FOCUSME: BONE-CONDUCTION-GUIDED SPEECH ENHANCEMENT IN CHALLENGING ACOUSTIC SCENARIOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recovering a wearer's speech from an air-conduction (AC) mixture requires both suppressing interference and identifying the target speaker, particularly when the interference includes other voices. We propose FocusMe, a causal speech enhancement system that uses synchronized bone-conduction (BC) speech for two complementary purposes: acoustic reconstruction and enrollment-free speaker conditioning. Cross-Attention Feature Fusion retrieves BC information relevant to the AC mixture and adaptively combines the two signals, while a causal extractor supplies frame-level BC speaker representations to the reconstruction network. The system is trained jointly for speech reconstruction and speaker classification. Across 13 test conditions constructed from ABCS and ESMB, FocusMe achieves the best average perceptual quality, intelligibility, and recognition accuracy among the compared systems, reducing mean character error rate from 22.85% for DC-CRN to 17.03%. These results demonstrate the effectiveness of jointly exploiting BC speech for acoustic reconstruction and speaker conditioning under severe noise and competing speech.

1 INTRODUCTION

In wearable devices, recovering the wearer's voice requires suppressing environmental noise and competing speakers while preserving speech quality and intelligibility (He et al., 2023). Air-conduction (AC) microphones capture full-band speech, but strong background noise can obscure the acoustic details needed for reconstruction (Wang et al., 2022a). Competing speech introduces a further challenge: the system must determine which speaker to preserve, even though both target and interfering voices share speech-like structure. This requires a cue that identifies the target speaker (Žmolíková et al., 2017).

Target-speaker extraction (TSE) provides this cue by conditioning speech recovery on information about the desired speaker. Enrollment-based approaches derive a speaker representation from a separately recorded utterance and use it as a speaker cue for extraction (Žmolíková et al., 2017; Wang et al., 2019; Ge et al., 2020). Short or noisy enrollment utterances can weaken the target representation (Veluri et al., 2024), while variation across utterances from the same speaker can affect extraction reliability (Sato et al., 2022).

Bone-conduction (BC) sensors capture speech-induced vibrations through the wearer's skull and tissues, providing a synchronized observation of the wearer's speech that is less sensitive to airborne noise and competing speech (Wang et al., 2022a; He et al., 2023). Existing systems combine BC and AC signals through learned cross-modal fusion (Yu et al., 2020; Wang et al., 2022b;a; Wu et al., 2026; He et al., 2023). However, high-frequency attenuation limits the spectral detail available in BC speech, making its contribution to reconstruction uneven across frequency bands (Wang et al., 2022a). These fusion-based approaches primarily learn joint acoustic representations, leaving the target-speaker cue implicit rather than explicitly supervised. Under competing speech, target selection therefore relies on acoustic reconstruction without an explicit speaker-discriminative objective.

To address these limitations, we propose FocusMe, a causal AC–BC speech enhancement framework that combines acoustic fusion with an explicit, enrollment-free conditioning cue. As shown in Figure 1, a frame-level speaker embedding (FSE) extractor learns speaker representations from

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

synchronized BC speech and provides time-aligned speaker cues to the reconstruction backbone. A Cross-Attention Feature Fusion (CAFF) module uses AC features to retrieve complementary BC information and adaptively fuses the modalities for complex-domain reconstruction. We jointly train the FSE extractor and reconstruction backbone with speaker classification and speech reconstruction objectives, allowing BC speech to provide both complementary acoustic information and speaker-discriminative cues.

Experiments on ABCS (Wang et al., 2023) and ESMB (Wang et al., 2022a) show improvements in speech quality, intelligibility, and recognition accuracy. At SNR -15 dB and SIR -10 dB, the most challenging test condition, FocusMe reduced character error rate (CER) by 26.9% relative to DCCRN (Wang et al., 2022a). These results support the benefit of combining acoustic fusion with BC-derived speaker conditioning for perceptual quality, intelligibility, and recognition.

2 Related Work

2.1 Target-Speaker Extraction

Target-speaker extraction uses auxiliary cues to recover a desired voice from a speech mixture. These cues can encode speaker characteristics, spatial location, or visual information such as lip movements (Žmolíková et al., 2023; Li et al., 2026a;b; 2025b; 2024b;a). Enrollment-based methods derive speaker information from a reference utterance and use it as a speaker cue for extraction (Žmolíková et al., 2017; Wang et al., 2019; Ge et al., 2020). Recent work studies the reliability and acquisition of enrollment cues under varied recording conditions (Sato et al., 2022; Veluri et al., 2024; Shen et al., 2025). These approaches still require a separate reference recording. FocusMe instead derives frame-level conditioning representations from BC speech recorded synchronously with the AC mixture, eliminating the need for a separate enrollment utterance.

2.2 Bone-Conduction-Assisted Speech Enhancement

BC sensing reduces sensitivity to airborne interference, but attenuates the high-frequency components needed for natural speech reproduction (Wang et al., 2022a). Bandwidth extension methods such as EBEN reconstruct wideband speech from noise-resilient body-conduction recordings (Hauret et al., 2023). Joint AC-BC enhancement combines complementary information from both signals (Yu et al., 2020; Wang et al., 2022b;a; He et al., 2023). Recent approaches further develop cross-modal interaction and representation learning. DBMIF combines iterative attention, gated information exchange, and a balanced-interaction bottleneck to refine AC-BC representations (Wu et al., 2026). BiNet uses separate encoders and contrastive regularization to encourage similarity between the two modalities in a shared embedding space (Li et al., 2025a). MBENet also applies BC-AC fusion specifically to target-speaker extraction (Zhang et al., 2025). These studies establish acoustic fusion and cross-modal representation learning as important components of BC-assisted speech recovery. FocusMe combines adaptive spectral fusion with a dedicated BC branch that produces speaker cues. Its frame-level speaker classification objective explicitly supervises speaker discrimination, while CAFF retrieves complementary BC information for AC reconstruction.

3 FocusMe

3.1 Overall Pipeline

FocusMe recovers the target speaker's clean AC speech from an audio mixture guided by synchronized BC speech. The mixture is

$$\mathbf{x}_{\mathrm{ac}} = \mathbf{s} + \mathbf{n} + \mathbf{i}, \tag{1}$$

where $\mathbf{s}$, $\mathbf{n}$, $\mathbf{i}$ denote target speech, background noise, and interfering speech.

As shown in Figure 1, an audio encoder applies an STFT to $\mathbf{x}_{\mathrm{ac}}$ and $\mathbf{x}_{\mathrm{bc}}$, yielding the real/imaginary components $\mathbf{X}_{\mathrm{ac}}, \mathbf{X}_{\mathrm{bc}} \in \mathbb{R}^{T\times F\times 2}$, where T and F are the numbers of time frames and frequency bins. CAFF fuses them into acoustic features $\mathbf{X}_{\mathrm{feat}} \in \mathbb{R}^{T\times F\times 6}$. In parallel, the frame-level speaker embedding (FSE) extractor maps BC log-Mel features $\mathbf{M}_{\mathrm{bc}} \in \mathbb{R}^{C_m\times T}$ to aligned embeddings $\mathbf{E}_{\mathrm{bc}} \in \mathbb{R}^{D_e\times T}$, with C_m and D_e the Mel-bin and embedding widths. The reconstruction backbone

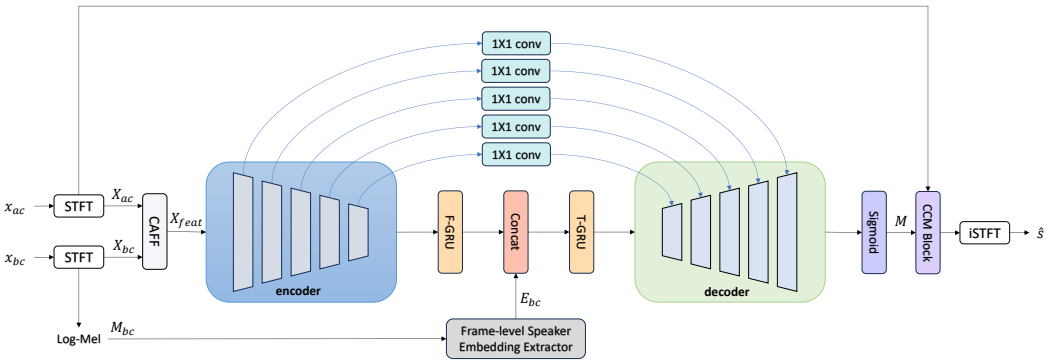


Figure 1: Overall pipeline of FocusMe. The audio encoder produces AC and BC spectra for CAFF. In parallel, the FSE extractor provides speaker embeddings to the reconstruction backbone, whose predictions are converted to speech by the audio decoder.

processes $\mathbf{X}_{\text{feat}}$ with $\mathbf{E}_{\text{bc}}$ injected into its temporal bottleneck to predict reconstruction coefficients $\mathbf{M} \in \mathbb{R}^{C_o \times T \times F}$, where C_o is the number of coefficients per time–frequency bin. Finally, the audio decoder combines $\mathbf{M}$ with $\mathbf{X}_{\text{ac}}$ through masking and inverse STFT (iSTFT) to recover the estimated speech $\widehat{\mathbf{s}} \in \mathbb{R}^L$.

3.2 Cross-Attention Feature Fusion (CAFF)

CAFF uses AC features to query a compact BC representation and learns a per-bin modality combination (Figure 2).

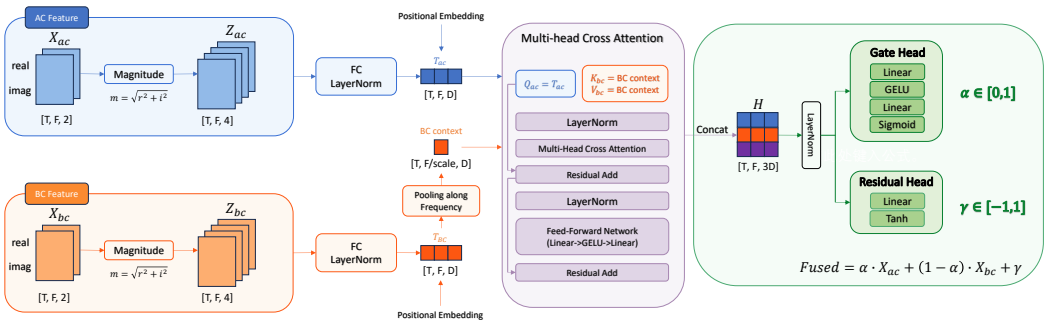


Figure 2: CAFF uses position-augmented tokens to retrieve pooled BC information with AC queries, then fuses the spectra through a learned gate and bounded residual.

