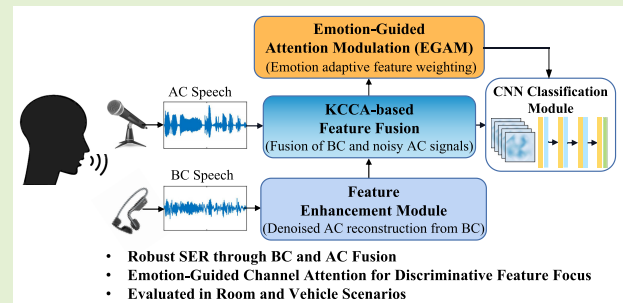


Multimodal Sensor Fusion of Bone- and Air-Conducted Speech for Robust Emotion Recognition

Shujie Zhao^{ID}, Lijun Zhang^{ID}, Israel Cohen^{ID}, *Fellow, IEEE*, and Yan Yang^{ID}

Abstract—Environmental noise poses a significant challenge to speech emotion recognition (SER) systems, as it distorts acoustic features and masks critical emotional cues. While traditional speech enhancement techniques aim to mitigate noise, they often introduce artifacts that compromise emotional information. To address this challenge, we propose a bimodal sensor fusion framework that integrates bone-conducted (BC) and air-conducted (AC) speech signals, leveraging their complementary strengths. BC speech exhibits robust noise immunity in low-frequency bands, while AC speech retains full-band spectral detail essential for emotion perception. The proposed framework includes two key modules: 1) a feature enhancement network that reconstructs clean AC features from BC inputs using hybrid loss functions and 2) a channel attention mechanism that dynamically prioritizes emotion-relevant channels in the fused BC–AC representation. Experiments conducted under diverse acoustic conditions, including simulated enclosed-room and in-vehicle environments, demonstrate that the proposed method outperforms conventional fusion approaches in recognition accuracy and noise robustness. These findings underscore the promise of emotion-aware multimodal feature fusion in developing robust SER systems suitable for real-world noisy environments, including smart vehicles, wearable devices, and industrial human–machine interfaces.

Index Terms—Adaptive signal processing, bone-conducted (BC) and air-conducted (AC) microphones, multimodal sensor fusion, robust speech emotion recognition (SER).



I. INTRODUCTION

SPEECH emotion recognition (SER) has emerged as a crucial technology in various real-world applications, including human–machine interaction (HMI), industrial safety monitoring, and intelligent transportation systems. In high-stakes environments, such as automated control centers, smart vehicles, and emergency response operations, accurate assessment of emotional states can enhance decision-making, improve user experience, and ensure operational safety. Tra-

ditional SER systems primarily rely on air-conducted (AC) speech signals, which are highly susceptible to environmental distortions such as noise, reverberation, and signal loss, posing significant challenges for accurately capturing emotional cues.

To address these limitations, researchers have explored two main approaches: advanced speech enhancement algorithms and alternative speech acquisition modalities. Advanced single-channel speech enhancement algorithms (e.g., CMGAN, CDiffuSE, LiSenNet, and FullSubNet [1], [2], [3], [4]) have shown effectiveness in mitigating distortions caused by noise and reverberation. However, these methods primarily optimize speech quality metrics such as perceptual evaluation of speech quality (PESQ) and short-time objective intelligibility (STOI), potentially suppressing critical acoustic cues essential for emotion classification. Consequently, their performance remains limited under severe nonstationary conditions, often failing to preserve subtle emotional cues essential for SER.

Recent SER advances have focused on emotion-aware architectures and multimodal fusion. Models like ABMD [5] and DBMER [6] demonstrate improved performance through

Received 27 May 2025; revised 23 June 2025; accepted 29 June 2025. Date of publication 10 July 2025; date of current version 15 August 2025. The work of Yan Yang was supported in part by the Natural Science Foundation of China (NSFC) under Grant 61771403 and Grant N2018KF0157. The associate editor coordinating the review of this article and approving it for publication was Dr. Liansheng Liu. (*Corresponding author: Lijun Zhang.*)

Shujie Zhao, Lijun Zhang, and Yan Yang are with the Center of Intelligent Acoustics and Immersive Communications, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China (e-mail: sjzhao@mail.nwpu.edu.cn; zhanglj7385@nwpu.edu.cn).

Israel Cohen is with the Technion–Israel Institute of Technology, Haifa 3200003, Israel (e-mail: icohen@ee.technion.ac.il).

Digital Object Identifier 10.1109/JSEN.2025.3585068

attention mechanisms and audio–text fusion. While these approaches have advanced emotion-specific representation learning, they primarily focus on single-modality enhancement or text-audio fusion, without exploring the complementary strengths of different speech acquisition modalities.

Among alternative speech acquisition modalities, bone-conducted (BC) speech has emerged as a particularly promising solution to these challenges. Transmitted directly via skull vibrations, BC speech inherently exhibits robustness against environmental interference, significantly reducing susceptibility to background noise and reverberation. This characteristic has been extensively exploited for robust speech enhancement and reliable communication under adverse acoustic conditions. Nevertheless, BC speech alone has limited bandwidth (approximately 0–1 kHz) and lacks the high-frequency information necessary for fine-grained emotion discrimination, making it insufficient for effective SER when used independently.

