compute the tail noise power as:

$$N^{\text{tail}} = \frac{1}{|\mathcal{E}^{\text{tail}}|}\sum_{x\in\mathcal{E}^{\text{tail}}} x, \quad \mathcal{E}^{\text{tail}} = \{x \in \mathcal{E} \mid x \geq \tau\}. \tag{3}$$

Accordingly, the tail-aware SNR is $Q = 10\log_{10}\frac{S}{N^{\text{tail}}}$. As illustrated in Fig. 8(c), we score all candidate bandwidths using this metric and select the one with the highest score.

Initialization Gain. We evaluate the effectiveness of user-adaptive initialization across users with different body builds (150-185 cm tall) who naturally hold the device 1-3 cm from the ear. Across all 54 participants, the selected radar bandwidths predominantly cluster at 2.5 or 3 GHz within the 2-3.5 GHz sweep range, indicating a similar adaptation pattern despite user-specific variation. Compared with the default 4 GHz configuration, our initialization substantially improves the quality of mmWave-captured bone-conducted vibrations, raising the average-power SNR from 0.96 to 10.34 dB.

Operating Distance. After a single initialization, we evaluate mmWave bone-conducted vibration sensing across different phone-to-ear distances using the selected radar configuration. Fig. 10 shows that *BoneWave* maintains high average-power SNRs at typical handheld phone-call distances of 1-3 cm, achieving around 11 dB on average. The performance remains stable at 6 cm and decreases slightly to 9.84 dB at 10 cm due to increased interference from off-target reflections. These results suggest that the sensing design is well suited to normal handheld phone-call distances and remains effective at larger-than-typical distances.

3.2 Motion Interference Suppression

Compared with micromotion during vocalization, such as facial movements and subtle tremors, handheld use introduces larger incidental hand or head motions. These small but fast motions can introduce distortion overlapping with bone-conducted vibrations (Fig. 11(a)). Prior work [47] attributes this distortion to coupling between body motion and the radar's zero-frequency DC component. However, directly filtering DC in the bin's IQ signal can disrupt signal continuity during static periods, where DC overlaps with static body reflections, causing spectral leakage (Appendix B). This motivates selective DC decoupling only during motion periods. Prior work [47] detects motion from Doppler velocity in the frequency-domain IQ signal, introducing non-negligible time-frequency transformation overhead over long static segments. In contrast, we observe that motion can be localized directly from time-domain phase differences, avoiding this overhead: *their amplitudes reflect displacement magnitude, producing pronounced fluctuations during motion but remaining stable during static periods (Fig. 11(b)).* Based on this, we design an efficient motion distortion suppression method that detects motion periods in the time domain and confines DC removal to them, refining bone-conducted vibrations.

Motion-active window localization. We compute the phase difference $\Delta\phi[n]$ between adjacent IQ samples. Since incidental motion induces larger phase variations than static periods, we define the motion score of each sliding window as the median amplitude of phase increments: $m = \text{median}(\{|\Delta\phi[n]|\}_{n=s}^{s+L-1})$, where L is the window length and s is the start index. This median-based score is robust to isolated phase spikes while capturing the overall motion level. As shown in Fig. 12, motion periods consistently exhibit higher scores than static periods. We thus set a threshold η slightly above the upper bound of the static-period distribution and label each window as motion-active if $m > \eta$, triggering DC suppression only for these windows. Fig. 13(a) shows that this effectively localizes motion-active periods.

Motion-gated DC suppression. For each motion-active window, we compute the spectrum of its windowed IQ samples and follow the smooth band-stop filtering strategy in [47] to remove the DC component. We then transform the processed motion-active windows back to the IQ time domain, merge them with static-period

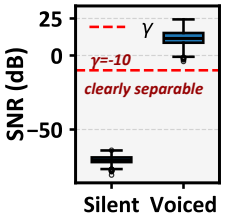


Figure 14: Distribution of silent and voiced frames.

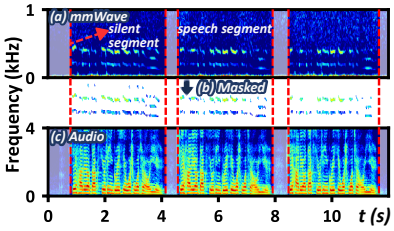


Figure 15: (a) mmWave spectrogram, (b) masked result and (c) aligned audio spectrogram.

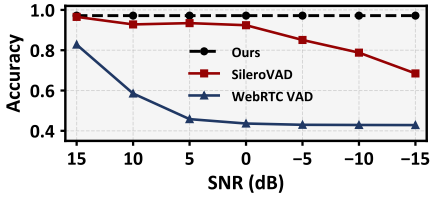


Figure 16: Speech-segment detection accuracy across SNRs.

signals, and compute the phase-difference signal that represents bone-conducted vibrations. Fig. 13(b) shows that the resulting signal has reduced overlapping distortion while preserving vibrations. Compared with frequency-domain Doppler-guided DC removal, our method avoids transformation overhead over long static segments, yielding about a 1.6× speedup during handheld activities.

3.3 Speech Enhancement Framework

For efficient speech enhancement on resource-constrained phones, *BoneWave* combines data refinement with lightweight cross-modal enhancement. On the data side, radar noise floors in mmWave signals burden feature extraction, while long non-speech periods introduce redundant computation. We isolate bone-conducted vibrations as a cleaner mmWave input and use them to gate subsequent processing to speech-active segments. On the model side, simple concatenation cannot effectively model cross-modal relationships under a tight size. We thus design a lightweight transformer-based model that leverages the selective interaction capability of cross-attention [42, 49] to enable efficient complementary fusion. Inspired by time-frequency (T-F) masking in speech enhancement [1, 8, 45], *BoneWave* estimates a T-F mask to suppress noise and uses a joint T-F loss to improve waveform fidelity and spectral accuracy.

3.3.1 Data-Side Refinement. We upsample the mmWave phase-difference signal to 16 kHz for alignment with the microphone audio, and convert both signals into the T-F domain. Based on Sec. 3.1.1, speech-related mmWave vibrations appear as local energy prominences along the frequency axis. We therefore apply CA-CFAR to the mmWave spectrogram within 0-1 kHz to obtain a speech-related binary mask. Active bins indicate the vibrations and are used to filter the original mmWave spectrogram, yielding a cleaner representation (Fig. 15(b)) that reduces the subsequent feature-extraction burden. Since bone-conducted vibrations are synchronized with vocal activity, we use them as noise-robust cues to identify speech-active periods. For each frame t, we compute an average-power SNR score r_t as the ratio between the average power