For modality $m \in \{\text{ac}, \text{bc}\}$, let $\mathbf{X}_{m,\text{r}}, \mathbf{X}_{m,\text{i}} \in \mathbb{R}^{T \times F}$ be the real and imaginary components and $\mathbf{X}_{m,\text{mag}} = |\mathbf{X}_m|$. We form four-channel features

$$\mathbf{Z}_m = [\mathbf{X}_{m,\text{r}}, \mathbf{X}_{m,\text{i}}, \mathbf{X}_{m,\text{mag}}, \log(1 + \mathbf{X}_{m,\text{mag}})] \tag{2}$$

and map them per time–frequency bin to tokens $\mathbf{T}_m \in \mathbb{R}^{T \times F \times D}$ by a pointwise projection ($\mathbf{W}_m \in \mathbb{R}^{D \times 4}$, bias $\mathbf{b}_m$) plus positional embeddings $\mathbf{P}_m$:

$$\mathbf{T}_m = \mathcal{N}(\mathbf{Z}_m \mathbf{W}_m^\top + \mathbf{b}_m) + \mathbf{P}_m, \tag{3}$$

where D is the token width, $\mathbf{P}_m$ matches the projected tokens, and $\mathcal{N}$ is channel-wise layer normalization (Ba et al., 2016).

To retrieve BC information for each AC bin, we use $\mathbf{Q}_{\text{ac}} = \mathbf{T}_{\text{ac}}$ as queries and adaptively average-pool $\mathbf{T}_{\text{bc}}$ to F_s frequency bins for keys and values $\mathbf{K}_{\text{bc}}, \mathbf{V}_{\text{bc}} \in \mathbb{R}^{T \times F_s \times D}$. Cross-attention along

frequency within each frame, followed by attention and feed-forward residual blocks, produces AC-aligned features

$$\widetilde{\mathbf{C}} = \mathbf{Q}_{\mathrm{ac}} + \mathcal{D}(\mathcal{A}(\mathcal{N}(\mathbf{Q}_{\mathrm{ac}}), \mathcal{N}(\mathbf{K}_{\mathrm{bc}}), \mathcal{N}(\mathbf{V}_{\mathrm{bc}}))), \tag{4}$$

$$\mathbf{C}' = \widetilde{\mathbf{C}} + \mathcal{D}\Big(\mathcal{H}\Big(\mathcal{N}(\widetilde{\mathbf{C}})\Big)\Big), \tag{5}$$

where $\mathcal{A}$ is multi-head attention, $\mathcal{H}$ a two-layer feed-forward network, and $\mathcal{D}$ dropout. The residual AC queries keep $\mathbf{C}'$ on the AC frequency grid, so attention only adds pooled BC information while the unpooled BC tokens remain available for fusion. Concatenating the AC-aligned features with the original tokens gives $\mathbf{H} \in \mathbb{R}^{T\times F\times 3D}$:

$$\mathbf{H} = \mathcal{N}(\mathrm{Concat}(\mathbf{T}_{\mathrm{ac}}, \mathbf{T}_{\mathrm{bc}}, \mathbf{C}')). \tag{6}$$

A gate head $g : \mathbb{R}^{3D} \to \mathbb{R}^{D} \to \mathbb{R}$ (two linear layers separated by GELU) predicts the relative modality contribution, and a residual head $r : \mathbb{R}^{3D} \to \mathbb{R}^2$ predicts a two-channel correction:

$$\boldsymbol{\alpha} = \sigma(g(\mathbf{H})), \qquad \boldsymbol{\gamma} = \eta \tanh(r(\mathbf{H})), \tag{7}$$

with $\boldsymbol{\alpha} \in (0,1)^{T\times F}$, $\boldsymbol{\gamma} \in (-\eta, \eta)^{T\times F\times 2}$, $\eta > 0$ bounding the residual, and the gate shared across real and imaginary channels. The fused spectrum and final features are

$$Fused = \boldsymbol{\alpha} \odot \mathbf{X}_{\mathrm{ac}} + (1 - \boldsymbol{\alpha}) \odot \mathbf{X}_{\mathrm{bc}} + \boldsymbol{\gamma}, \tag{8}$$

$$\mathbf{X}_{\mathrm{feat}} = \mathrm{Concat}(\mathbf{X}_{\mathrm{ac}}, \mathbf{X}_{\mathrm{bc}}, Fused) \in \mathbb{R}^{T\times F\times 6}, \tag{9}$$

where $\odot$ broadcasts over channels, so both original spectra are retained alongside their adaptive fusion.

3.3 Frame-Level Speaker Embedding (FSE) Extractor

The FSE extractor maps BC log-Mel features $\mathbf{M}_{\mathrm{bc}}$ to frame-level speaker embeddings $\mathbf{E}_{\mathrm{bc}}$, providing a time-aligned speaker cue from the synchronized BC recording without a separate enrollment utterance. The causal temporal convolutions in Figure 3 preserve frame resolution without utterance-level pooling.

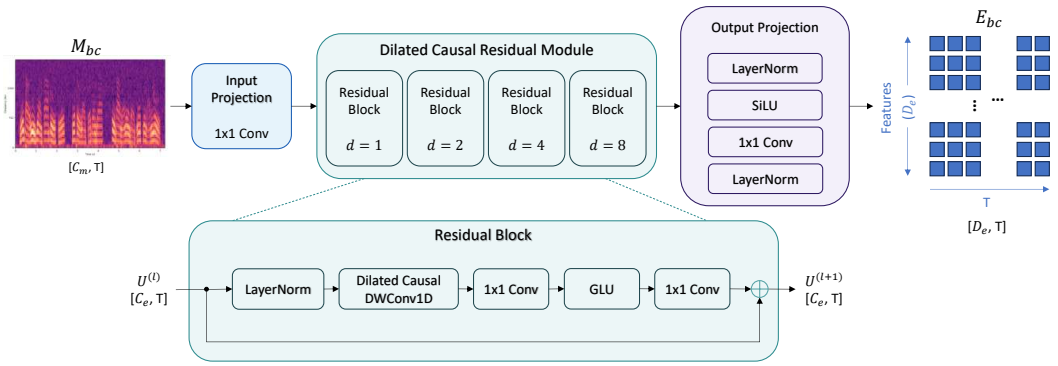


Figure 3: The FSE extractor preserves the temporal resolution through causal residual blocks and produces one speaker representation per frame.

A first pointwise projection $\mathcal{P}_0 : \mathbb{R}^{C_m} \to \mathbb{R}^{C_e}$ maps each BC log-Mel frame to a shared feature space of width C_e,

$$\mathbf{U}^{(0)} = \mathcal{P}_0(\mathbf{M}_{\mathrm{bc}}) \in \mathbb{R}^{C_e\times T}, \tag{10}$$

and N_e residual blocks aggregate temporal context at full frame resolution. For block $l \in \{1, \ldots, N_e\}$,

$$\mathbf{H}^{(l)} = \mathcal{N}(\mathbf{U}^{(l-1)}), \qquad \mathbf{Z}_1^{(l)} = \mathcal{W}_l(\mathbf{H}^{(l)}), \tag{11}$$

$$\mathbf{Z}_2^{(l)} = \mathrm{GLU}(\mathcal{P}_l^{\mathrm{in}}(\mathbf{Z}_1^{(l)})), \qquad \mathbf{U}^{(l)} = \mathbf{U}^{(l-1)} + \mathcal{P}_l^{\mathrm{out}}(\mathbf{Z}_2^{(l)}), \tag{12}$$

where $\mathcal{W}_l$ is a depthwise temporal convolution with kernel size K_e and dilation d_l, using $(K_e-1)d_l$ left-padded frames so each output depends only on the current and preceding frames; GLU is the gated linear unit; and $\mathcal{P}_l^{\text{in}} : \mathbb{R}^{C_e} \to \mathbb{R}^{2C_e}$, $\mathcal{P}_l^{\text{out}} : \mathbb{R}^{C_e} \to \mathbb{R}^{C_e}$ expand channels before gating and return to the residual path. A final projection $\mathcal{P}_e : \mathbb{R}^{C_e} \to \mathbb{R}^{D_e}$ yields

$$\mathbf{E}_{\text{bc}} = \mathcal{N}\Big(\mathcal{P}_e\Big(\text{SiLU}(\mathcal{N}(\mathbf{U}^{(N_e)}))\Big)\Big), \tag{13}$$

so each frame t carries a speaker vector $\mathbf{e}_t \in \mathbb{R}^{D_e}$ aligned with the acoustic features and passed to the temporal bottleneck.

3.4 Speaker-Conditioned Reconstruction Backbone

The reconstruction backbone predicts coefficients from CAFF features and FSE embeddings using a causal convolutional encoder–decoder with two bottleneck recurrent units: a frequency GRU (F-GRU) models spectral structure within each frame, and a temporal GRU (T-GRU) integrates the speaker cue over time.

The convolutional encoder reads the six channels of $\mathbf{X}_{\text{feat}}$ in channel-first order and extracts local features through N causal 2-D convolutional stages, each applying convolution, batch normalization (Ioffe & Szegedy, 2015), and ELU activation (Clevert et al., 2016) while preserving time resolution and halving frequency resolution. Stage k outputs $\mathbf{A}^{(k)} \in \mathbb{R}^{C_k \times T \times F_k}$ with width C_k and frequency resolution F_k. At the bottleneck, a bidirectional F-GRU models within-frame frequency dependencies, followed by a layer-normalized projection to C_N channels and residual addition.