Recent commercial products have successfully integrated BC and AC speech signals to address these limitations. For example, Shokz’s latest open-ear headphones feature DualPitch¹ technology, combining BC and AC microphones with digital frequency division. The system includes a high-frequency unit with dual-voice coil and dual-spring diaphragm, paired with a low-frequency unit featuring an 18×11 mm ring diaphragm. This integration enhances speech clarity and reduces noise interference in challenging environments. Such commercial implementations demonstrate the practical viability of dual-modality approaches for enhanced speech recognition and emotion detection applications.

Consequently, the fusion of AC and BC speech signals offers a compelling approach to improving SER robustness and accuracy in challenging acoustic scenarios. While existing studies on AC–BC speech fusion (e.g., [7], [8], [9]) focus predominantly on enhancing speech intelligibility and robustness without explicitly addressing the preservation of emotional information, our work introduces an emotion-guided attention mechanism specifically designed to enhance SER accuracy. This novel approach selectively emphasizes emotion-relevant features within the fused AC–BC representations, rather than merely improving perceptual speech quality. Compared to traditional audio-visual emotion recognition, which can fail under dark, privacy-sensitive, or wearable-device contexts due to visual data unavailability, multimodal AC–BC fusion provides a reliable and privacy-preserving alternative. By effectively preserving critical emotional cues across the frequency spectrum, this fusion strategy expands the practical applicability of SER systems and ensures consistent performance in diverse, challenging acoustic environments.

II. BACKGROUND AND RELATED WORK

The integration of AC and BC modalities offers a promising multimodal approach by combining the noise robustness of BC speech [10], [11] with the spectral richness of AC speech [12], [13], [14], [15]. This complementary fusion

enables enhanced SER performance resilience across diverse acoustic environments. AC speech captures a full-bandwidth signal (0–8 kHz) with pronounced high-frequency content and rich temporal dynamics, while BC speech provides smoother, low-frequency-dominant representations with reduced interference [18], [19]. Notably, both share a similar low-frequency energy distribution, making them well-suited for complementary fusion [20].

Significant efforts have been made to address the limitations of BC speech using signal processing techniques such as filtering [18], [21], transfer function estimation [22], [23], and analysis-synthesis models [24]. More recently, deep learning approaches have significantly advanced this research domain. Neural architectures including multilayer perceptrons (MLPs) [25], deep autoencoders [26], bidirectional long short-term memory (BLSTM) networks [27], and CycleGANs [28] have shown improved performance in BC speech enhancement and reconstruction [29], [30], [31].

Despite these advancements, two key challenges persist. First, BC speech inherently lacks high-frequency information due to the physical characteristics of bone transmission, limiting its standalone applicability to SER. Second, recent BC-based SER studies, such as Hosain et al. [11], primarily rely on synthetic BC signals and evaluations in clean environments, raising concerns about generalization to real-world conditions and robustness under noise. These limitations highlight the need for a more robust, adaptable multimodal SER framework capable of handling environmental variability and acoustic complexity effectively.

To assess the effectiveness of BC–AC speech fusion, we consider two representative acoustic scenarios with distinct environmental challenges.

- 1) The first involves large enclosed environments (such as industrial sites, control rooms, conference rooms) characterized by long reverberation time (T_{60} up to 1.6 s) and multisource noise [32] (factory noise, speech babble, and white noise). AC microphones suffer significant degradation in such conditions, while BC-augmented approaches provide a more stable assessment of operator stress and cognitive load.
- 2) The second scenario models small enclosed environments (vehicle cabins, emergency response units) with short reverberation time ($T_{60} < 0.4$ s) but high-intensity noise. Small in-vehicle spaces were emulated using real-world impulse responses and noise sources from the ASR-CABNOIS dataset [33]. Two types of in-vehicle noise were selected: mechanical noise (including engine, wind, and road components) and speech-based background noise (in-car babble). In automotive HMI systems, robust SER contributes to driver state monitoring and safety, whereas in emergency operations, BC–AC fusion enables reliable emotional state under dynamic conditions. These scenarios reflect practical applications where conventional AC-based SER systems are often compromised, demonstrating the potential of multimodal speech fusion for improving robustness across diverse acoustic settings.

¹Trademark.