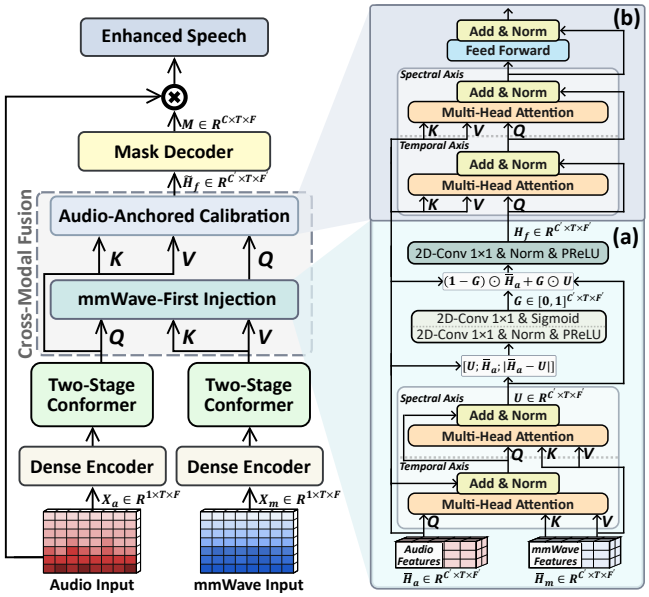


Figure 17: Network structure. The fusion module adopts (a) an mmWave-first injection and (b) an audio-anchored calibration strategy via cross-attention.

inside the active bins and that in the remaining background bins. As shown in Fig. 14, voiced and silent frames exhibit clearly separated SNR levels. We therefore set a threshold γ to label frame t as speech-active if $r_t > \gamma$, and as non-speech otherwise. Consecutive speech-active frames are merged into speech segments:

$$\mathcal{P} = \{(t_s, t_e) \mid r_t > \gamma,\ \forall t \in [t_s, t_e]\}, \tag{4}$$

where t_s and t_e denote the segment boundaries. We further dilate detected speech segments to preserve weakly vibrating consonants and merge short silent gaps to avoid over-segmentation. The boundaries are applied to the corresponding speech spectrogram (Fig. 15(c)), while non-speech frame indices are retained for complete utterance reconstruction.

Detection Performance. We compare mmWave-based speech-segment extraction with two audio-based baselines, SileroVAD [37] and WebRTC VAD [5] (Fig. 16). Under relatively clean conditions, i.e., at 15 dB SNR, our method and SileroVAD both achieve an accuracy of 0.97, while WebRTC VAD reaches 0.83. As noise increases, our method remains stable at 0.97 due to the noise-resilient mmWave sensing, whereas SileroVAD and WebRTC VAD drop substantially to 0.79 and 0.43 at −10 dB. This confirms the robustness of mmWave-based detection under noise, with details in Appendix C.

3.3.2 Lightweight Network. As shown in Fig. 17, we design a transformer-based model for cross-modal fusion, followed by T-F mask prediction for noise suppression. Since the output lies in the audio space, direct audio-mmWave fusion may introduce cross-modal misalignment. We therefore first use cross-attention to selectively inject noise-robust mmWave cues into the audio representation, and then use audio-anchored attention to recalibrate the fused representation toward the audio-space enhancement target. A joint T-F loss further preserves waveform fidelity and spectral details.

The audio and mmWave segments, denoted as $X_a, X_m \in \mathbb{R}^{C\times T\times F}$, are fed into two modality-specific branches with the same architecture, where $C = 1$; T and F denote the temporal and frequency dimensions, respectively. Motivated by the effectiveness of Dense Encoder [23] and Conformer [7, 23] in speech modeling, each branch uses a Dense Encoder followed by two two-stage Conformer Blocks to extract audio and mmWave features H_a and H_m for cross-modal fusion, respectively, with architectures provided in Appendix D.

Cross-Modal Fusion. We fuse audio and mmWave features using a cross-attention-based strategy that selectively integrates complementary cues from the two modalities (Fig. 17(a)). Both features are first projected into a shared latent space by separate 1×1 2D convolutions. Modality-specific positional encodings are then generated by 2D convolutions and added to the projected features, yielding $\bar{H}_a, \bar{H}_m \in \mathbb{R}^{C'\times T\times F'}$. We apply cross-attention in two stages, first along the temporal axis and then along the spectral axis, with the audio feature as the query (Q) and the mmWave feature as the key (K) and value (V). This selectively retrieves noise-robust mmWave cues and progressively injects them into the audio representation via residual addition, producing $U \in \mathbb{R}^{C'\times T\times F'}$. We further balance the fusion by estimating weights $G \in [0,1]^{C'\times T\times F'}$ from $[U; \bar{H}_a; |\bar{H}_a - U|]$ using a sigmoid-gated 2D convolutional projection, yielding the fused feature:

$$H_f = P_o\left((1-G)\odot \bar{H}_a + G\odot U\right), \quad H_f \in \mathbb{R}^{C'\times T\times F'}, \tag{5}$$

where $\odot$ denotes element-wise multiplication, and $P_o(\cdot)$ is a 1×1 convolutional projection. We further use the pre-fusion audio feature $\bar{H}_a$ as an anchor to refine the fused feature H_f, suppressing modality-induced bias and aligning it with the audio-space enhancement target (Fig. 17(b)). We first perform temporal cross-attention with H_f as Q and $\bar{H}_a$ as K and V, using residual addition to preserve fused information while injecting audio-space structural cues. We then use the temporally refined feature as Q and apply the same audio-anchored refinement along the spectral axis. Finally, a feed-forward module produces the refined fused feature $\tilde{H}_f$.

The refined feature $\tilde{H}_f$ is decoded into a T-F mask $M \in \mathbb{R}^{C\times T\times F}$ using a 2D transposed convolution, normalization, PReLU activation, and a final 1×1 convolutional projection. The mask is then applied to the noisy audio spectrogram, followed by inverse STFT to reconstruct the enhanced waveform. *Overall, BoneWave is lightweight, with only 1.7 M parameters.*

Loss Function. Enhanced speech quality depends on both waveform fidelity and spectral detail, which jointly influence intelligibility and perceptual quality. We therefore optimize *BoneWave* in both the time and frequency domains. In the time domain, we use the scale-invariant signal-to-distortion ratio (SI-SDR) to measure signal fidelity independent of amplitude scaling:

$$\text{SI-SDR} = 10\log_{10}\frac{|S_{\text{target}}|^2}{|S_{\text{error}}|^2}, \quad \begin{cases} S_{\text{error}} = \hat{S} - S_{\text{target}}, \\ S_{\text{target}} = \frac{\hat{S}\cdot S}{|S|^2}S, \end{cases} \tag{6}$$

where $\hat{S}$ and S denote the estimated speech and ground-truth speech, respectively. Since SI-SDR is a metric to be maximized, we minimize its negative value, i.e., $L_{\text{SI-SDR}} = -\text{SI-SDR}$. In the frequency domain, we use an L1 loss on the magnitude spectrogram to improve spectral accuracy and recover speech details:



Figure 18: Experimental setup. (a) Phone microphones and mmWave radars used in our evaluation; (b) Near-ear phone-call setting; (c) Handheld device placements.

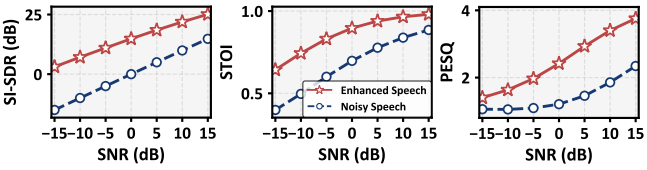


Figure 19: Performance across noise levels.

$L_{\text{spec}} = \left\| |\text{Spec}(\hat{S})| - |\text{Spec}(S)| \right\|_1$. The joint loss is:

$$L = L_{\text{SI-SDR}} + \lambda L_{\text{spec}}, \tag{7}$$

where λ balances the loss scales and is set to 0.1.