Speaker conditioning enters before temporal modeling. Flattening the F-GRU output across channels and frequency gives $\mathbf{R} \in \mathbb{R}^{T \times (C_N F_N)}$, to which the aligned BC embeddings are concatenated:

$$\mathbf{V} = [\mathbf{R} \| \mathbf{E}_{\text{bc}}^{\top}]. \tag{14}$$

A unidirectional T-GRU processes $\mathbf{V}$; its output is projected to $C_N F_N$ dimensions, layer-normalized, and reshaped for residual addition to the acoustic features.

The convolutional decoder restores full frequency resolution with N sub-pixel convolutional stages (Shi et al., 2016). Let $\mathbf{D}^{(k)} \in \mathbb{R}^{G_k \times T \times J_k}$ denote its stage-k features (width G_k, frequency resolution J_k), initialized by the bottleneck output with $G_0 = C_N$ and $J_0 = F_N$. To retain encoder information at each resolution, the decoder uses projected additive skip connections:

$$\mathbf{D}^{(k)} = \mathcal{U}_k\Big(\mathbf{D}^{(k-1)} + \mathcal{B}\Big(\mathcal{P}_k^{\text{skip}}(\mathbf{A}^{(N-k+1)})\Big)\Big). \tag{15}$$

Here $\mathcal{P}_k^{\text{skip}} : \mathbb{R}^{C_{N-k+1}} \to \mathbb{R}^{G_{k-1}}$ is a pointwise projection, $\mathcal{B}$ batch normalization, and $\mathcal{U}_k$ the sub-pixel stage; matched resolutions satisfy $F_{N-k+1} = J_{k-1}$, and each stage doubles the frequency resolution. Batch normalization and ELU follow every stage except the output stage, which restores $J_N = F$ with $G_N = C_o$ channels. A frequency-dependent learnable sigmoid ς (Zhang & Woodland, 2015) finally produces $\mathbf{M} \in (0, a)^{C_o \times T \times F}$ with bound $a > 0$:

$$\mathbf{M} = a\,\varsigma(\mathbf{D}^{(N)}). \tag{16}$$

These coefficients are passed to the audio decoder together with the AC spectrum.

3.5 Audio Decoder

The audio decoder reconstructs the enhanced waveform from the coefficients and the AC spectrum. Its complex reconstruction operator $\mathcal{K}$ (Ristea et al., 2023) maps $\mathbf{X}_{\text{ac}}$ and $\mathbf{M}$ to a compressed target-spectrum estimate $\widetilde{\mathbf{S}}_{\text{tgt}} \in \mathbb{C}^{T \times F}$:

$$\widetilde{\mathbf{S}}_{\text{tgt}} = \mathcal{K}(\mathbf{X}_{\text{ac}}, \mathbf{M}). \tag{17}$$

Magnitude compression $\mathcal{C}_p$ raises spectral magnitudes to a power $p \in (0, 1)$ while preserving phase; its inverse recovers $\widehat{\mathbf{S}}_{\text{tgt}} \in \mathbb{C}^{T \times F}$:

$$\widehat{\mathbf{S}}_{\text{tgt}} = |\widetilde{\mathbf{S}}_{\text{tgt}}|^{1/p-1} \odot \widetilde{\mathbf{S}}_{\text{tgt}}. \tag{18}$$

Finally, iSTFT converts $\widehat{\mathbf{S}}_{\text{tgt}}$ into $\widehat{\mathbf{s}}$.

3.6 JOINT TRAINING OBJECTIVE

Joint training couples target-speech reconstruction with speaker discrimination: reconstruction losses supervise the estimated spectrum and waveform, while an auxiliary speaker loss encourages the BC embeddings to distinguish identities. The FSE extractor and backbone are optimized together so that the embeddings support target-speech recovery.

Let $\mathbf{S}_{\mathrm{tgt}} = \mathrm{STFT}(\mathbf{s}) \in \mathbb{C}^{T\times F}$ denote the clean target spectrum. We constrain compressed magnitudes, complex spectral values, and time-domain samples using

$$\mathcal{L}_{\mathrm{mag}} = \frac{1}{TF}\left\| |\mathcal{C}_p(\widehat{\mathbf{S}}_{\mathrm{tgt}})| - |\mathcal{C}_p(\mathbf{S}_{\mathrm{tgt}})| \right\|_F^2, \tag{19}$$

$$\mathcal{L}_{\mathrm{RI}} = \frac{1}{TF}\left\| \mathcal{C}_p(\widehat{\mathbf{S}}_{\mathrm{tgt}}) - \mathcal{C}_p(\mathbf{S}_{\mathrm{tgt}}) \right\|_F^2, \tag{20}$$

$$\mathcal{L}_{\mathrm{wav}} = \frac{1}{L}\|\widehat{\mathbf{s}} - \mathbf{s}\|_2^2, \tag{21}$$

where $\|\cdot\|_F$ and $\|\cdot\|_2$ are the Frobenius and Euclidean norms; the complex squared norm in $\mathcal{L}_{\mathrm{RI}}$ sums the real- and imaginary-part squared errors, giving phase-sensitive supervision alongside $\mathcal{L}_{\mathrm{mag}}$.

Speaker supervision is applied only to speech-active frames, so silent intervals do not bias identity learning. Let $\mathcal{V} \subseteq \{1,\dots,T\}$ be the frames selected by voice activity detection on clean AC speech, K the number of training speakers, y the target class, and $\mathbf{w}_k \in \mathbb{R}^{D_e}$ the unit-norm weight for class k. With cosine similarity $c_{t,k} = \mathbf{e}_t^\top \mathbf{w}_k / \|\mathbf{e}_t\|_2 \in [-1,1]$, margin $\mu > 0$, scale $\tau > 0$, and $\psi(c) = \cos(\arccos(c) + \mu)$, AAM-Softmax gives

$$\mathcal{L}_{\mathrm{spk}} = -\frac{1}{|\mathcal{V}|}\sum_{t\in\mathcal{V}} \log \frac{\exp(\tau\psi(c_{t,y}))}{\exp(\tau\psi(c_{t,y})) + \sum_{k\neq y}\exp(\tau c_{t,k})}. \tag{22}$$

With $\mathcal{L}_{\mathrm{rec}} = \mathcal{L}_{\mathrm{mag}} + \lambda_{\mathrm{RI}}\mathcal{L}_{\mathrm{RI}} + \mathcal{L}_{\mathrm{wav}}$ ($\lambda_{\mathrm{RI}} > 0$), the joint objective is $\mathcal{L} = \mathcal{L}_{\mathrm{rec}} + \mathcal{L}_{\mathrm{spk}}$.

4 EXPERIMENTAL SETUP

4.1 DATASETS

We constructed the training, validation, and test sets from synchronized AC and BC recordings in ABCS (Wang et al., 2023) and ESMB (Wang et al., 2022a). We mixed the target AC recordings with environmental noise and competing speech and retained the corresponding BC recordings without added corruption. All waveforms were sampled at 16 kHz, and training examples were cropped or zero-padded to 3 s segments.

We partitioned the data by speaker in an $8:1:1$ ratio, with no speaker overlap across splits. The training, validation, and test sets contained 120 h, 15 h, and 15 h of audio, respectively. The test set comprised $1,300$ mixtures across 13 conditions with varying signal-to-noise ratios (SNRs) and signal-to-interference ratios (SIRs). FocusMe and all baselines were trained and evaluated on the same data splits. Appendix A provides details of the data sources and mixture construction.

4.2 IMPLEMENTATION DETAILS

Model configuration. The audio encoder used a 32 ms Hann window, 50% overlap, and a 512-point FFT for the STFT, with magnitude compression exponent $p = 0.3$. CAFF used $C_s = 2$ real–imaginary channels, $C_z = 4$ augmented features, token width $D = 16$, and 4 attention heads. BC keys and values were pooled along frequency by a factor of 4. Both residual branches used a dropout probability of 0.1, and the spectral residual bound was $\eta = 0.25$. Concatenating the AC, BC, and fused representations yielded $C_f = 6$ input channels for the backbone.

The FSE extractor used $C_m = 80$ Mel bins spanning 20–2000 Hz, channel width $C_e = 144$, and embedding width $D_e = 192$. Its $N_e = 4$ residual blocks used temporal kernel size $K_e = 5$ and dilation rates 1, 2, 4, and 8, respectively. The backbone used $N = 5$ stages in both the convolutional encoder and decoder. The encoder used 4×3 kernels with stride $(1, 2)$ and channel widths 32, 64, 64, 64, and 128. The F-GRU and T-GRU used 128 and 512 hidden units, respectively. The decoder channel widths were 64, 64, 64, 32, and 27, and the coefficient output bound was $a = 2$.

Training configuration. We trained FocusMe for 150 epochs on eight NVIDIA RTX 4090 GPUs with an effective batch size of 32. Adam used an initial learning rate of 1×10^{-3} and zero weight decay. The learning rate was reduced by a factor of 0.9 when the training objective averaged across GPUs plateaued, with a patience of three epochs. We jointly optimized the FSE extractor and reconstruction backbone using the objective in Section 3.6, with $\lambda_{\mathrm{RI}} = 0.2$, AAM-Softmax margin $\mu = 0.2$, and scale $\tau = 30$. We selected the checkpoint with the highest validation SI-SDR.

4.3 Baselines and Evaluation Metrics

Baselines. We compared FocusMe with five AC–BC baselines: DBMIF (Wu et al., 2026), FCN (Yu et al., 2020), MMINet (Wang et al., 2022b), DCCRN (Wang et al., 2022a), and VibVoice (He et al., 2023). Appendix B describes these models.

Evaluation metrics. We evaluated waveform fidelity with SI-SDR (Le Roux et al., 2019), perceptual quality with PESQ (Rix et al., 2001) and DNSMOS (Reddy et al., 2021), and intelligibility with STOI (Taal et al., 2011) and ESTOI (Jensen & Taal, 2016). We measured recognition accuracy using the CER of Whisper (Radford et al., 2023) transcriptions. For efficiency, we reported parameter counts (in millions) and multiply–accumulate operations (MACs, in billions), profiled using the `ptflops` package[1]. Profiling used 1 s of paired AC and BC audio at 16 kHz. Table 3 also includes CPU and GPU inference time in milliseconds. Appendix C provides metric definitions, score aggregation, and timing settings.