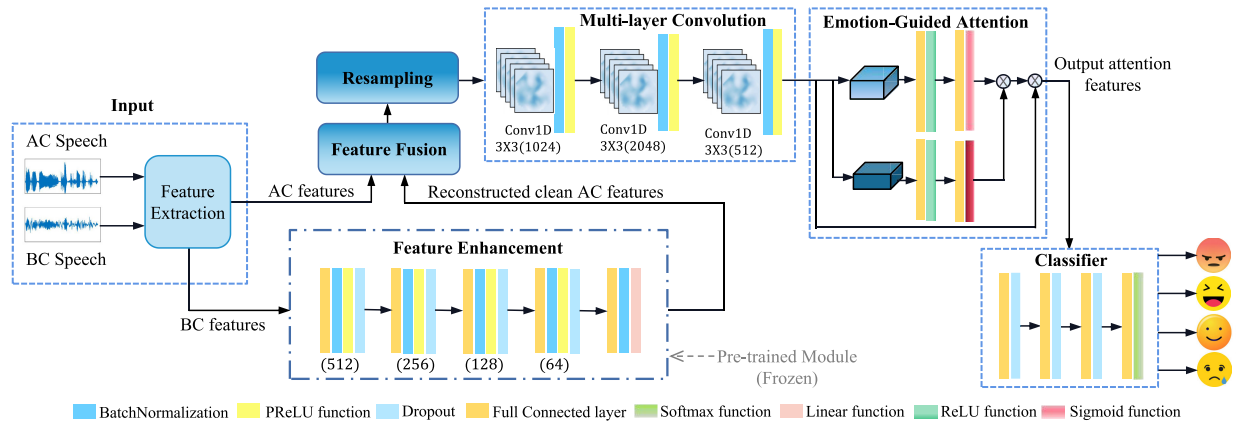


Fig. 1. Overview of the proposed multimodal SER framework, which integrates BC–AC feature extraction, reconstruction, fusion, attention-based enhancement, and final emotion classification.

To address the challenges of environmental noise and reverberation, this study proposes a novel multimodal sensor fusion approach with emotion-aware channel attention, leveraging the complementary strengths of BC and AC speech to enhance both emotional fidelity and noise robustness. The primary contributions are as follows.

- 1) *Feature Enhancement Module*: A fully connected neural network is trained to reconstruct clean AC features from BC inputs, using a hybrid loss function that combines linear reconstruction and correlation-based constraints to preserve emotion-related content while suppressing noise.
- 2) *KCCA-Based Feature Fusion*: A kernel canonical correlation analysis (KCCA) framework [34] is introduced to effectively combine enhanced AC features with BC features, capturing both shared and complementary information for emotion representation.
- 3) *Channel Attention Mechanism*: An emotion-aware attention module adaptively reweights fused feature channels based on their emotional relevance, enabling the model to highlight informative features while suppressing irrelevant noise components.

This represents, to our knowledge, the first SER study to explore BC–AC speech fusion under both high-reverberation and high-noise scenarios, highlighting its potential for sensor-driven, adaptive SER in practical environments.

III. PROPOSED METHOD

To address the limitations posed by noise and reverberation, we propose an attention-guided multimodal feature fusion network for robust SER, as illustrated in Fig. 1. The proposed SER framework comprises four key modules that address specific challenges in noise-robust emotion recognition. The feature enhancement module reconstructs clean AC features using BC inputs as guidance, recovering emotion-relevant content while mitigating noise. The KCCA-based fusion module integrates enhanced AC and BC features, capturing shared and complementary emotional information. The resampling module addresses class imbalance to achieve balanced training across emotion categories. Finally, the

emotion-aware attention module highlights emotion-relevant channels while suppressing noise-corrupted features, improving robustness and recognition performance. Detailed formulations and implementation of each module are presented in Sections III-A–III-D.

A. Feature Enhancement Module

Adverse acoustic conditions can significantly degrade the emotional content of speech signals. To address this, we introduce a feature enhancement module that reconstructs clean AC features using the more noise-resilient BC speech as a reference. Implemented as a fully connected neural network with progressive dimensionality reduction, the module learns complex nonlinear mappings between BC and clean AC features. A hybrid loss function combining reconstruction loss (e.g., mean squared error) and correlation-based constraints is used to ensure fidelity while preserving emotion-relevant characteristics. Unlike conventional denoising approaches that may distort affective information, this module leverages the noise robustness of BC speech to estimate clean, emotionally expressive AC features. The enhanced features are then passed to the fusion stage, providing more reliable emotional cues under adverse acoustic conditions.

In the preprocessing stage, dual-channel speech recordings are separated into BC and AC streams and processed at the sentence level. Each signal undergoes feature extraction to generate a comprehensive acoustic representation including Mel-frequency cepstral coefficients (MFCCs), spectral contrast, chroma, tonnetz, pitch, and statistical measures to capture both low-level and high-level speech characteristics [36]. Features like loudness and energy are augmented with statistical descriptors: mean, median, minimum, maximum, variance, skewness, kurtosis, percentiles (25th, 50th, and 75th), and interquartile range. These statistical measures enable detailed analysis of the feature distributions, providing a robust foundation for subsequent processing tasks. The complete feature list and dimensional breakdown are provided in Table I.