4 PERFORMANCE EVALUATION

4.1 Experimental Setup

Hardware. *BoneWave* is built on a phone microphone and an mmWave radar. Specifically, we use three types of phone microphones (Huawei P30, Xiaomi 14, and iPhone 15), together with two types of radars, TI IWR6843ISK [41] and Infineon BGT60TR13C [16], as shown in Fig. 18. The radar operates in a SISO configuration with one transmit and one receive antenna, sampling at 2.5 kHz to capture bone-conducted vibrations, as detailed in Table 1.

Software. Leveraging the mmWave recording tool [50], we develop Python scripts for radar bandwidth sweeps during sensing initialization (Sec. 3.1) and bone-conducted vibration capture. We train *BoneWave* on a server with NVIDIA GeForce RTX 4090 GPUs using PyTorch 2.1.1. The model uses 2-s input segments with a batch size of 6 and the Adam optimizer at a learning rate of 1×10^{-4}.

Dataset. We recruited 54 volunteers (27 males and 27 females) to record paired speech and mmWave data under natural near-ear handheld use (Fig. 18). Participants were aged 19-29 years, approximately 150-185 cm tall, with BMIs predominantly between 17 and 28, except one at 32.7. All studies were approved by the Institutional Review Board (IRB) of our institution. Each participant randomly selected three sentences from the TIMIT corpus [3] and read each sentence 10 times, yielding approximately 4 hours of recordings. We use the TUT dataset [15] for noise mixing, which contains 10 common noisy scenes (e.g., shopping malls, public squares, streets,

Table 1: mmWave radar configuration.

Frame Periodicity	0.4 ms	**Idle Time**	40 μs
Chirps/Frame	1	**Ramp End Time**	60 μs

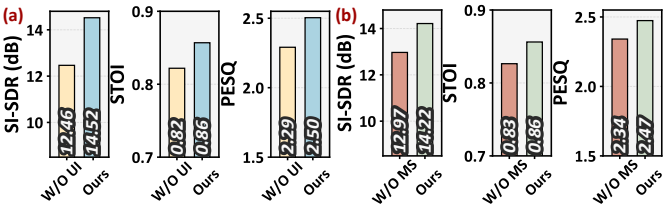


Figure 20: Ablation of (a) user initialization and (b) motion distortion suppression.

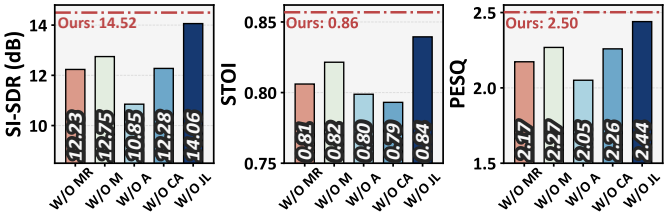


Figure 21: Ablation of speech enhancement framework components.

buses, and subways) covering common daily scenarios. Each participant's speech is uniformly mixed with different noise types to generate noisy speech with SNRs from −15 dB to 15 dB. We use a strict test protocol: 80% of participants are randomly assigned to the training set and the remaining 20% (12 participants) to the test set, with no overlap in participants, speech content, or noise.

Metrics. We evaluate *BoneWave* using three speech-quality metrics, all with higher scores indicating better quality.

- *Scale-Invariant Signal-to-Distortion Ratio (SI-SDR)* [19] measures the distortion of enhanced speech relative to clean speech, reflecting similarity to the clean reference.
- *Short-Time Objective Intelligibility Metric (STOI)* [40] evaluates the intelligibility of speech, ranging from 0 to 1.
- *Perceptual Evaluation of Speech Quality (PESQ)* [35] evaluates the quality of speech, ranging from -0.5 to 4.5.

4.2 Overall Performance

As described in Sec. 4.1, we evaluate *BoneWave* on 12 unseen users with noisy inputs spanning -15 to 15 dB SNR. Table 2 shows that *BoneWave* significantly improves noisy speech in signal fidelity, intelligibility, and perceptual quality, increasing SI-SDR by 14.57 dB on average and achieving STOI and PESQ scores of 0.86 and 2.50, corresponding to relative gains of 28% and 74%, respectively.

A breakdown by noise level in Fig. 19 shows that *BoneWave* significantly improves SI-SDR, STOI, and PESQ across all SNRs, with especially large gains under severe noise. At −15 dB SNR, it improves SI-SDR by 18.10 dB over the noisy input.

Performance Comparison. We compare *BoneWave* with three state-of-the-art audio-only speech enhancement models, MP-SENet [23], PHASEN [51], and PrimeK-Net [20], all leveraging T-F representations. We further reproduce two mmWave-based multimodal models, RadioSES [30] and mmMUSE [45], which enhance speech by jointly fusing mmWave and audio signals. Notably, their original sensing configurations do not yield usable bone-conducted vibrations at near-ear distances, causing the systems to fail in this regime. We thus apply the same user-adaptive initialization to both baselines for a fair model-level comparison, isolating performance from sensing differences. All baselines are trained and evaluated on our dataset under the same settings as *BoneWave*. As shown in Table 2, *BoneWave* outperforms all baselines, with average gains of 2.24 dB in SI-SDR, 0.03 in STOI, and 0.21 in PESQ. Compared with audio-only models, it improves SI-SDR by 2.74 dB, demonstrating the benefit of noise-robust mmWave cues for complementary speech enhancement. Compared with RadioSES and mmMUSE, it achieves an average SI-SDR gain of 1.51 dB, showing the effectiveness of our transformer-based fusion. Unlike these multimodal baselines, which optimize enhanced speech only with time-domain losses, *BoneWave* uses a joint T-F loss to better preserve speech details, improving PESQ by 0.25 (11%) on average.

Table 2: Performance comparison.

	SI-SDR (dB)	STOI	PESQ
Noisy speech	-0.05	0.67	1.44
PHASEN [51]	12.82	0.83	2.34
MP-SENet [23]	11.83	0.82	2.34
PrimeK-Net [20]	10.70	0.83	2.28
RadioSES [30]	12.56	0.83	2.20
mmMUSE [45]	13.47	0.83	2.31
BoneWave	**14.52**	**0.86**	**2.50**

4.3 Ablation Study

In this section, we conduct an ablation study to quantify the effectiveness of each design in *BoneWave*.

Ablation of User-Adaptive Sensing Initialization. We ablate the user-adaptive initialization in *BoneWave*. *W/O UI* disables this module and uses the default 4 GHz radar bandwidth. Without this initialization, close-range sensing introduces severe spectral artifacts that overwhelm bone-conducted vibrations. Fig. 20(a) shows that *W/O UI* degrades all metrics to 12.46 dB SI-SDR, 0.82 STOI, and 2.29 PESQ. This confirms the importance of user-adaptive initialization for reliable near-ear mmWave sensing.

Ablation of Motion Distortion Suppression. We evaluate the motion distortion suppression module under incidental handheld motion, including walking, stair climbing, and jogging. Fig. 20(b) shows that removing this module (*W/O MS*) substantially reduces SI-SDR, STOI, and PESQ to 12.97 dB, 0.83, and 2.34, respectively. Motion-induced distortion corrupts mmWave-captured bone-conducted vibrations and weakens their complementarity with audio, making *W/O MS* perform similarly to audio-only methods.