5 Results

5.1 Comparison with State-of-the-Art Models

Table 1: Comparison with AC–BC systems across 13 test conditions with environmental noise and competing speech. Scores were averaged with equal weight assigned to each condition. Bold and underlined values indicate the best and second-best results in each column, respectively; arrows indicate whether higher (↑) or lower (↓) values are preferred.

Model	SI-SDR (dB) ↑	PESQ ↑	STOI ↑	ESTOI ↑	DNSMOS ↑	CER (%) ↓
FCN (Yu et al., 2020)	4.78	1.26	0.69	0.52	2.50	50.69
DCCRN (Wang et al., 2022a)	9.74	<u>2.29</u>	<u>0.86</u>	<u>0.78</u>	<u>3.11</u>	<u>22.85</u>
MMINet (Wang et al., 2022b)	**11.44**	1.76	0.83	0.73	2.81	29.22
VibVoice (He et al., 2023)	8.40	1.96	0.82	0.73	2.95	34.59
DBMIF (Wu et al., 2026)	7.57	1.59	0.79	0.69	2.64	36.82
FocusMe	<u>11.27</u>	**2.48**	**0.88**	**0.81**	**3.17**	**17.03**

Performance. Table 1 reports the mean scores across 13 equally weighted test conditions. FocusMe ranked first in perceptual quality (PESQ and DNSMOS), intelligibility (STOI and ESTOI), and recognition accuracy (CER). Compared with DCCRN, the strongest baseline on these five mean scores, FocusMe increased PESQ from 2.29 to 2.48 and reduced CER from 22.85% to 17.03%. At SNR -15 dB and SIR -10 dB, where both environmental noise and competing speech exceeded the target in power, FocusMe also led on these five measures. Its CER advantage over DCCRN widened from 5.82 percentage points on average to 16.71 percentage points in this condition (Table 2). Appendix D reports the results for individual SNR and SIR conditions. The spectrograms in Appendix D.3 show that FocusMe preserves target-aligned harmonic bands and higher-frequency speech structure while suppressing diffuse residual energy from competing speech at SNR 10 dB and SIR -10 dB (Figure 5) and environmental noise at SNR -15 dB and SIR 10 dB (Figure 6). In addition, we have generated audio enhanced using different methods, and sample audio is available for listening [2]. From a listening perspective, FocusMe exhibits less distortion and fewer instances of noise leakage.

[1] `https://github.com/sovrasov/flops-counter.pytorch`

[2] `https://focusme-paper.github.io/FocusMe/`

Table 2: Target-speech recovery under joint environmental noise and competing speech at SNR -15 dB and SIR -10 dB.

Model	SI-SDR (dB) ↑	PESQ ↑	STOI ↑	ESTOI ↑	DNSMOS ↑	CER (%) ↓
FCN (Yu et al., 2020)	-4.90	1.07	0.52	0.27	2.26	103.55
DCCRN (Wang et al., 2022a)	-0.74	<u>1.51</u>	<u>0.72</u>	<u>0.59</u>	2.52	<u>62.21</u>
MMINet (Wang et al., 2022b)	**4.63**	1.35	0.71	0.56	<u>2.58</u>	81.09
VibVoice (He et al., 2023)	-2.71	1.32	0.64	0.50	2.44	95.92
DBMIF (Wu et al., 2026)	-3.04	1.14	0.59	0.44	2.30	90.32
FocusMe	<u>1.30</u>	**1.59**	**0.76**	**0.64**	**2.63**	**45.50**

Efficiency. As shown in Table 3, FocusMe used 12.42% fewer parameters and 8.54% fewer MACs than DCCRN while improving perceptual quality, intelligibility, and recognition accuracy (Table 1). FocusMe had the lowest mean CPU inference time (250.479 ms) and the second-lowest mean GPU inference time (6.739 ms), behind FCN (1.586 ms). Relative to DCCRN, FocusMe combined better target-speech recovery with lower model complexity and faster inference.

Table 3: Model complexity and inference time for the compared systems. Parameter counts are in millions and MACs in billions; CPU and GPU columns denote inference time in milliseconds.

Model	Params. (M) ↓	MACs (G) ↓	CPU (ms) ↓	GPU (ms) ↓
FCN (Yu et al., 2020)	**0.30**	4.84	453.929 ± 5.261	**1.586 ± 0.005**
DCCRN (Wang et al., 2022a)	4.91	1.64	287.397 ± 2.675	93.617 ± 0.077
MMINet (Wang et al., 2022b)	3.70	7.42	1117.096 ± 33.351	9.043 ± 0.021
VibVoice (He et al., 2023)	<u>3.10</u>	3.15	<u>277.369 ± 3.087</u>	6.977 ± 0.108
DBMIF (Wu et al., 2026)	10.41	**1.49**	404.126 ± 2.470	92.259 ± 0.201
FocusMe	4.30	<u>1.50</u>	**250.479 ± 3.802**	<u>6.739 ± 0.019</u>

We ablate adaptive AC–BC fusion and BC-derived speaker cues at SNR -15 dB and SIR -10 dB, where environmental noise and competing speech are both strong (Table 4).

Table 4: Ablation study of acoustic fusion and BC speaker conditioning at SNR -15 dB and SIR -10 dB. Here w/o and w/ mean without and with; CAFF denotes cross-attention feature fusion; add and concat denote elementwise addition and channel-wise concatenation; AFF denotes the attention-fused feature of Wang et al. (2022a); FSE denotes frame-level speaker embedding; spk. loss denotes the speaker classification loss; and N_e is the number of FSE extractor residual blocks. Bold marks the best value within each ablation group and underlining the second-best.

Variant	Params. (M) ↓	MACs (G) ↓	SI-SDR (dB) ↑	PESQ ↑	STOI ↑	ESTOI ↑	DNSMOS ↑	CER (%) ↓
w/o CAFF	4.29	**1.39**	0.47	1.47	0.71	0.57	2.52	65.15
w/ add	4.29	**1.39**	0.45	1.38	0.68	0.55	2.55	72.68
w/ concat	4.29	1.39	0.77	1.48	0.72	0.59	2.54	52.38
w/ AFF	4.30	1.50	0.77	1.50	0.73	0.60	2.55	<u>48.83</u>
w/o FSE	**3.71**	<u>1.47</u>	0.86	1.51	0.72	0.59	2.54	62.08
w/o spk. loss	4.30	1.50	<u>0.98</u>	<u>1.54</u>	<u>0.74</u>	<u>0.61</u>	<u>2.55</u>	51.33
FocusMe ($N_e = 4$)	4.30	1.50	**1.30**	**1.59**	**0.76**	**0.64**	**2.63**	**45.50**
FocusMe ($N_e = 3$)	4.24	1.50	0.96	1.53	0.74	0.61	2.55	54.88
FocusMe ($N_e = 2$)	<u>4.17</u>	1.50	0.91	1.52	0.73	0.60	2.53	57.76

Acoustic fusion. We compared CAFF with four alternatives: no fusion, elementwise addition of AC and BC features, channel-wise concatenation, and AFF (Wang et al., 2022a). Removing fusion reduced SI-SDR from 1.30 to 0.47 dB and increased CER from 45.50% to 65.15%, and direct addition degraded both further (0.45 dB, 72.68%), showing that simple combination did not provide effective fusion under this condition. Channel-wise concatenation improved all six metrics over removing fusion (0.77 dB, 52.38%), and AFF reached the same SI-SDR with a lower CER of 48.83% despite

matching the full model's parameter and MAC counts at the reported precision. CAFF nevertheless achieved the best results across all six metrics, supporting its effectiveness in combining the paired AC and BC features.

Speaker conditioning. Removing the FSE extractor reduced SI-SDR to 0.86 dB and increased CER to 62.08%; omitting the speaker classification loss while retaining the extractor raised CER to 51.33% and degraded all six metrics. This distinction supports both the contribution of the BC conditioning branch and the value of explicit speaker supervision for learning its representations.

Extractor depth. Reducing N_e from 4 to 3 and 2 lowered the parameter count from 4.30 to 4.24 and 4.17 M, with MAC counts unchanged at 1.50 G at the reported precision, but SI-SDR decreased from 1.30 to 0.96 and 0.91 dB, and CER increased from 45.50% to 54.88% and 57.76%. Among the tested depths, four blocks thus gave the highest SI-SDR and the lowest CER at a modest parameter cost.

Speaker representation analysis. We extracted FSE embeddings from valid speech frames for three speakers excluded from training (100 utterances each) and projected them into two dimensions with principal component analysis (PCA). In Figure 4, points are colored by ground-truth speaker identity on the left and by unsupervised K-means assignments ($K = 3$) on the right. Frames from the same speaker occupied compact regions, and the clustering assignments largely agreed with the speaker labels, with some overlap near the cluster boundaries. This agreement suggests that frame-level BC embeddings retain speaker-discriminative information for unseen speakers.

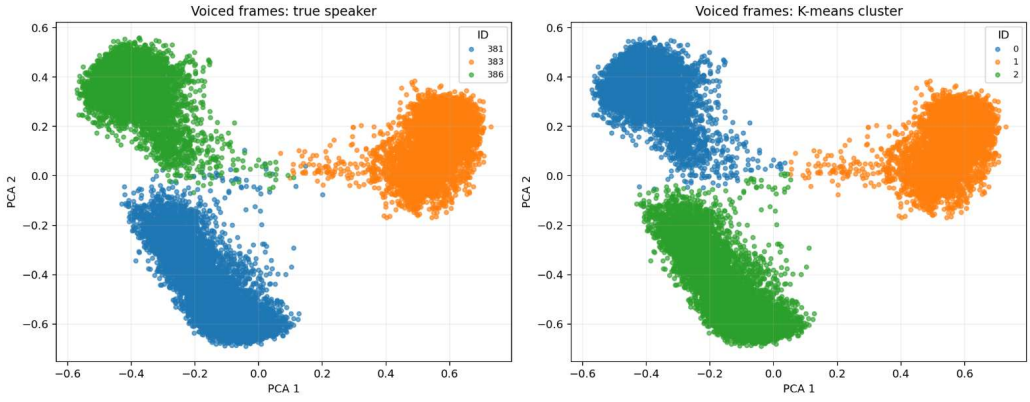


Figure 4: PCA visualization of frame-level BC speaker embeddings from three unseen speakers. Points are colored by ground-truth speaker identity (left) and K-means assignments with $K = 3$ (right).