During the pretraining phase, the feature enhancement module is optimized using a multiobjective loss function that minimizes reconstruction error and maximizes linear and

TABLE I

SUMMARY OF EXTRACTED FEATURES AND THEIR DIMENSIONS

Feature Type	Description	Dimensions
MFCCs	MFCCs with deltas and delta-deltas	120
Chroma	Energy distribution across pitch classes	12
Mel-Spectrogram	Spectral power distribution across mel bands	128
Spectral Contrast	Difference between spectral peaks and valleys	7
Tonnetz	Tonal centroid features	6
Loudness	RMS energy with statistical descriptors	11
Energy	Short-term energy with statistical descriptors	11

nonlinear correlations between the reconstructed and target AC features. This ensures that the enhanced outputs maintain both numerical accuracy and the structural/emotional characteristics necessary for robust SER. The mean absolute error (MAE) is employed to quantify reconstruction accuracy due to its robustness to outliers and ability to capture absolute deviations. In addition to the reconstruction loss \mathcal{L}_r , two correlation-based penalties are introduced: a linear correlation loss \mathcal{L}_p based on the Pearson correlation coefficient (PCC) and a nonlinear loss \mathcal{L}_k derived from KCCA. These terms encourage the network to preserve statistical relationships and latent structures between the BC and AC feature representations. The overall loss function $\mathcal{L}_{\text{total}}$ is defined as a weighted sum of these three terms, guiding the model to generate features that are both accurate and preserve emotional discriminability, even under adverse acoustic conditions

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \mathcal{L}_r + w_1 \mathcal{L}_p + w_2 \mathcal{L}_k \\ &= \frac{1}{N} \sum_{i=1}^N \|\mathbf{x}_i - \hat{\mathbf{x}}_i\|_{L_1} + w_1 (1 - |r_p|) + w_2 (1 - |r_k|) \end{aligned} \quad (1)$$

where N is the number of samples in each training batch, $\mathbf{x}_i \in \mathbb{R}^{295}$ denotes the 295-D feature vector of the clean AC speech for the i th sample, and $\hat{\mathbf{x}}_i \in \mathbb{R}^{295}$ is the corresponding reconstructed feature vector obtained from BC speech. Here, r_p and r_k represent PCC and the KCCA coefficient, respectively, quantifying the linear and nonlinear relationships between the true and reconstructed features. w_1 and w_2 are tunable hyperparameters that adjust the relative weights of the linear and nonlinear correlation penalties.

B. Feature Fusion Module

Following the feature enhancement stage, the reconstructed AC features are fused with the original noisy AC features to leverage complementary information and prior signal context, enhancing robustness against environmental noise. This strategy retains the structure informed by BC features while preserving AC characteristics, enabling adaptive and noise-resilient representations. To achieve this, the fusion module employs the Softmax function for adaptive weight allocation, transforming input relevance scores (such as correlation coefficients or signal energies) into a normalized probability distribution within the range $[0, 1]$, summing to unity. Softmax ensures smaller inputs receive nonzero weights, avoiding overemphasis of dominant features or neglect of weaker ones. As a result, the model maintains the relative relationships among feature sources while achieving smooth, interpretable, and balanced fusion.

In our implementation, sentence-level fusion is performed between the enhanced AC features derived from BC speech and the original noisy AC features. Let \mathbf{A} and \mathbf{B} denote the original AC and BC feature vectors, respectively. The feature enhancement module produces a reconstructed AC feature vector $\hat{\mathbf{A}}$ from BC speech. To guide the fusion process, we compute the kernel correlation coefficient r between \mathbf{A} and $\hat{\mathbf{A}}$ using KCCA. This coefficient reflects the nonlinear similarity between the original and reconstructed features. Denoting $r_A = r$ and $r_{\hat{A}} = 1 - r$, the fusion weights w_A and $w_{\hat{A}}$ are obtained using the Softmax function

$$w_A = \frac{\exp(r_A)}{\exp(r_A) + \exp(r_{\hat{A}})}, \quad w_{\hat{A}} = \frac{\exp(r_{\hat{A}})}{\exp(r_A) + \exp(r_{\hat{A}})}. \quad (2)$$

These weights are used for computing the final fused feature vector $\mathbf{F}_{\text{fusion}}$

$$\mathbf{F}_{\text{fusion}} = w_A \mathbf{A} + w_{\hat{A}} \hat{\mathbf{A}}. \quad (3)$$

This adaptive fusion mechanism allows the system to dynamically balance noise-affected and noise-suppressed features, ensuring emotional fidelity and stable performance under noise. The fused features are subsequently forwarded to downstream modules for attention-driven refinement and classification.

C. Resampling Module

The proposed model employs a targeted resampling strategy to address class imbalance in the dataset. As the neutral emotion serves as a baseline for other emotional states, it is oversampled by duplicating existing samples when it is under-represented, ensuring that the neutral class size matches or exceeds that of other emotion categories. This approach effectively corrects the imbalance while maintaining the critical role of neutral emotion as a reference point for emotional analysis. By addressing the imbalance while preserving data integrity, the model achieves better generalization across all emotion categories, particularly preventing the neutral emotion from being overlooked during classification.