Ablation of Speech Enhancement Framework. We ablate *BoneWave*'s data-side refinement and network design.

- ***W/O MR*** removes mmWave refinement and uses the original mmWave spectrogram without bone-conducted vibration separation in both training and testing.
- ***W/O M*** removes the mmWave input and its subnetwork.
- ***W/O A*** removes the audio input and its subnetwork.
- ***W/O CA*** replaces cross-attention fusion modules with ResNet-style residual blocks [13] and fuses concatenated cross-modal features through residual processing, with a comparable size.

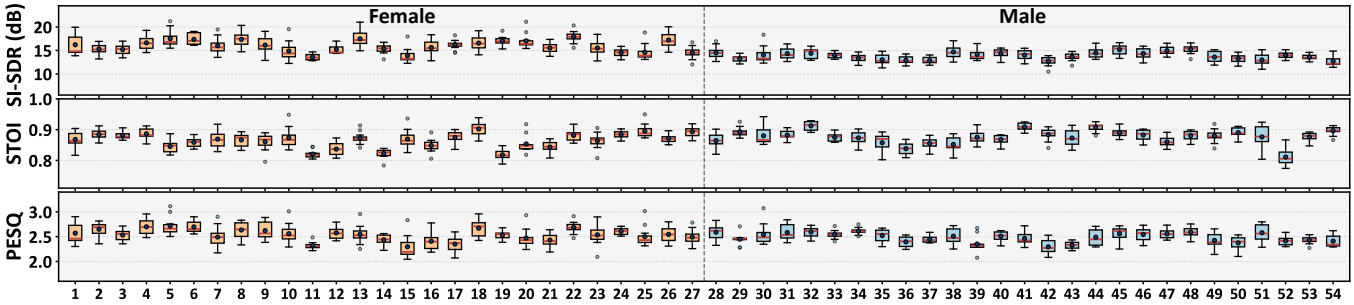


Figure 22: Performance across unseen users.

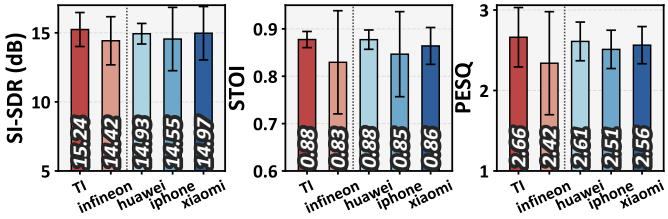


Figure 23: Performance under device variations.

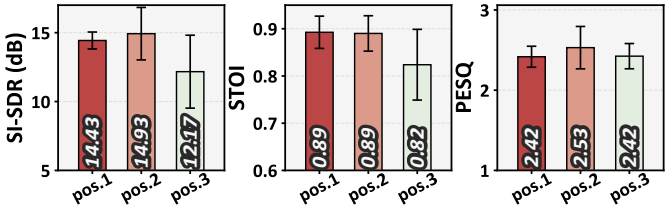


Figure 24: Performance under varied device placements.

- ***W/O JL*** removes the spectral loss from the joint objective, using only the time-domain SI-SDR loss.

Fig. 21 shows the ablation results of *BoneWave*. Since radar noise floors burden mmWave feature extraction, removing mmWave refinement in *W/O MR* degrades all metrics, reducing SI-SDR, STOI, and PESQ by 2.29 dB, 6%, and 13%, respectively. This confirms that mmWave refinement provides cleaner input for effective lightweight feature extraction. Removing the mmWave input in *W/O M* causes a clear performance drop, reducing the average SI-SDR to 12.75 dB, a decrease of 1.77 dB. Together with the audio-only baselines, this verifies the effectiveness of noise-robust bone-conducted vibration cues from mmWave. In contrast, *W/O A* cannot fully recover speech, as mmWave alone captures incomplete speech harmonics, leading to degradation across all metrics. Replacing cross-attention in *W/O CA* reduces SI-SDR, STOI, and PESQ by 2.24 dB, 8%, and 10%, respectively, demonstrating the effectiveness of our fusion module while keeping the model lightweight. Finally, *W/O JL* degrades performance, indicating that preserving spectral details beyond signal fidelity is important for speech quality.

4.4 Robustness Evaluation

We evaluate *BoneWave*'s robustness across unseen users and daily-use conditions, including device variations, placement changes, handheld motion during daily activities, and common sensing obstructions. Specifically, we test it with two commercial mmWave radars (TI IWR6843ISK and Infineon BGT60TR13C) and three smartphones (Huawei P30, Xiaomi 14, and iPhone 15), as shown in Fig. 18(a). We also consider three handheld placements (Fig. 18(c)) and three handheld-use motion scenarios: walking, stair climbing, and jogging. Finally, we evaluate daily-use sensing obstructions, including sweat, clothing occlusion by a wool hat, and hair occlusion. Since our radar initialization depends on both the user and device, which jointly shape near-field reflections, we rerun it when switching radars or smartphones. For all other robustness conditions, we reuse a single initialized configuration without recalibration.

Unseen Users. We conduct user-grouped five-fold cross-validation on our 54-user dataset, with no overlap in participants, speech content, or noise between training and test sets. Specifically, we place users with overlapping speech content in the same fold, yielding five mutually exclusive folds with similar user counts. Although noise categories may recur, TUT [15] provides substantial within-category variability across locations and time periods. As shown in Fig. 22, *BoneWave* generalizes well to unseen users, achieving average SI-SDR, STOI, and PESQ of 14.90 ± 1.44 dB, 0.87 ± 0.02, and 2.51 ± 0.11, respectively, with comparable performance across gender groups. Notably, we observe attenuated bone-conducted signals in users with obesity, consistent with prior findings on soft-tissue attenuation [25]. For example, an SI-SDR of 12.77 dB is observed for an unseen user with a BMI of 32.7, 14% below the overall average.

Device Variation. Fig. 23 shows that *BoneWave* is robust across sensing and recording devices. With the two radars, it achieves SI-SDRs of 15.24 dB and 14.42 dB, respectively. For the BGT60TR13C, STOI and PESQ are slightly lower, at 0.83 and 2.42, mainly because it shares a single cable for power supply and data transmission, which may reduce signal stability. Across smartphones, *BoneWave* achieves 14.55-14.97 dB SI-SDR with stable STOI and PESQ. These results show that *BoneWave* is robust to device variations.

Device Placement. As shown in Fig. 24, *BoneWave* achieves strong performance at Positions 1 and 2, with SI-SDRs of 14.43 dB and 14.93 dB, respectively. Performance drops at Position 3, resulting in 12.17 dB SI-SDR, 0.82 STOI, and 2.42 PESQ. This is mainly because it spans the temple, upper-ear, and posterior-lateral regions, where bone-conducted vibrations are weak [14]. Consistently, we observe little discernible vibration in the mmWave signal. Nevertheless, our survey shows that this placement is rarely used in handheld calls.