6 Conclusion

We presented FocusMe, a causal speech enhancement framework that uses synchronized BC speech as both complementary acoustic evidence and an enrollment-free speaker cue. Adaptive cross-attention fusion combines AC and BC features, while frame-level BC representations learned with speaker supervision condition target-speech reconstruction. On speaker-disjoint mixtures constructed from ABCS and ESMB, FocusMe achieved the best average perceptual quality, intelligibility, and recognition accuracy among the evaluated systems. Ablations under severe interference showed that acoustic fusion and BC conditioning each contribute to performance, with further gains from explicit speaker supervision. These findings support jointly exploiting the acoustic and speaker-discriminative information in BC speech for target-speech recovery under strong noise and competing speech.

Limitations. Our evaluation used synthetic AC mixtures with synchronized BC recordings and no added BC corruption. The respective contributions of speaker identity and synchronized phonetic content in the BC representations remain unresolved. Future work will investigate this distinction and evaluate robustness to BC noise, temporal misalignment, contact variation, and device changes, alongside end-to-end streaming latency in real recording conditions.

AI USE STATEMENT

We used generative AI tools for language editing, terminology checks, and LaTeX formatting of the manuscript. The research method, experiments, and reported results were supplied by the authors, who take responsibility for the final text and scientific claims.

ETHICS STATEMENT

The experiments use previously released speech and noise corpora; no participants were recruited and no new voice recordings were collected. The paired AC and BC recordings contain speaker information, so we report quantitative results in aggregate and refer readers to the original dataset releases for access.

REPRODUCIBILITY STATEMENT

Section 3 specifies the model architecture and training objective, and Section 4 reports the implementation and evaluation settings. Appendix A documents the source corpora, speaker partitions, and mixture construction; Appendices B and C describe the baseline configurations and evaluation metrics. Results for each test condition are reported in Appendix D. The code and pre-trained models will be made publicly available upon acceptance of the paper.

REFERENCES

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. AISHELL-1: An open-source mandarin speech corpus and a speech recognition baseline. In *20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*, pp. 1–5, 2017.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (ELUs). In *International Conference on Learning Representations*, 2016.

David Dean, Sridha Sridharan, Robert Vogt, and Michael Mason. The QUT-NOISE-TIMIT corpus for the evaluation of voice activity detection algorithms. In *Proc. Interspeech*, pp. 3110–3113, 2010.

Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, and Haizhou Li. SpEx+: A complete time domain speaker extraction network. In *Proc. Interspeech*, pp. 1406–1410, 2020.

Julien Hauret, Thomas Joubaud, Véronique Zimpfer, and Éric Bavu. EBEN: Extreme bandwidth extension network applied to speech signals captured with noise-resilient body-conduction microphones. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2023.

Lixing He, Haozheng Hou, Shuyao Shi, Xian Shuai, and Zhenyu Yan. Towards bone-conducted vibration speech enhancement on head-mounted wearables. In *Proceedings of the 21st Annual International Conference on Mobile Systems, Applications and Services*, pp. 14–27, 2023.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 448–456. PMLR, 2015.

Jesper Jensen and Cees H. Taal. An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):2009–2022, 2016.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. SDR – half-baked or well done? In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 626–630, 2019.

Changtao Li, Feiran Yang, and Jun Yang. Bone conduction-aided speech enhancement with two-tower network and contrastive learning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:163–174, 2025a.

Kai Li, Fenghua Xie, Hang Chen, Kexin Yuan, and Xiaolin Hu. An audio-visual speech separation model inspired by cortico-thalamo-cortical circuits. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(10):6637–6651, 2024a. doi: 10.1109/TPAMI.2024.3384034.

Kai Li, Runxuan Yang, Fuchun Sun, and Xiaolin Hu. IIANet: An intra- and inter-modality attention network for audio-visual speech separation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 29181–29200. PMLR, 2024b.

Kai Li, Guo Chen, Wendi Sang, Yi Luo, Zhuo Chen, Shuai Wang, Shulin He, Zhong-Qiu Wang, Andong Li, Zhiyong Wu, and Xiaolin Hu. Advances in speech separation: Techniques, challenges, and future trends, 2025b.

Kai Li, Jintao Cheng, Chang Zeng, Zijun Yan, Helin Wang, Zixiong Su, Bo Zheng, and Xiaolin Hu. A semantically consistent dataset for data-efficient query-based universal sound separation, 2026a.

Kai Li, Kejun Gao, and Xiaolin Hu. Efficient audio-visual speech separation with discrete lip semantics and multi-scale global-local attention. In *International Conference on Learning Representations*, 2026b.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023.

Chandan K. A. Reddy, Vishak Gopal, Ross Cutler, Ebrahim Beyrami, Roger Cheng, Harishchandra Dubey, Sergiy Matusevych, Robert Aichner, Ashkan Aazami, Sebastian Braun, Puneet Rana, Sriram Srinivasan, and Johannes Gehrke. The INTERSPEECH 2020 deep noise suppression challenge: Datasets, subjective testing framework, and challenge results. In *Proc. Interspeech*, pp. 2492–2496, 2020.

Chandan K. A. Reddy, Vishak Gopal, and Ross Cutler. DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6493–6497, 2021.

Nicolae Catalin Ristea, Evgenii Indenbom, Ando Saabas, Tanel Pärnamaa, Jegor Guzhvin, and Ross Cutler. DeepVQE: Real time deep voice quality enhancement for joint acoustic echo cancellation, noise suppression and dereverberation. In *Proc. Interspeech*, pp. 3819–3823, 2023.

Antony W. Rix, John G. Beerends, Michael P. Hollier, and Andries P. Hekstra. Perceptual evaluation of speech quality (PESQ)—a new method for speech quality assessment of telephone networks and codecs. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. 749–752, 2001.

Hiroshi Sato, Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Takafumi Moriya, Naoki Makishima, Mana Ihori, Tomohiro Tanaka, and Ryo Masumura. Strategies to improve robustness of target speech extraction to enrollment variations. In *Proc. Interspeech*, pp. 996–1000, 2022.

Pengjie Shen, Kangrui Chen, Shulin He, Pengru Chen, Shuqi Yuan, He Kong, Xueliang Zhang, and Zhong-Qiu Wang. Listen to extract: Onset-prompted target speaker extraction. *IEEE Transactions on Audio, Speech and Language Processing*, 33:4832–4843, 2025.

Wenzhe Shi, Jose Caballero, Ferenc Huszar, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1874–1883, 2016.

Cees H. Taal, Richard C. Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(7):2125–2136, 2011.

Joachim Thiemann, Nobutaka Ito, and Emmanuel Vincent. DEMAND: A collection of multi-channel recordings of acoustic noise in diverse environments. Zenodo, 2013.

Bandhav Veluri, Malek Itani, Tuochao Chen, Takuya Yoshioka, and Shyamnath Gollakota. Look once to hear: Target speech hearing with noisy examples. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 1–16, 2024.

Heming Wang, Xueliang Zhang, and DeLiang Wang. Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:3134–3143, 2022a.

Mou Wang, Junqi Chen, Xiaolei Zhang, Zhiyong Huang, and Susanto Rahardja. Multi-modal speech enhancement with bone-conducted speech in time domain. *Applied Acoustics*, 200:109058, 2022b.

Mou Wang, Junqi Chen, Xiao-Lei Zhang, and Susanto Rahardja. End-to-end multi-modal speech recognition on an air and bone conducted speech corpus. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:513–524, 2023.

Quan Wang, Hannah Muckenhirn, Kevin Wilson, Prashant Sridhar, Zelin Wu, John R. Hershey, Rif A. Saurous, Ron J. Weiss, Ye Jia, and Ignacio Lopez Moreno. VoiceFilter: Targeted voice separation by speaker-conditioned spectrogram masking. In *Proc. Interspeech*, pp. 2728–2732, 2019.

Yilei Wu, Changyan Zheng, Xingyu Zhang, Yakun Zhang, Chengshi Zheng, Shuang Yang, Ye Yan, and Erwei Yin. DBMIF: A deep balanced multimodal iterative fusion framework for air- and bone-conduction speech enhancement. *Applied Intelligence*, 56(5):165, 2026.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. CSTR VCTK corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92). University of Edinburgh, The Centre for Speech Technology Research, 2019.

Cheng Yu, Kuo-Hsuan Hung, Syu-Siang Wang, Yu Tsao, and Jeih-weih Hung. Time-domain multi-modal bone/air conducted speech enhancement. *IEEE Signal Processing Letters*, 27:1035–1039, 2020.

C. Zhang and Philip C. Woodland. Parameterised sigmoid and ReLU hidden activation functions for DNN acoustic modelling. In *Proc. Interspeech*, pp. 3224–3228, 2015.

Chen Zhang, Linfeng Feng, Zhi Liu, Xiao-Lei Zhang, and Xuelong Li. MBENet: Bone-conduction and air-conduction fusion network for target speaker extraction. In *IEEE Automatic Speech Recognition and Understanding Workshop*, pp. 1–5, 2025.

Kateřina Žmolíková, Marc Delcroix, Keisuke Kinoshita, Takuya Higuchi, Atsunori Ogawa, and Tomohiro Nakatani. Speaker-aware neural network based beamformer for speaker extraction in speech mixtures. In *Proc. Interspeech*, pp. 2655–2659, 2017.

Kateřina Žmolíková, Marc Delcroix, Tsubasa Ochiai, Keisuke Kinoshita, Jan Černocký, and Dong Yu. Neural target speech extraction: An overview. *IEEE Signal Processing Magazine*, 40(3): 8–29, 2023.