D. Emotion-Guided Attention Modulation

To leverage the complementary characteristics of BC and AC speech features, we introduce an emotion-guided attention modulation (EGAM) module. Unlike conventional channel attention mechanisms [37] that rely on activation statistics, EGAM integrates emotion context by modulating attention weights based on emotion representation vectors extracted from intermediate features. This approach highlights emotionally discriminative channels and enhances recognition performance in ambiguous or noisy conditions where statistical activations may not reflect emotional relevance. Consequently, EGAM improves adaptability and robustness across diverse acoustic environments. The mechanism operates in two main stages:

Emotion Vector Construction: Given the input feature tensor $\mathbf{X} \in \mathbb{R}^{B \times T \times C}$ (where B , T , and C represent batch size, temporal length, and channel number), we first obtain a global

representation of the emotional context by applying global average pooling over time

$$\mathbf{g} = \frac{1}{T} \sum_{t=1}^T \mathbf{X}_{:,t,:} \in \mathbb{R}^{B \times C}. \quad (4)$$

Here, \mathbf{g} represents the global average-pooled features across batches, where each row corresponds to a C -dimensional summary vector for a single input sample. The global vector \mathbf{g} is then passed through a series of fully connected layers to generate an abstract emotion embedding vector $\mathbf{e} \in \mathbb{R}^{B \times D}$, where D denotes the embedding dimension

$$\mathbf{e} = \phi \left[\delta \left(\mathbf{g} \mathbf{W}_e^{(1)} + \mathbf{b}_e^{(1)} \right) \mathbf{W}_e^{(2)} + \mathbf{b}_e^{(2)} \right]. \quad (5)$$

Here, $\delta(\cdot)$ denotes the rectified linear unit (ReLU) activation function, $\mathbf{W}_e^{(1)}$ and $\mathbf{W}_e^{(2)}$ are learnable weights, and $\mathbf{b}_e^{(1)}$ and $\mathbf{b}_e^{(2)}$ are the corresponding bias vectors of the two fully connected layers. The function $\phi(\cdot)$ denotes the activation function applied to the final projection layer, typically tanh.

Emotion-Guided Attention Generation: The conventional channel attention mechanism, inspired by the squeeze-and-excitation (SE) block [37], generates attention weights based solely on channel activation statistics. Given an input tensor $\mathbf{X} \in \mathbb{R}^{B \times T \times C}$, global average pooling produces a channel descriptor $\mathbf{f} \in \mathbb{R}^{B \times C}$ that is passed through a two-layer network to generate normalized attention weights

$$\mathbf{z}_1 = \delta(\mathbf{f} \mathbf{W}^{(1)}), \quad \mathbf{W}_{\text{base}} = \sigma(\mathbf{z}_1 \mathbf{W}^{(2)}) \quad (6)$$

where $\sigma(\cdot)$ denotes the sigmoid function, and $\mathbf{W}^{(1)}$, $\mathbf{W}^{(2)}$ are learnable weights. The attention weights are then applied to reweight the channels based on activation strength. While effective, this approach lacks emotional context awareness, which motivates our emotion-guided attention mechanism.

To introduce emotion awareness, the emotion vector \mathbf{e} is projected to match the channel dimension via

$$\mathbf{m} = \sigma(\mathbf{e} \mathbf{W}_m + \mathbf{b}_m) \in \mathbb{R}^{B \times C}. \quad (7)$$

The final emotion-guided attention weights are computed through element-wise modulation

$$\mathbf{W}_{\text{EGAM}} = \mathbf{W}_{\text{base}} \odot \mathbf{m} \quad (8)$$

where \odot denotes element-wise product. The attention weights \mathbf{W}_{EGAM} are applied to the input tensor \mathbf{X} via channel-wise multiplication to produce the recalibrated feature tensor $\mathbf{X}' \in \mathbb{R}^{B \times T \times C}$. Specifically, for each time frame t in the b th sample, attention is applied across all channels as

$$\mathbf{X}'_{b,t,:} = \mathbf{X}_{b,t,:} \odot \mathbf{W}_{\text{EGAM},b,:} \quad (9)$$

where $\mathbf{X}_{b,t,:}$ and $\mathbf{W}_{\text{EGAM},b,:}$ are vectors of length C representing the original features and the emotion-guided attention weights for the corresponding time frame. This operation performs channel-wise attention modulation at each time step, allowing the model to dynamically recalibrate feature importance based on emotional context.



Fig. 2. Recording setup illustration. (a) Anechoic chamber used for dual-channel speech collection. (b) Participant wearing synchronized AC and BC microphones. EEG and EOG sensors were used for auxiliary multimodal acquisition. Facial features are blurred to preserve privacy.

IV. EXPERIMENTS

A. Experimental Setup

1) **BC-AC Multimodal Emotional Speech Corpus:** To support the study of multimodal SER, we constructed a dual-channel emotional speech database containing synchronously recorded both BC and AC speech signals. Recordings were conducted in an anechoic chamber with 100 volunteers (gender-balanced). Following the protocol in [35], AC signals were captured via the left channel using a digital microphone positioned approximately 80 cm in front of the participant, while BC signals were recorded from the right channel using an in-ear BC microphone. Recordings were sampled at 44.1 kHz to ensure high fidelity. Fig. 2 shows the data collection setup: (a) the acoustic environment and (b) a participant wearing the recording equipment. Additional sensors, including an electroencephalography (EEG) cap and four face-mounted electrooculography (EOG) electrodes, were applied for parallel multimodal data acquisition, though this study focuses solely on the analysis of speech signals.