Daily Handheld Motion. Fig. 25 shows that *BoneWave* remains robust under daily handheld motion. Across three common activities, its performance is comparable to the still setting, with SI-SDRs of 13.95-14.54 dB and relative variations of only 0.67% in STOI and

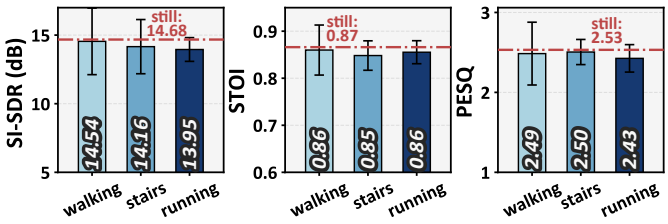


Figure 25: Performance under handheld motion.

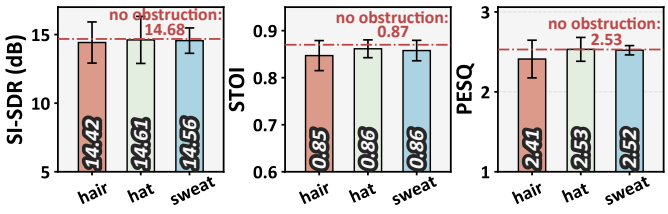


Figure 26: Performance under obstructions.

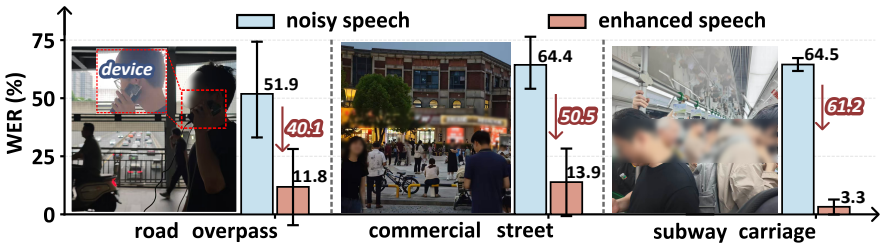


Figure 27: Evaluation of *BoneWave* in real-world scenarios.

Table 3: Run-time latency.

	Preproc.	Infer.	Total	RTF
Server	9.19 ± 0.18 ms	19.59 ± 0.50 ms	28.78 ± 0.58 ms	0.01
Laptop	6.61 ± 0.93 ms	569.13 ± 8.35 ms	575.75 ± 8.55 ms	0.29
Phone	3.12 ± 0.07 ms	1194.04 ± 10.73 ms	1197.16 ± 10.69 ms	0.60

1.53% in PESQ. Although performance slightly decreases during running, *BoneWave* still maintains strong enhancement quality.

Daily Sensing Obstruction. Fig. 26 shows that *BoneWave* remains robust under common daily sensing obstructions, including hair occlusion, a wool hat covering the ear, and sweat. It achieves SI-SDRs of 14.42-14.61 dB, while maintaining STOI and PESQ above 0.85 and 2.41, respectively, comparable to the unobstructed setting.

We discuss potential improvements for body- or placement-related signal attenuation in Sec. 6.2.

4.5 Real-World Evaluation

We further validate *BoneWave* under real-world conditions with diverse noise and natural handheld use. As shown in Fig. 27, we evaluate it in three scenarios: a road overpass, a commercial street, and a subway carriage. The road overpass features high-intensity noise from wind and intermittent traffic hum, with levels of around 73-78 dB. The other two scenarios contain mixed human voices and varying ambient noise at around 58-76 dB and 53-74 dB, respectively. The commercial street features more dynamic human speech, whereas the subway is dominated by mechanical noise varying mainly in intensity. The evaluation involves 10 unseen users (5 males and 5 females, aged 22-29 years and about 158-185 cm tall), each recording around 3-5 minutes of continuous, unrehearsed speech per scenario, comprising random TIMIT sentences and spontaneous utterances. We transcribe the speech using Whisper [33] and quantify enhancement effectiveness with word error rate (WER).

Fig. 27 shows that *BoneWave* effectively enhances noisy speech across all three scenarios, achieving an average WER of 9.6%. Specifically, the subway scenario yields the best post-enhancement WER (3.3±3.5%) with the smallest variation, whereas the commercial street (13.9±15.9%) and road overpass (11.8±17.9%) exhibit greater variability. Compared with noisy inputs, *BoneWave* reduces WER by 61.2, 50.5, and 40.1 percentage points on average in the subway, commercial street, and road overpass, respectively. The better performance in the subway may be attributed to its relatively stable environment, compared with the transient traffic noise at the road overpass and more active human speech in the commercial street. Despite considerable scene differences, *BoneWave* maintains effective enhancement for natural speech, supporting generalization beyond the controlled setting.

4.6 Computational Overhead

Model Parameters. *BoneWave* is lightweight, with only 1.7 M parameters. It is smaller than audio-only models such as MP-SENet (2.1 M) and PHASEN (7.7 M), comparable to PrimeK-Net (1.4 M), and also smaller than multimodal models such as RadioSES (2.1 M) and mmMUSE (2.1 M). This compact design benefits from complementary mmWave cues, reducing reliance on audio alone and highlighting the effectiveness of our fusion module. As modern smartphones continue to improve in computation and memory capacity [6], models at this scale are practical for mobile deployment.

Time Consumption. We evaluate *BoneWave*'s computational efficiency, including preprocessing (motion interference suppression and data-side refinement) and model inference, on three platforms: a server with an AMD EPYC 7763 CPU and an NVIDIA GeForce RTX 4090 GPU, a laptop with an AMD Ryzen AI 9 HX 370 CPU, and a smartphone with a Snapdragon 8 Gen 3 SoC. For smartphones, we export the model to ONNX and integrate it with preprocessing in an Android Studio app for offline on-device inference. The server runs preprocessing on the CPU and inference on the GPU, while the laptop and smartphone run both on the CPU. We run 100 iterations on synchronized 2-s mmWave-audio segments. Table 3 shows that *BoneWave* achieves real-time factors (RTFs) of 0.01, 0.29, and 0.60 on the three platforms, respectively, indicating real-time computational capability. To account for the impact of buffering latency on phone-call interactivity, we further benchmark it on the smartphone using finer-grained 500-, 100-, and 50-ms segments, with RTFs remaining stable and below 1. In particular, a 50-ms segment takes an average of 30.64 ms to process, corresponding to an estimated local block latency of ≈81 ms including buffering and computation. These results provide a computational basis for integrating *BoneWave* into practical streaming phone-call pipelines toward the sub-150-ms mouth-to-ear delay guideline in ITU-T G.114 [18], with limitations discussed in Sec. 6.2.

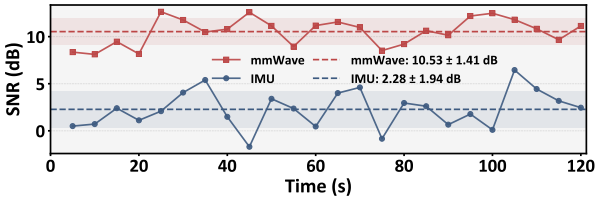


Figure 28: Comparison of bone-conducted signal quality captured by mmWave radar and a smartphone IMU.

5 RELATED WORK

5.1 Audio-only Speech Enhancement

Learning-based speech enhancement methods have substantially improved speech recovery by mapping noisy speech to clean speech in the time domain [31, 32] or the T-F domain [20, 23, 51]. However, relying solely on audio, their effectiveness is fundamentally limited in complex real-world environments, particularly under multi-speaker interference [39] or extremely low SNRs [1, 31].