A Datasets and Mixture Construction

Source corpora. The Air- and Bone-Conducted synchronized Speech (ABCS) corpus (Wang et al., 2023)[3] contains 47,182 paired AC and BC utterances from 100 speakers, totaling approximately 42 h. Participants wore a headset and read text prompts in an anechoic chamber. Each speaker contributed approximately 25 min of speech, and individual utterances lasted 1–5 s. The synchronized recordings provide two observations of the same utterance, allowing the AC recording to serve as the reconstruction reference and the BC recording as the auxiliary input.

[3] ABCS corpus: `https://github.com/wangmou21/abcs`.

The Elevoc Simultaneously-recorded Microphone/Bone-sensor (ESMB) corpus (Wang et al., 2022a)[4] contains 128 h of Chinese read speech from 287 speakers (131 male and 156 female). Speech was captured with a pair of Elevoc Clear earbuds. Each earbud included an ST25ba BC sensor near the ear canal to capture articulation-induced vibrations and an external AC sensor operating as a close-talk microphone. Participants read Chinese prompts for approximately 20 min each. The released recordings were sampled at 16 kHz, with stereo channels corresponding to the two earbuds. In our experiments, the target speech and its BC conditioning signal were taken from the same synchronized recording.

Noise and interfering-speech sources. Environmental noise for training and validation was drawn from the DNS Challenge (Reddy et al., 2020) and DEMAND (Thiemann et al., 2013), with disjoint noise-file sets used for the two splits. Evaluation noise was drawn from QUT-NOISE (Dean et al., 2010). Competing speech was drawn from AISHELL-1 (Bu et al., 2017) and VCTK (Yamagishi et al., 2019), providing Mandarin and English interference, respectively, with no overlap in interfering-speaker identities across the training, validation, and evaluation splits. Within each split, the available noise files were pooled across noise datasets, and the available interfering-speech files were pooled across speech datasets. Files were sampled uniformly from each pool. These recordings supplied the additive interference in the AC mixtures, and the paired ABCS and ESMB recordings supplied the target speaker.

Speaker partitions. We pooled the target speakers from ABCS and ESMB and randomly partitioned them into training, validation, and test sets using a random seed of 2666. The resulting splits contained 310, 38, and 38 speakers, respectively, corresponding to an approximately $8:1:1$ ratio. All utterances and both recording modalities associated with each target speaker were assigned to the same split, ensuring that no target speaker appeared in more than one split. FocusMe and all baselines used these same partitions.

Waveform preparation. All waveforms were processed at 16 kHz. Training examples consisted of 3 s segments, corresponding to $48{,}000$ samples. Paired AC and BC recordings were required to have equal lengths after loading and were subjected to identical cropping and padding operations. For utterances longer than 3 s, a contiguous segment was extracted with its starting sample selected uniformly from all valid positions. Utterances shorter than 3 s were zero-padded to the required length, with the left-padding length selected uniformly from zero to the total padding length and the remaining padding applied on the right. Utterances of exactly 3 s were retained unchanged. Noise and interfering-speech recordings were independently cropped or zero-padded using the same length-adjustment procedure. No additional intermodal delay estimation or compensation was performed by the data loader. During validation and inference, each supplied audio file was loaded at its full duration without additional cropping.

Mixture construction and level calibration. For an example containing L audio samples, let $\mathbf{s}, \mathbf{x}_{\mathrm{bc}}, \mathbf{n}_0, \mathbf{i}_0 \in \mathbb{R}^L$ denote the clean AC target, its synchronized BC recording, an unscaled noise segment, and unscaled competing speech, respectively. The AC mixture $\mathbf{x}_{\mathrm{ac}} \in \mathbb{R}^L$ is expressed as

$$\mathbf{x}_{\mathrm{ac}} = \mathbf{s} + \alpha_n \mathbf{n}_0 + \alpha_i \mathbf{i}_0, \tag{23}$$

where $\alpha_n, \alpha_i \in \mathbb{R}_{>0}$ control the noise and interference levels. Let $\mathcal{P} : \mathbb{R}^L \to \mathbb{R}_{\geq 0}$ denote the signal-power estimator, and let $\rho_n, \rho_i \in \mathbb{R}$ denote the desired SNR and SIR in decibels. For nonzero source powers, the corresponding amplitude gains are

$$\alpha_n = \sqrt{\frac{\mathcal{P}(\mathbf{s})}{\mathcal{P}(\mathbf{n}_0)}}\, 10^{-\rho_n/20}, \qquad \alpha_i = \sqrt{\frac{\mathcal{P}(\mathbf{s})}{\mathcal{P}(\mathbf{i}_0)}}\, 10^{-\rho_i/20}. \tag{24}$$

Both ratios are defined relative to the clean target speech. Each model received the pair $(\mathbf{x}_{\mathrm{ac}}, \mathbf{x}_{\mathrm{bc}})$ and was trained to reconstruct $\mathbf{s}$. No environmental noise or competing speech was added to the BC recording.

The data-generation ranges were -15 to 20 dB for SNR and -10 to 20 dB for SIR. Signal power was estimated as the mean squared amplitude over the entire segment, including silent and zero-padded samples, without voice-activity-based exclusion. Before mixing, the target speech, noise, and interfering speech were peak-normalized and adjusted to a common RMS level. Noise and

[4]ESMB corpus: `https://github.com/elevoctech/ESMB-corpus`.

interfering speech were then scaled separately relative to the target speech to obtain the sampled SNR and SIR, respectively. SNR and SIR values were drawn independently and uniformly from their respective configured candidate lists.

For each training example, one noise file and one interfering-speech file were sampled uniformly from their respective file lists. Each selected recording was independently adjusted to the training-segment length: longer recordings were cropped using a uniformly sampled valid starting position, whereas shorter recordings were zero-padded with a uniformly sampled left-padding length. The resulting signals were added over the same segment, with their active overlap determined by the sampled crops and padding positions.

After mixing, the mixture RMS level was randomly selected in integer-dB steps from $L_0 - 10$ to $L_0 + 9$ dBFS, where L_0 denotes the configured nominal RMS level. The resulting gain was also applied to the AC target. If the mixture peak exceeded 0.999, both the mixture and the AC target were scaled down by a common factor to place the mixture peak just below 0.99. Training mixtures were generated online during data loading.

Test conditions and shared evaluation. The test set comprised $1,300$ mixtures across the 13 SNR–SIR combinations in Table 5, with 100 mixtures per combination. Six conditions varied environmental noise at a fixed SIR, and six varied competing speech at a fixed SNR. The final condition combined the lowest SNR and SIR. All systems were evaluated on the same test mixtures and clean AC references. The same 100 sets of source segments were reused across all 13 SNR–SIR conditions, with the SNR/SIR-dependent gains adjusted for each condition.

Table 5: Test configuration for evaluating robustness to environmental noise and competing speech. The 13 SNR–SIR pairs comprised six noise conditions, six competing-speech conditions, and one joint condition. Each pair contained 100 mixtures, yielding 1,300 test mixtures shared by all models.

Condition group	SNR (dB)	SIR (dB)	Mixtures
Environmental noise	$\{-15, -10, -5, 0, 5, 15\}$	10	600
Competing speech	10	$\{-10, -5, 0, 5, 10, 15\}$	600
Joint interference	-15	-10	100

B Baseline Models

We retrained all five baselines on the data splits in Appendix A, following the model architectures and training settings reported in their original papers.

DBMIF. The Deep Balanced Multimodal Iterative Fusion framework (Wu et al., 2026)[5] enhances speech through separate AC and BC branches and a central fusion branch. A filter bank first decomposes the input waveforms into subbands, which are processed by a multiscale encoder–decoder. Iterative attention combines the modality features, while cross-branch gates allow the fusion stream and the two sensor streams to modulate one another. A balanced-interaction bottleneck further refines the joint representation. The estimated subbands are then synthesized into the enhanced waveform.

FCN. The fully convolutional network (Yu et al., 2020)[6] maps paired AC and BC waveforms to clean AC speech in the time domain. Its early-fusion variant concatenates the two waveforms at the input and processes them with a shared convolutional network. The late-fusion variant first performs AC enhancement and BC-to-AC conversion in separate networks, then concatenates their outputs and applies an additional convolutional fusion network. The variants therefore differ in whether sensor information is combined before or after modality-specific reconstruction. We used the early fusion configuration.

MMINet. MMINet (Wang et al., 2022b) uses synchronized AC and BC speech in an end-to-end encoder–decoder architecture. The encoder converts the waveforms into a learned representation, and an involution network uses information from both modalities to estimate a mask for the clean

[5]Code: `https://github.com/wyl516w/dbmif`.

[6]Code: `https://github.com/WilliamYu1993/BAMSE`.

speech component. Applying this mask suppresses unwanted components in the encoded features. The decoder then maps the masked representation back to an enhanced AC waveform.

DCCRN. DCCRN denotes the AC-BC sensor-fusion system of Wang et al. (2022a) in our comparison. The system operates on complex spectra and extracts features from the two sensors before fusion. Its attention module combines local and global channel context to determine the contributions of AC and BC features. The fused representation is mapped to the real and imaginary components of the clean AC spectrum, allowing the reconstructed speech to use an estimated phase as well as an estimated magnitude.

VibVoice. VibVoice (He et al., 2023)[7] combines microphone audio with bone-conducted vibrations captured on head-mounted wearables. Separate encoders extract features from the audio and vibration spectrograms. A fusion decoder combines these features and predicts a magnitude mask for the noisy audio spectrum. During training, an auxiliary decoder predicts the low-frequency clean speech spectrum from the vibration features alone, providing supervision for the vibration branch. The masked audio spectrum is used for waveform reconstruction.

C Evaluation Metrics

Let $\mathbf{s}, \widehat{\mathbf{s}} \in \mathbb{R}^L$ denote the clean target and enhanced waveforms, respectively, where L is the number of audio samples. We evaluated waveform fidelity, perceptual quality, intelligibility, and recognition accuracy using the metrics below.