Each recording session lasted approximately 60 min and consisted of two phases: emotion induction and emotion expression. In the first phase, participants viewed a 2–5 min emotion-inducing video clip designed to elicit a target emotional state. Subsequently, they completed a self-assessment by selecting a discrete emotion label and rating its intensity on a scale from 0 (no emotion) to 5 (maximum intensity). In the second phase, participants read predefined textual prompts while maintaining the induced emotion, producing emotionally labeled utterances for subsequent analysis.

The text material consisted of 480 emotional utterances (120 for each of the four target emotions: angry, happy, neutral, and sad), selected from a standardized corpus provided by the iFLYTEK Speech Group. To reduce cognitive fatigue and maintain emotional diversity, we alternated between emotional and neutral utterances within each balanced subset.

We conducted a manual perceptual evaluation to validate the recording quality, resulting in recognition accuracies of 93.56% for emotional speech and 82.38% for neutral speech. After segmenting recordings into individual utterances, the final dataset contained 24 191 utterances. The corpus is partitioned into training, validation, and test subsets in a 70%–15%–15% split, ensuring speaker independence and balanced gender distribution across all categories.

2) *Data Augmentation Strategies*: To improve generalization and noise robustness, we developed a targeted data augmentation strategy for both AC and BC speech signals. For AC speech, we applied a diverse set of augmentations, including speed perturbation, pitch shifting, volume scaling adapted from [36], as well as additional techniques such as frequency and time masking, noise injection, and reverberation simulation tailored for this work. These transformations enhanced the diversity of training data and expanded the dataset eightfold.

We simulated reverberation by convolving clean AC speech with room impulse responses generated using the image method [38]. Additionally, we injected white noise at various signal-to-noise ratios (SNRs) ranging from 30 to -5 dB to enhance system robustness across diverse noise conditions. In contrast, we did not apply reverberation to BC speech due to its inherent resistance to reflected sound. Instead, we employed a pre-filtered noise injection strategy, applying low-pass filtering to align the noise spectrum with the typical 0–1 kHz frequency range of BC signals. We also used additional augmentations, including pitch and speed perturbation, volume scaling, and masking, to simulate realistic distortions while preserving BC’s low-frequency characteristics.

3) *Large Enclosed Test Environment (Room Scenario)*: Processing speech signals in large enclosed environments, such as industrial sites, laboratories, and conference rooms, presents significant acoustic challenges. In these complex spaces, different types of speech signals propagate in significantly different ways. While AC speech is highly susceptible to reverberation-induced distortions due to multiple surface reflections, BC speech maintains signal integrity by bypassing the air medium, resulting in more consistent emotional feature extraction. Based on these inherent signal characteristics, our experimental design aims to assess the stability of BC speech under high-reverberation conditions and to validate the potential of AC–BC speech fusion for SER tasks. To this end, we engineered an experimental environment reflecting the complex acoustic dynamics of large indoor spaces.

To simulate real-world acoustics, we developed a framework that systematically varied room size, reverberation, and noise to generate diverse acoustic conditions. We carefully selected noise sources from the NoiseX-92 database [32], a standard corpus of environmental noise for speech testing to represent realistic environmental interference, including babble, factory, and white noise. The speech source was centrally positioned, and microphones were precisely placed to capture AC and BC speech signals simultaneously.

4) *Small Enclosed Test Environment (In-Vehicle Scenario)*: The unique acoustic characteristics of in-vehicle environments led us to select a small enclosed space that simulates typical vehicular acoustic conditions. This environment features extremely short reverberation times ($T_{60} < 0.4$ s) and complex noise conditions, encompassing engine, wind, road, and in-car babble noise. The primary objective of this experiment was to analyze speech signal propagation across different seating positions and to evaluate the effectiveness of AC and BC speech fusion under realistic driving conditions.

We used an expanded reverberation range for training to enhance model robustness and generalization in real-world

TABLE II
EXPERIMENTAL ENVIRONMENT CONFIGURATIONS

Environment	Room sizes (m)	T_{60} (s)	Noise sources	Speaker position	AC microphone placement
Large enclosed environment	6-8(L)	0.6 - 1.6	Babble, factory, and white noise	Central in a room	Table-mounted
	4-6(W)				
	2.5-3.5(H)				
Small enclosed environment	2.0-2.5(L)	0.2 - 0.4	Mechanical and in-car babble noise	Driver and rear-seat passenger	Rearview mirror and ceiling-mounted
	1.5-1.8(W)				
	1.2-1.5(H)				

in-vehicle environments. Specifically, the training set included T_{60} values ranging from 0.1 to 0.6 s, providing broader coverage of reverberant conditions. For validation, T_{60} values were set between 0.15 and 0.45 s to introduce moderate variation beyond the test conditions while maintaining consistency. The test set T_{60} was fixed at 0.2–0.4 s, reflecting typical in-vehicle reverberation. This setup exposes the model to diverse acoustic conditions while ensuring realistic evaluation, thereby improving robustness and generalization in SER tasks.