5.2 Multi-Modal Speech Enhancement

With advances in multimodal fusion, modalities such as vision, ultrasound, and vocal-induced vibrations have increasingly been explored as noise-robust complements to audio.

5.2.1 Articulatory Gestures. Cameras and ultrasound can capture articulatory gestures, such as lip and tongue movements, to provide complementary cues for speech enhancement. However, vision-based methods [4, 38] face privacy and lighting issues, while ultrasound-based methods [1, 39] are vulnerable to environmental disturbances such as wind [46]. Both require near-frontal sensing for sufficient articulatory information, a geometry that conflicts with the near-ear holding posture of phone calls.

5.2.2 Vocal-Induced Vibrations. During near-ear handheld calls, bone-conducted vibrations concentrate around the ear, making them well suited for this setting. We systematically compare prior contact-based and mmWave-based methods for speech enhancement in Table 4 to assess *BoneWave*'s advantages in this scenario.

Contact-Based Sensing. Prior work captures such vibrations with contact-based IMUs [14, 46] or piezoelectric sensors [24]. VibVoice [14] targets head-mounted wearables, while AccCall [46] uses a smartphone's built-in IMU to sense bone-conducted vibrations for phone-call enhancement. AccCall captures vibrations along the IMU axis normal to the face, with robustness to low-frequency motion largely confined to this axis; yet, activity-induced orientation changes may redistribute vibrations across axes, degrading signal quality. We further compare the quality of synchronously captured bone-conducted signals from *BoneWave* and a smartphone's built-in IMU as in AccCall, under natural handheld use while walking (Appendix E). Fig. 28 shows that *BoneWave* maintains high-quality signals with a consistently higher average-power SNR (10.53±1.41 vs. 2.28±1.94 dB), whereas the IMU signal severely degrades and fluctuates markedly as phone-skin contact naturally varies. *BoneWave* does not depend on sustained stable contact, accommodating phone shifts and grip changes for flexible phone-call use. It is also beneficial under near-ear occlusions from hats or hairstyles, or when users prefer not to press the phone against sweaty skin. Our evaluations substantiate these practical benefits, with the placement and motion results further supporting natural handheld use (Sec. 4.4).

Table 4: Comparison of speech enhancement systems.

	Aux. Modality	Near-field Adapt.	Motion Robustness	Enh. Framework		Contact-less
				Data Refin.	Model Size	
VibVoice [14]	IMU	N/A	✓	✗	3.5M	✗
AccCall [46]	IMU	N/A	✓	✗	2.1M	✗
RadioSES [30]	mmWave	✗	✗	✗	2.1M	✓
mmMUSE [45]	mmWave	✗	✓	✓	2.1M	✓
BoneWave	mmWave	✓	✓	✓	1.7M	✓

✓: included; ✗: not included; N/A: not applicable.

mmWave Sensing. Compared with contact-based sensing, mmWave-based contactless sensing has gained attention in diverse applications [10–12, 26, 27, 43, 44, 47]. It enables flexible vibration sensing by capturing minute displacements in the environment. Although prior work has used mmWave sensing for speech recognition [21, 53] and enhancement [30, 45], these methods primarily rely on frontal vocal-cord vibration sensing beyond 10 cm. In contrast, phone calls require ear-side sensing at a much closer range of 1-3 cm under handheld use, introducing two challenges: near-field spectral artifacts and distortion from small but rapid incidental motions. We further characterize these artifacts and accordingly design a user-adaptive initialization, enabling effective near-field mmWave sensing of bone-conducted vibrations beyond prior sensing regimes. Targeting resource-constrained phones, *BoneWave* develops a pipeline with component-level efficiency optimizations over RadioSES [30] and mmMUSE [45] (Table 4). For motion interference, we follow the suppression strategy in prior work [47], recovering the signal by DC removal during motion-active periods. We detect these periods directly from time-domain phase variations, avoiding T-F transformations over long static segments required by Doppler-based motion detection in [47], yielding about a 1.6× speedup overall. In contrast, mmMUSE compensates for Doppler shifts in the T-F representation of radar IQ signals to reconstruct vocal spectrograms, incurring substantial overhead; *BoneWave* instead applies efficient DC removal only where needed, avoiding costly and redundant processing and achieving about a fivefold speedup. At the model level, *BoneWave* efficiently fuses complementary audio and mmWave cues with a transformer-based model, achieving stronger enhancement with fewer parameters.

6 DISCUSSION

6.1 Deployment Considerations

Feasibility and Trends. *BoneWave* enhances phone calls while supporting flexible handheld use, but raises considerations regarding hardware cost, integration, and power consumption. Advances in mmWave radar have helped lower these barriers to mobile deployment. Consumer-grade radar chips from ICLEGEND MICRO cost about $2 [28], while Infineon and TI modules are priced around $10–20 [16, 17, 41], indicating affordable integration. *BoneWave* uses a SISO radar and a low-power MEMS microphone; smartphone-grade MEMS microphones consume about 1.8 mW [29], while compact radars operate from sub-watt to watt-level power, e.g., 1.19-1.75 W for IWR6843 [41] and up to ≈0.36 W for BGT60TR13C [16]. These estimates provide a rough power budget for mobile deployment.

Meanwhile, mmWave radars have already been deployed in consumer mobile devices. For example, Google Pixel 4 phone [34] integrates Soli radar with three receive antennas for gesture interaction, meeting strict mobile size and power constraints. In contrast, *BoneWave*'s SISO design may offer lower power potential. Portable devices such as AIBI PocketPet [22] combine low-power radar sensing with on-device processing for interactive applications. These suggest potential for future mobile integration of *BoneWave*.

Future Ear-Worn Use. As wireless earbuds become increasingly common for hands-free calls, conventional in-ear and semi-in-ear designs maintain stable contact with the ear canal, making IMU-based bone-conduction sensing a natural choice for speech enhancement. In contrast, emerging form factors, such as ear-clip (worn on the outer ear) and open-ear designs, may lack effective contact with vibration-rich regions near the ear. *BoneWave*'s contactless, centimeter-scale mmWave sensing and motion robustness offer potential for such devices. While currently focused on handheld calls, the proposed technique may extend to suitable ear-worn devices.

6.2 Limitations and Future Work

Initialization Persistence and Recalibration. Our user-adaptive initialization mitigates mmWave artifacts at typical near-ear phone-call distances (1-3 cm). In controlled evaluations, *BoneWave* consistently captures usable bone-conducted vibrations across common device placements, daily activities, and sensing obstructions after a single initialization (Sec. 4.4). The configuration also maintains high-quality sensing across phone-to-ear distances of 1-10 cm. Real-world tests further demonstrate its effectiveness during continuous, unrehearsed use, potentially involving incidental changes in grip, phone pose, and left/right-ear use. Yet, longer-term persistence across calls or days remains untested, as changes in hairstyle, body weight, or phone surface may alter near-ear reflections and require recalibration. Future work will develop a monitoring mechanism that tracks signal-quality shifts over time to identify persistent deviations from the baseline and trigger re-initialization when needed.