SI-SDR. Scale-invariant signal-to-distortion ratio (Le Roux et al., 2019) evaluates waveform reconstruction after compensating for a global gain difference. With the optimal projection gain $\alpha_\star \in \mathbb{R}$, the score is

$$\alpha_\star = \frac{\widehat{\mathbf{s}}^\top \mathbf{s}}{\|\mathbf{s}\|_2^2}, \qquad \mathrm{SI{-}SDR} = 10 \log_{10} \frac{\|\alpha_\star \mathbf{s}\|_2^2}{\|\widehat{\mathbf{s}} - \alpha_\star \mathbf{s}\|_2^2}. \tag{25}$$

The numerator and denominator measure projected target energy and residual error energy. Higher SI-SDR (in decibels) indicates better waveform fidelity.

PESQ. Perceptual evaluation of speech quality (Rix et al., 2001) estimates perceived quality from the enhanced signal and a clean reference. The algorithm aligns the signals, transforms them into auditory representations, and aggregates perceptual disturbances across time and frequency. Its symmetric and asymmetric disturbance terms account for differences in the perceptual effects of signal distortion. Higher scores indicate better predicted quality. We used the Python `pesq` package in wideband mode (ITU-T P.862.2) at a sampling rate of 16 kHz.

STOI. Short-time objective intelligibility (Taal et al., 2011) compares temporal envelopes in one-third-octave bands after removing silent intervals. Let $J, Q, U \in \mathbb{N}_{>0}$ denote the number of bands, the segment length in frames, and the number of segments, respectively. For band $j \in \{1, \ldots, J\}$ and segment $m \in \{1, \ldots, U\}$, the reference and enhanced envelopes are $\mathbf{a}_{j,m}, \mathbf{b}_{j,m} \in \mathbb{R}_+^Q$. With a clipping threshold $\beta \in \mathbb{R}$ in decibels, the enhanced envelope is normalized and clipped to give $\widetilde{\mathbf{b}}_{j,m} \in \mathbb{R}^Q$:

$$\widetilde{\mathbf{b}}_{j,m} = \min\left(\frac{\|\mathbf{a}_{j,m}\|_2}{\|\mathbf{b}_{j,m}\|_2} \mathbf{b}_{j,m}, (1 + 10^{-\beta/20}) \mathbf{a}_{j,m} \right), \tag{26}$$

where the minimum is taken elementwise. Let $\mathbf{a}^\circ_{j,m}, \mathbf{b}^\circ_{j,m} \in \mathbb{R}^Q$ be the mean-centered versions of $\mathbf{a}_{j,m}$ and $\widetilde{\mathbf{b}}_{j,m}$. STOI averages their normalized correlations:

$$\mathrm{STOI} = \frac{1}{JU} \sum_{j=1}^{J} \sum_{m=1}^{U} \frac{(\mathbf{a}^\circ_{j,m})^\top \mathbf{b}^\circ_{j,m}}{\|\mathbf{a}^\circ_{j,m}\|_2 \|\mathbf{b}^\circ_{j,m}\|_2}. \tag{27}$$

Higher values indicate better preservation of the temporal envelopes associated with intelligibility.

ESTOI. Extended STOI (Jensen & Taal, 2016) incorporates correlations across frequency bands to account for temporally modulated interference. For each segment m, let $\mathbf{A}_m, \mathbf{B}_m \in \mathbb{R}^{J \times Q}$ collect the reference and enhanced envelopes. Each row is mean-centered and normalized to unit

[7] Code: `https://github.com/CUHK-AIoT-Sensing/vibvoice`.

Euclidean norm, followed by the same operations on each column. Denoting the resulting matrices by $\overline{\mathbf{A}}_m, \overline{\mathbf{B}}_m \in \mathbb{R}^{J\times Q}$, the score is

$$\mathrm{ESTOI} = \frac{1}{UQ} \sum_{m=1}^{U} \left\langle \overline{\mathbf{A}}_m, \overline{\mathbf{B}}_m \right\rangle_F , \tag{28}$$

where $\langle \cdot, \cdot \rangle_F$ denotes the sum of elementwise products. The score averages the similarity of the normalized spectral patterns over all segments; higher values indicate better predicted intelligibility.

We computed STOI and extended STOI (ESTOI) scores using the Python `pystoi` package, supplying the reference and enhanced signals at 16 kHz.

DNSMOS. DNSMOS (Reddy et al., 2021) is a learned predictor of human speech-quality ratings. It estimates quality from spectral features of the enhanced speech and requires no clean reference. The original predictor was trained with human mean opinion scores and a multistage self-teaching procedure. We used the DNSMOS P.808 model checkpoint `model_v8.onnx` and reported its predicted P.808 mean opinion score (MOS), with higher values indicating better predicted perceptual quality.

CER. We measured recognition accuracy by transcribing the enhanced speech with Whisper (Radford et al., 2023). Let $n_s, n_d, n_i \in \mathbb{N}_0$ denote the numbers of character substitutions, deletions, and insertions in the minimum-edit alignment with the reference transcript, and let $n_r \in \mathbb{N}_{>0}$ be the number of reference characters. Character error rate is

$$\mathrm{CER} = \frac{n_s + n_d + n_i}{n_r} \times 100\%. \tag{29}$$

Lower CER indicates fewer transcription errors relative to the reference transcript. We used OpenAI Whisper large with the language set to Chinese and the default transcription and decoding settings. Both reference and hypothesis transcripts were converted to Simplified Chinese using OpenCC, followed by Unicode NFKC normalization, lowercasing, and removal of characters in Unicode categories P (punctuation), Z (separators), and C (other, including control and formatting characters). Character substitutions, deletions, and insertions were obtained through utterance-level Levenshtein alignment.

Aggregation across test conditions. The means in Table 1 were computed from the condition-level scores in Tables 6, 7, and 2. Each of the 13 conditions contributed equally: six varied the SNR, six varied the SIR, and one combined the lowest SNR and SIR. Mean CER was the arithmetic mean of the 13 condition-level error rates.

Efficiency. Parameter counts are reported in millions, MACs in billions, and inference time in milliseconds for the input duration specified in Section 4.3. Inference runtime was evaluated on an Intel Xeon Gold 5320 CPU (2.20 GHz) and a single NVIDIA GeForce RTX 4090 GPU using PyTorch 2.2.2 with CUDA 12.1 and cuDNN 8.9.2. The model was executed in eager evaluation mode under `torch.inference_mode()`, using FP32 precision with automatic mixed precision and TF32 disabled. The inference batch size was 1. Each input consisted of a 3-s air-conduction/bone-conduction waveform pair sampled at 16 kHz. PyTorch intra-op and inter-op thread counts were both set to 1, and the OpenMP and numerical-library thread limits were also set to 1. cuDNN autotuning was disabled. For each device and timing scope, 200 warm-up iterations were followed by 500 timed iterations on the same prerecorded segment. Wall-clock time was measured using `time.perf_counter_ns()`, with explicit CUDA synchronization immediately before and after each timed GPU call.

D Results under Individual Test Conditions

Tables 6 and 7 report the results for the SNR and SIR sweeps, respectively. Together with the results under joint noise and interference in Table 2, these scores were used to compute the means in Table 1. Each SNR–SIR condition comprised 100 test mixtures, and all models were evaluated on the same mixtures.

Table 6: Performance under varying environmental noise, with SIR fixed at 10 dB and SNR ranging from −15 to 15 dB. Each SNR condition contained 100 test mixtures shared by all models. Within each condition, bold and underlined values indicate the best and second-best results for each metric, respectively; arrows indicate the preferred direction.

SNR (dB)	Model	SI-SDR (dB) ↑	PESQ ↑	STOI ↑	ESTOI ↑	DNSMOS ↑	CER (%) ↓
−15	FCN	-3.05	1.08	0.56	0.30	2.26	96.52
	DCCRN	0.60	1.57	0.73	0.60	2.56	55.48
	MMINet	**5.35**	1.39	0.72	0.57	2.52	65.36
	VibVoice	-1.05	1.34	0.65	0.50	2.45	84.73
	DBMIF	-0.44	1.19	0.63	0.48	2.31	82.83
	FocusMe	2.40	**1.66**	**0.77**	**0.65**	**2.67**	**41.57**
−10	FCN	0.66	1.12	0.61	0.37	2.29	82.79
	DCCRN	4.56	1.79	0.78	0.68	2.79	38.83
	MMINet	**7.95**	1.48	0.76	0.61	2.63	47.36
	VibVoice	3.32	1.53	0.72	0.59	2.56	59.94
	DBMIF	3.41	1.33	0.70	0.56	2.37	67.03
	FocusMe	6.14	**1.90**	**0.82**	**0.72**	**2.89**	**30.52**
−5	FCN	3.69	1.16	0.66	0.44	2.29	70.01
	DCCRN	7.79	2.03	0.83	0.73	2.90	29.02
	MMINet	**9.63**	1.60	0.79	0.67	2.68	38.56
	VibVoice	6.66	1.72	0.78	0.66	2.67	48.18
	DBMIF	6.35	1.47	0.76	0.63	2.44	46.35
	FocusMe	9.22	**2.19**	**0.86**	**0.77**	**2.98**	**21.50**
0	FCN	6.61	1.23	0.72	0.52	2.35	43.44
	DCCRN	10.71	2.26	0.87	0.79	3.11	19.29
	MMINet	11.62	1.71	0.83	0.72	2.76	25.93
	VibVoice	9.70	1.92	0.83	0.73	2.94	28.44
	DBMIF	9.08	1.62	0.81	0.71	2.60	30.96
	FocusMe	**12.02**	**2.44**	**0.90**	**0.82**	**3.17**	**15.07**
5	FCN	8.80	1.30	0.75	0.59	2.43	23.06
	DCCRN	13.00	2.50	0.90	0.84	3.28	12.45
	MMINet	13.30	1.87	0.87	0.74	2.86	16.01
	VibVoice	12.03	2.14	0.87	0.79	3.12	18.03
	DBMIF	11.14	1.74	0.85	0.76	2.72	19.16
	FocusMe	**14.28**	**2.76**	**0.92**	**0.86**	**3.32**	**9.38**
15	FCN	11.61	1.53	0.82	0.71	2.74	12.03
	DCCRN	16.25	2.94	0.94	0.90	3.48	6.39
	MMINet	16.03	2.19	0.93	0.86	3.10	6.47
	VibVoice	15.07	2.51	0.93	0.88	3.34	8.22
	DBMIF	14.12	2.03	0.91	0.84	2.96	8.96
	FocusMe	**17.62**	**3.18**	**0.96**	**0.92**	**3.49**	**4.65**

D.1 Robustness to Environmental Noise

At a fixed SIR of 10 dB, FocusMe achieved the highest PESQ, STOI, ESTOI, and DNSMOS scores and the lowest CER across all six SNRs from −15 to 15 dB (Table 6). At −15 dB SNR, it improved PESQ from 1.57 to 1.66 and reduced CER from 55.48% to 41.57% (13.91 percentage points) compared with DCCRN, the strongest baseline on both metrics. These results showed improved intelligibility and recognition under strong environmental noise.