To capture the spatial variability of real in-vehicle communication, the experimental setup incorporated speech sources from both the driver and rear-seat passenger positions. In the first experiment, the driver acted as the speaker to simulate the typical SER system during driving. AC microphones were strategically placed near the rearview mirror and ceiling (common locations in automotive systems), while BC signals were captured using near-talk sensors worn by the speaker. In the second experiment, the rear-seat passenger acted as the speaker to investigate speech signal propagation characteristics. This setting reflects real-world scenarios such as family trips or conversations involving multiple passengers. This comprehensive design enables the evaluation of acoustic propagation and speech quality under different speaker locations. The configuration ensures synchronized dual-channel acquisition and robust modeling of complex in-cabin acoustics. The detailed parameter settings for the two experimental environments are documented in Table II.

5) *Parameter Configuration*: The experiments were conducted using a deep learning model implemented in the TensorFlow environment with the following hyperparameter configuration: a batch size of 128, an initial learning rate of $1e^{-3}$, Adadelta optimizer, L1 and L2 regularization coefficients of 0.01, and a cross-entropy loss function. We used two callback functions to enhance training efficiency and reduce overfitting: EarlyStopping, which halts training when validation performance plateaus, and ReduceLROnPlateau, which adaptively decreases the learning rate during performance stagnation.

B. Results and Analysis

1) *Noise Robustness Evaluation*: To evaluate the noise robustness of the proposed framework, we conducted experiments under two distinct environments: a large enclosed room and a small in-vehicle setting. Each environment involved multiple noise conditions and spatial configurations to assess performance stability under real-world acoustic challenges.

a) *Large enclosed environment*: Fig. 3 shows the SER performance in a large enclosed environment under three noise

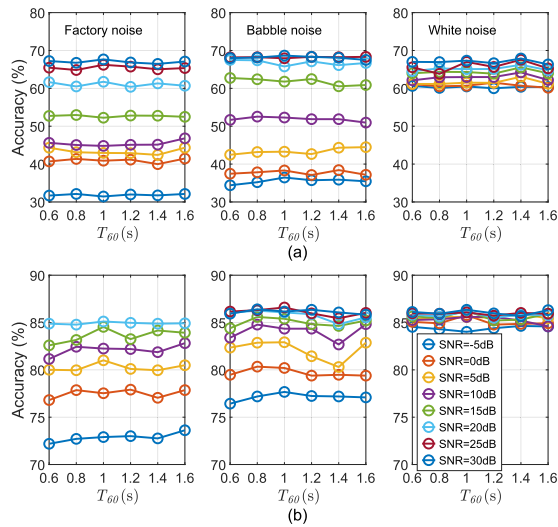


Fig. 3. SER performance in a large enclosed room under various SNR levels, noise types, and T_{60} . (a) Single AC modality results. (b) AC-BC fusion results.

types (factory, babble, and white), multiple SNR levels, and various T_{60} . The experimental results reveal the following key findings.

- 1) The AC-BC fusion consistently outperforms the single AC modality across all noise types, SNR levels, and T_{60} , with the most significant gains at low SNRs. For example, at -5 dB, the fusion model maintains over 70% accuracy, whereas the AC-only system drops below 40%, demonstrating superior noise robustness through multimodal integration.
- 2) Factory noise has the most adverse effect on AC-only systems, particularly at low SNRs, whereas white noise leads to comparatively better performance. With fusion, white noise results in competitive or even better performance, suggesting that the BC channel effectively mitigates additive noise interference.
- 3) Accuracy improves steadily with increasing SNR. In the AC-only case, factory noise shows an improvement of about 30% points as SNR increases from -5 to 30 dB. In contrast, the fusion model exhibits more stable performance across SNRs, indicating reduced sensitivity to noise variations.
- 4) Reverberation time has minimal impact on accuracy at fixed SNRs, especially in the fusion setting. The overall trend remains stable, with slight performance fluctuations observed across T_{60} values. For example, under white noise at -5 dB, the fusion system still achieves approximately 84% accuracy across all T_{60} . This suggests that low SNR affects performance more than reverberation does, with variations in T_{60} having only minor effects.

b) *In-vehicle environment*: Fig. 4 illustrates the SER performance in challenging in-vehicle acoustic environments with the driver as the speech source. AC-only systems exhibit limited accuracy under mechanical noise and in-car babble, with performance degrading significantly at lower SNRs and increased T_{60} . In contrast, the AC-BC fusion approach demonstrates remarkable robustness, outperforming single-modality