Unvoiced Speech. Unvoiced consonants produce little or no vibration, limiting their capture by *BoneWave*. To quantify this, we derive speech-component masks from paired microphone audio and mmWave spectrograms and measure the temporal portions of the audio speech mask absent from the mmWave mask. These intervals are brief, intermittent, and interspersed with voiced speech, consistent with prior findings [2], collectively accounting for 14.9±6.7% of natural speech duration on average. Voiced components such as vowels typically exhibit higher energy and longer duration, enabling *BoneWave* to effectively enhance natural speech. Yet, we do not separately evaluate the enhancement of unvoiced consonants. Moreover, *BoneWave* fails for whispering, which relies primarily on turbulent airflow rather than vocal-cord vibration.

Signal Attenuation. We observe that near-ear bone-conducted vibrations can weaken for users with obesity, consistent with prior findings on soft-tissue attenuation [25]. Notably, the lateral neck region can still provide detectable vibration cues, but is more sensitive to handheld pose changes, as varying radar viewing angles may make effective reflections sparse or unstable. In addition, *BoneWave* degrades at some rarely used placements, such as the temple, upper-ear, and posterior-lateral regions, where bone-conducted vibrations are weak. We consider that phone inertial sensing can help estimate handheld pose, while a multi-antenna mmWave design, e.g., with differently oriented antennas, can potentially broaden sensing coverage. Future work will combine them through pose-guided radar beamforming across sensing regions to adaptively locate vibration-rich regions under severe body- or placement-related attenuation.

Integration and Optimization. *BoneWave* is not yet fully integrated into smartphones, with its mouth-to-ear latency and power consumption yet to be validated. While our offline smartphone benchmark confirms real-time computation, it does not capture additional delays from factors such as codec processing, data transmission, and playback. Future work will explore custom radar platforms for tighter smartphone integration, considering antenna placement, enclosure materials, and hardware/software interfacing. We will adapt *BoneWave*'s signal-processing pipeline and model for streaming with fine-grained buffering, toward transparent phone-call interaction. We will also characterize system-level power to guide power-efficient configurations for bone-conduction sensing while preserving enhancement quality.

7 CONCLUSION

This paper presents *BoneWave*, leveraging contactless mmWave bone-conduction sensing for flexible phone-call quality enhancement. It shows the feasibility of close-range mmWave sensing for near-ear calls, enabled by adaptive radar initialization that mitigates near-field artifacts under user variability and robust suppression of motion interference during handheld use. It further improves speech enhancement efficiency through data-side refinement and a lightweight transformer-based model that effectively fuses detail-rich audio cues with noise-robust mmWave cues. Extensive evaluation demonstrates the robustness and practicality of *BoneWave*, making it a promising solution for phone-call enhancement.

ACKNOWLEDGMENTS

This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (No. JYB2025XDXM113) and the National Natural Science Foundation of China (No. 62332016).

REFERENCES

[1] 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.

[2] 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.

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

[4] 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).

[5] Google. 2023. Google WebRTC. [Online]. https://webrtc.org/.

[6] Yole Group. 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/.

[7] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Interspeech 2020*. 5036–5040.

[8] 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.

[9] Fredric J Harris. 1978. On the use of windows for harmonic analysis with the discrete Fourier transform. *Proc. IEEE* 66, 1 (1978), 51–83.

[10] Chenming He, Chengzhen Meng, Xiaoran Fan, Dequan Wang, Lu Zhang, Yubo Yan, and Yanyong Zhang. 2026. mmOVD$^+$: Detecting Fully Occluded Vehicles With Millimeter-Wave Signals. *IEEE Transactions on Mobile Computing (TMC)* (2026).

[11] 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.

[12] Chenming He, Rui Xia, Chengzhen Meng, Xiaoran Fan, Dequan Wang, Haojie Ren, Jianmin Ji, and Yanyong Zhang. 2025. Ghost Points Matter: Far-Range Vehicle Detection with a Single mmWave Radar in Tunnel. In *Proceedings of the 31st Annual International Conference on Mobile Computing and Networking (MobiCom)*. 650–666.

[13] 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.

[14] 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.

[15] Toni Heittola, Annamaria Mesaros, and Tuomas Virtanen. 2018. TUT urban acoustic scenes 2018, development dataset. (2018).

[16] Infineon. 2023. BGT60TR13C. [Online]. https://www.infineon.com/part/BGT60TR13C.

[17] Infineon. 2025. BGT60CUTR13AIP. [Online]. https://www.infineon.com/part/BGT60CUTR13AIP.

[18] ITU-T. 2003. *G.114: One-Way Transmission Time.* ITU-T Recommendation G.114. International Telecommunication Union.

[19] 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.

[20] Zizhen Lin, Junyu Wang, Ruili Li, Fei Shen, and Xi Xuan. 2025. PrimeK-net: Multi-scale spectral learning via group prime-kernel convolutional neural networks for single channel speech enhancement. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

[21] 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.

[22] LivingAI. 2026. AIBI Pocket. [Online]. https://living.ai/product/aibi-pocket/.

[23] 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.

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

[25] Jameson K Mattingly, Nathaniel T Greene, Herman A Jenkins, Daniel J Tollin, James R Easter, and Stephen P Cass. 2015. Effects of skin thickness on cochlear input signal using transcutaneous bone conduction implants. *Otology & Neurotology* 36, 8 (2015), 1403–1411.

[26] 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.

[27] Chengzhen Meng, Chenming He, Dequan Wang, Yuxuan Xiao, Lingyu Wang, Xiaoran Fan, Lu Zhang, and Yanyong Zhang. 2025. Gr-fall: A fall detection system with gait recognition for indoor environments using siso mmwave radar. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 9, 3 (2025), 1–26.

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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/.

[35] 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.

[36] Hermann Rohling. 2007. Radar CFAR thresholding in clutter and multiple target situations. *IEEE transactions on aerospace and electronic systems* 4 (2007), 608–621.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] TI. 2020. IWR6843 intelligent mmWave sensor standard antenna plug-in module. [Online]. https://www.ti.com/tool/IWR6843ISK.

[42] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems (NeurIPS)* 30 (2017).

[43] Dequan Wang, Chenming He, Lingyu Wang, Chengzhen Meng, Xiaoran Fan, and Yanyong Zhang. 2026. FlowGait: Enabling Robust Long-Term Gait Recognition Across Real-World Covariates with mmWave Radar. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI 26)*. 1–23.

[44] 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.

[45] 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. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 9, 4 (2025), 1–33.

[46] Lei Wang, Xingwei Wang, Xi Zhang, Xiaolei Ma, Yu Zhang, Fusang Zhang, Tao Gu, and Haipeng Dai. 2025. AccCall: Enhancing Real-time Phone Call Quality with Smartphone's Built-in Accelerometer. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (IMWUT)* 9, 3 (2025), 1–33.

[47] Lingyu Wang, You Zuo, Dequan Wang, Chenming He, Chengzhen Meng, Xinran Zhang, Xiaoran Fan, and Yanyong Zhang. 2026. FeelWave: Enabling Emotion-Aware Voice Interaction through Noise-Robust mmWave Emotion Sensing. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI 26)*. 1–24.

[48] Wikipedia. 2024. Voice frequency - Wikipedia. [Online]. https://en.wikipedia.org/wiki/Voice_frequency.

[49] Peng Xu, Xiatian Zhu, and David A Clifton. 2023. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 45, 10 (2023), 12113–12132.