D.2 Robustness to Competing Speech

At a fixed SNR of 10 dB, FocusMe again achieved the best PESQ, STOI, ESTOI, DNSMOS, and CER across all six SIRs from −10 to 15 dB (Table 7). At −10 dB SIR, it reduced CER from 20.59% for DCCRN to 14.03%, a reduction of 6.56 percentage points. As the SIR increased to

Table 7: Performance under varying competing speech, with SNR fixed at 10 dB and SIR ranging from -10 to 15 dB. Each SIR condition contained 100 test mixtures shared by all models. Within each condition, bold and underlined values indicate the best and second-best results for each metric, respectively; arrows indicate the preferred direction.

SIR (dB)	Model	SI-SDR (dB)↑	PESQ↑	STOI↑	ESTOI↑	DNSMOS↑	CER (%)↓
-10	FCN	-0.80	1.15	0.61	0.45	2.71	77.22
	DCCRN	7.46	2.12	0.85	0.78	3.12	20.59
	MMINet	**10.49**	1.65	0.82	0.72	2.81	23.55
	VibVoice	5.47	1.79	0.80	0.72	2.91	30.62
	DBMIF	3.42	1.34	0.73	0.64	2.60	44.92
	FocusMe	9.42	**2.27**	**0.87**	**0.81**	**3.18**	**14.03**
-5	FCN	2.49	1.20	0.67	0.52	2.71	59.08
	DCCRN	10.00	2.35	0.87	0.81	3.24	16.11
	MMINet	**11.98**	1.76	0.84	0.74	2.86	32.07
	VibVoice	8.35	1.98	0.84	0.77	3.04	24.49
	DBMIF	6.44	1.52	0.78	0.69	2.71	33.10
	FocusMe	11.75	**2.52**	**0.90**	**0.84**	**3.30**	**10.58**
0	FCN	5.96	1.25	0.72	0.56	2.62	40.56
	DCCRN	11.93	2.51	0.89	0.83	3.26	11.61
	MMINet	12.88	1.84	0.86	0.77	2.85	13.39
	VibVoice	10.51	2.15	0.86	0.79	3.08	19.84
	DBMIF	9.05	1.65	0.82	0.73	2.71	21.13
	FocusMe	**13.51**	**2.72**	**0.92**	**0.86**	**3.28**	**9.03**
5	FCN	8.48	1.33	0.76	0.61	2.60	25.69
	DCCRN	13.31	2.56	0.91	0.85	3.34	11.02
	MMINet	13.88	1.93	0.88	0.80	2.93	13.53
	VibVoice	12.17	2.22	0.88	0.81	3.20	13.70
	DBMIF	11.03	1.77	0.86	0.77	2.82	15.15
	FocusMe	**14.82**	**2.84**	**0.93**	**0.88**	**3.38**	**8.27**
10	FCN	10.61	1.41	0.78	0.65	2.60	13.43
	DCCRN	15.02	2.75	0.92	0.87	3.41	7.52
	MMINet	14.95	2.03	0.90	0.83	2.98	9.38
	VibVoice	14.02	2.35	0.90	0.84	3.26	11.24
	DBMIF	13.00	1.87	0.88	0.81	2.84	10.67
	FocusMe	**16.30**	**3.01**	**0.94**	**0.89**	**3.43**	**6.47**
15	FCN	12.00	1.51	0.83	0.71	2.65	11.57
	DCCRN	16.68	2.87	0.94	0.89	3.44	6.50
	MMINet	16.00	2.09	0.91	0.85	3.03	7.15
	VibVoice	15.64	2.50	0.92	0.87	3.32	6.34
	DBMIF	14.91	2.03	0.91	0.85	2.89	8.03
	FocusMe	**17.71**	**3.13**	**0.95**	**0.91**	**3.46**	**4.82**

15 dB, FocusMe's CER decreased to 4.82% and its SI-SDR increased from 9.42 to 17.71 dB. These results demonstrated that the BC speaker cue helped preserve target-speech content under competing speech.

D.3 Spectrogram Comparison

We compared the enhanced spectra with the clean AC reference for the same test utterance under two interference conditions. All panels in both figures share the same STFT settings and the same magnitude scale, so the six enhanced outputs can be compared directly with the reference, the mixture, and the synchronized BC input. The two conditions isolate the two interference types examined in Tables 6 and 7.

In both interfering speech (Figure 5) and environmental noise scenarios (Figure 6), the AC mixture reveals that target speech is heavily masked by the interfering source and ambient noise. In contrast,

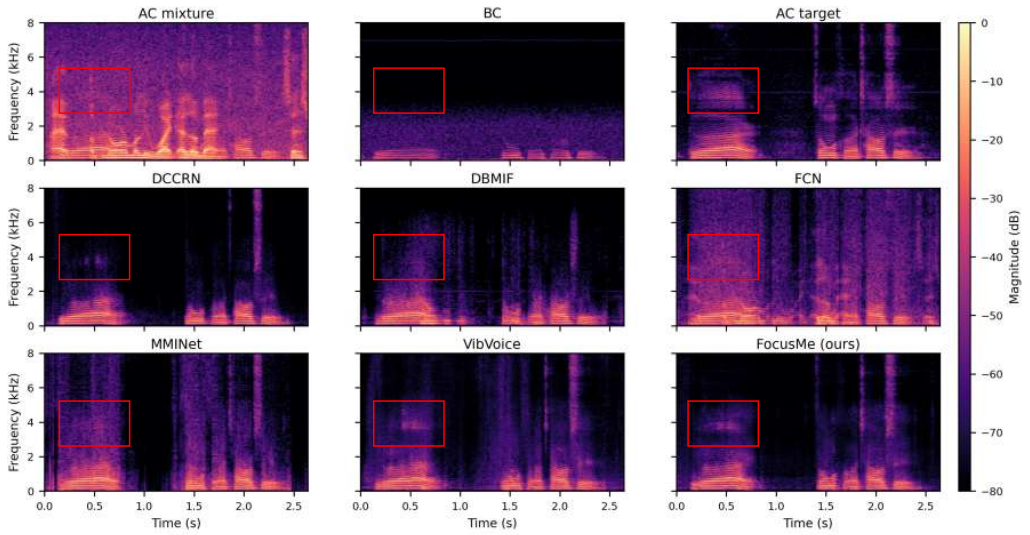


Figure 5: Enhanced spectra under interfering speech, with SNR fixed at 10 dB and SIR at -10 dB. The AC mixture and the AC target are shown together with the BC input, which contains no interfering speech and environmental noise, followed by the six enhanced outputs.

the BC signal concentrates its energy predominantly below 1 kHz—tracing the temporal envelope of the target speech—which indicates that BC recordings capture voiced target segments while offering negligible information at higher frequencies. FocusMe effectively retains distinct target harmonics while suppressing energy from both interference and noise. Conversely, DCCRN suffers from high-frequency harmonic attenuation above 2 kHz along with degraded fricatives, whereas other baseline models (DBMIF, FCN, MMINet, and VibVoice) excessively smooth spectral details, obscuring the underlying formant structures.

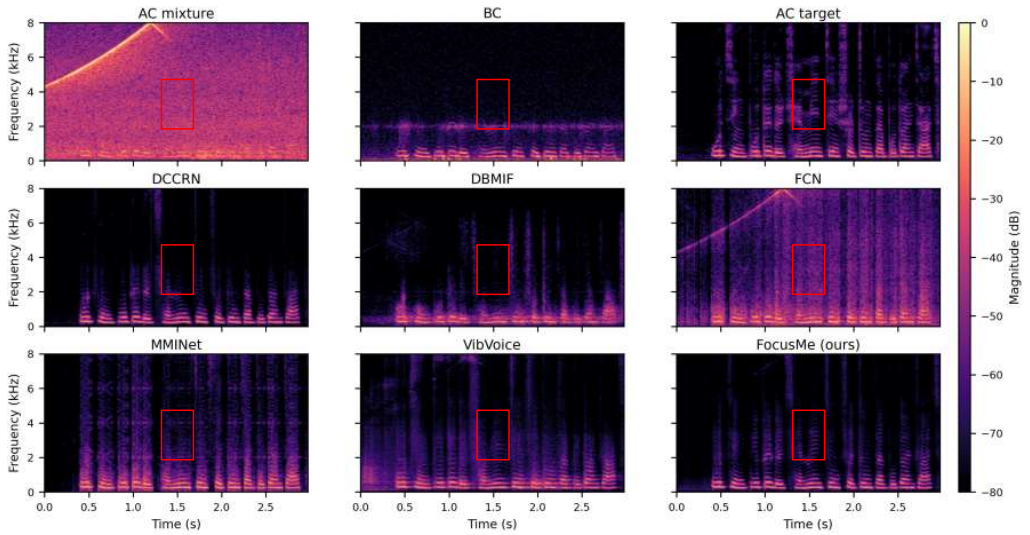


Figure 6: Enhanced spectra under environmental noise, with SNR fixed at -15 dB and SIR at 10 dB. The AC mixture and the AC target are shown together with the BC input, which contains no interfering speech and environmental noise, followed by the six enhanced outputs.