[50] Hongfei Xue, Yan Ju, Chenglin Miao, Yijiang Wang, Shiyang Wang, Aidong Zhang, and Lu Su. 2021. mmMesh: Towards 3D real-time dynamic human mesh construction using millimeter-wave. In *Proceedings of the 19th Annual International Conference on Mobile Systems, Applications, and Services (MobiSys)*. 269–282.

[51] 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.

[52] Shiming Zhang, Yosuke Sugiura, Nozomiko Yasui, and Tetsuya Shimamura. 2020. Quantifying noise robustness of bone-conducted speech. In *2020 IEEE 63rd international midwest symposium on circuits and systems (MWSCAS)*. IEEE, 582–585.

[53] 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.

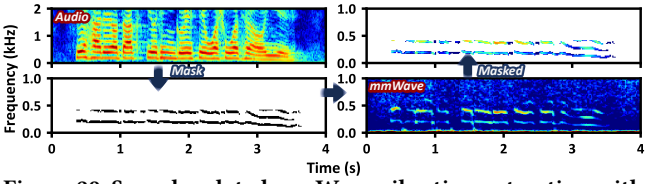


Figure 29: Speech-related mmWave vibration extraction with audio-derived masks.

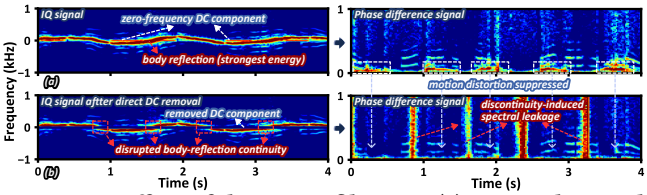


Figure 30: Effect of direct DC filtering. (a) Original IQ and phase-difference signals; (b) after DC removal.

A CA-CFAR PARAMETER SETTING

This appendix details the CA-CFAR parameter setting for constructing the speech-related frequency mask. In CA-CFAR, guard cells around the cell under test are excluded to avoid target-energy leakage, while training cells estimate the local background level.

Guard cells. We set the guard-cell width according to the STFT-broadened spread of a vocal peak. For the Hann window, the null-to-null mainlobe width of a stable spectral component is $\Delta f_{\text{mainlobe}} \approx 4f_s/L$, where f_s is the sampling rate and L is the window length. Since the frequency-bin spacing is $\Delta f_{\text{bin}} = f_s/L$, the mainlobe spans approximately $W_{\text{mainlobe}} = \Delta f_{\text{mainlobe}}/\Delta f_{\text{bin}} \approx 4$ bins. We thus use two guard cells on each side to cover the tested peak's main energy spread, excluding it from the background estimate.

Training cells. We set the training span based on vocal harmonic spacing. Adjacent vocal harmonics are separated by one fundamental-frequency interval, and the human fundamental frequency is typically above 90 Hz [48]. In our implementation, $f_s = 2500$ Hz and $L = 150$, giving $\Delta f_{\text{bin}} = \frac{f_s}{L} \approx 16.67$ Hz. We therefore use three training cells on each side, corresponding to approximately 50 Hz, which keeps the background estimate local while reducing the chance of including adjacent harmonic peaks.

Threshold coefficient. In CA-CFAR, the threshold for a target false-alarm rate P_{FA} is $\alpha = N(P_{\text{FA}}^{-1/N} - 1)$, where N is the total number of training cells. With three training cells on each side, $N = 6$; setting $P_{\text{FA}} = 10^{-3}$ yields $\alpha \approx 12.97$.

In Sec. 2.1, we derive the mask from the clean-audio spectrogram and apply it below 0.5 kHz to the corresponding mmWave spectrogram region to extract speech-related vibration components while avoiding false extraction of background components (Fig. 29).

B ANALYSIS OF DIRECT DC FILTERING

As shown in Fig. 30(a), the mmWave IQ spectrogram contains both the zero-frequency DC component and the body reflections, whose coupling during motion introduces distortion in the phase-difference signal. The body reflection often carries the strongest energy due to the large reflection area. During static periods or at motion onset, the Doppler velocity of the body reflection is close to zero, causing it to overlap with DC. Although direct DC removal suppresses motion-induced distortion in the phase-difference signal, it also removes part of the overlapping body reflection (Fig. 30(b)).

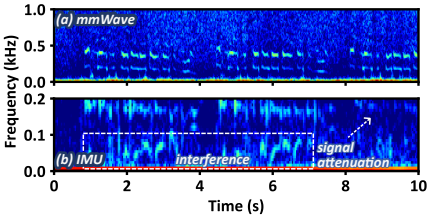


Figure 31: Bone-conducted vibrations captured by mmWave radar and a smartphone IMU.

This disrupts body-reflection continuity in the IQ signal and produces broadband spectral leakage after phase-difference computation. This motivates our motion-gated design, which confines DC suppression to motion periods while preserving signal continuity.

C EVALUATION OF SPEECH DETECTION

We evaluate our speech-segment extraction using paired speech and mmWave recordings. Since precise manual boundary annotation is challenging, we use frame-level SileroVAD outputs on clean speech as pseudo-ground-truth labels. For fair comparison, all methods use the same 100 ms temporal dilation and 60 ms minimum silence duration for merging short pauses. We evaluate frame-level speech-segment detection accuracy across SNRs from 15 dB to −15 dB. As shown in Fig. 16, our method remains stable at 0.97 across all SNRs, benefiting from noise-robust mmWave-captured bone-conducted vibrations. In contrast, the audio-based baselines degrade as noise increases, dropping from 0.97 to 0.68 for SileroVAD and from 0.83 to 0.43 for WebRTC VAD at -15 dB. This shows that mmWave-based speech-segment extraction is more reliable in noisy environments.

D FEATURE EXTRACTION ARCHITECTURE

BoneWave uses two modality-specific branches to process audio and mmWave inputs, respectively. Each branch consists of a Dense Encoder [23] followed by two Time-Spectral Conformer Blocks [7, 23] (TSCBs). The Dense Encoder applies a 1×1 convolution for channel projection, followed by a dilated DenseNet [23] and a frequency-strided convolution. The DenseNet contains four convolutions with dilation factors of $\{1, 2, 4, 8\}$, each concatenating all preceding layer outputs. Each TSCB then applies one Conformer along the temporal axis and another along the spectral axis. Within each Conformer, self-attention models long-range dependencies, while convolution captures local patterns, followed by feed-forward transformation.

E COMPARISON OF MMWAVE AND IMU SIGNAL QUALITY

With mmWave radar and a Xiaomi 14 smartphone's built-in IMU, we synchronously capture bone-conducted vibrations. We develop a bridge app using the Android Sensor Framework to acquire IMU measurements. Consistent with AccCall [46], we use the screen-normal IMU axis, sample at 400 Hz (close to the Xiaomi 14's IMU limit of 416 Hz), and apply a biquad filter to suppress motion interference. Participants walk and speak while holding the phone near the ear, maintaining natural phone-skin contact. A microphone simultaneously records clean speech to generate a shared speech mask. Given the IMU's sampling-rate limit, we calculate both signals' average-power SNRs over 0-200 Hz using this mask for a fair comparison. Fig. 28 shows variations in signal quality: contactless mmWave sensing maintains high-quality bone-conducted signals, whereas the IMU exhibits substantial signal attenuation (Fig. 31(b)) as phone-skin contact naturally varies in pressure and orientation.