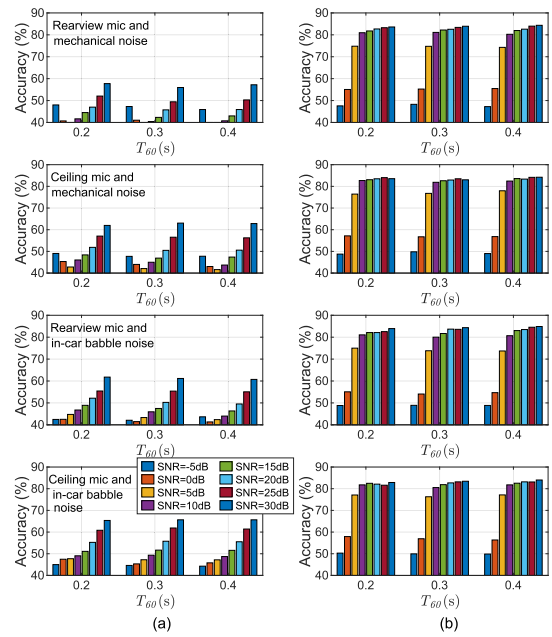


Fig. 4. SER performance in an in-vehicle environment with the driver as the speech source, evaluated with two microphone positions and two noise categories at varying SNR levels and T_{60} . The top two and bottom two rows represent mechanical and in-car babble noise conditions, respectively. (a) Single AC modality results. (b) AC-BC fusion results.

methods across all test conditions. Its effectiveness is particularly pronounced at higher T_{60} and lower SNRs, demonstrating effective acoustic interference mitigation through multimodal integration. Microphone placement also affects performance: for single-modality AC systems, the ceiling microphone generally achieves better accuracy than the rearview microphone, likely due to a more favorable acoustic path and reduced interference.

To validate performance across different speaker positions, we repeated the experiment with the passenger as the speaker. Results from the passenger experiments showed consistent performance trends with those from the driver experiments, confirming the effectiveness of our proposed fusion approach under varied in-vehicle seating configurations.

Fig. 5 illustrates the performance improvement of the proposed AC-BC feature fusion model over individual AC and BC models across four noise types: babble, factory, F16 (a type of aircraft engine noise with high spectral complexity), and white noise. Each bar group represents a different noise type, showing how the fusion approach performs under each condition. Several observations can be made.

- 1) The fusion method consistently outperforms baseline models, with the most pronounced gains at lower SNRs (below 10 dB), demonstrating enhanced noise robustness in challenging acoustic conditions.
- 2) The fusion approach shows the most significant improvement under F16 noise, indicating its effectiveness in handling complex broadband noise. The model shows moderate yet consistent gains under babble and factory noise, while improvements under white noise remain stable across SNRs, highlighting the model's broad applicability.

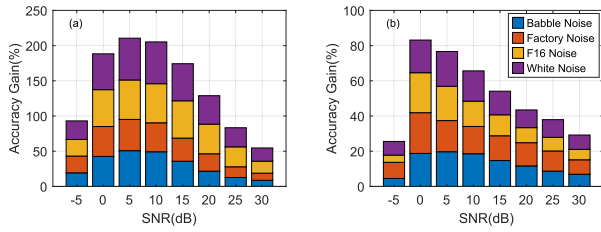


Fig. 5. Absolute performance improvement of the proposed AC-BC feature fusion model compared to the individual (a) AC and (b) BC models.

TABLE III
ABLATION STUDY OF THE ATTENTION-BASED
FEATURE FUSION MODEL

Module				Validation set (%)		
Enhancement	Fusion	Resampling	Attention	Accuracy	Precision	F1-score
✓	✓	✓	✓	82.61	82.57	82.59
✓	✗	✓	✓	85.22	85.23	85.23
✓	✓	✗	✓	74.27	75.22	69.90
✓	✓	✓	✗	86.81	86.82	86.81
✓	✓	✓	✓	87.19	87.22	87.20

- 3) The fusion model substantially outperforms the standalone BC model, particularly at low SNRs. Compared to the AC model, the improvements are even more significant, suggesting that AC-only systems are less robust than BC-only systems.

These results validate our BC-AC fusion approach, which demonstrates robust performance across challenging acoustic environments. The framework integrates BC and AC modalities to maintain stable emotion recognition under low SNR, high reverberation, and complex noise conditions. These findings underscore the method's potential to enhance SER reliability by mitigating acoustic interference and spatial variability. Using both AC and BC features provides clear advantages across diverse noise types by exploiting their complementary characteristics, consistently achieving higher accuracy than using a single modality alone.

2) *Ablation Study*: Table III presents the results of the ablation study. We systematically removed each module from the full model, including feature enhancement, fusion, resampling, and channel attention, to assess its contribution. Performance was evaluated on the validation set using accuracy, precision, and $F1$ -score (the harmonic mean of precision and recall). The complete model consistently outperformed its ablated variants, demonstrating that each component improves feature quality, class balance, and emotion-relevant feature selection.

3) *Comparison With Conventional Methods*: Fig. 6 compares the recognition performance of the proposed EGAMF fusion (EGAMF) method with five alternative strategies: data fusion (DF), weighted SNR fusion (WSNRF), canonical correlation analysis fusion (CCAF), kernel CCAF (KCCAF), and a CMGAN-based pre-enhancement approach (CMGAN + AC). In the CMGAN + AC method, noisy AC speech is first enhanced using the conformer-based metric GAN (CMGAN), and the resulting features are used for subsequent emotion recognition. This baseline assesses the benefit of applying advanced speech enhancement before recognition, without using BC speech. We selected babble noise—nonstationary

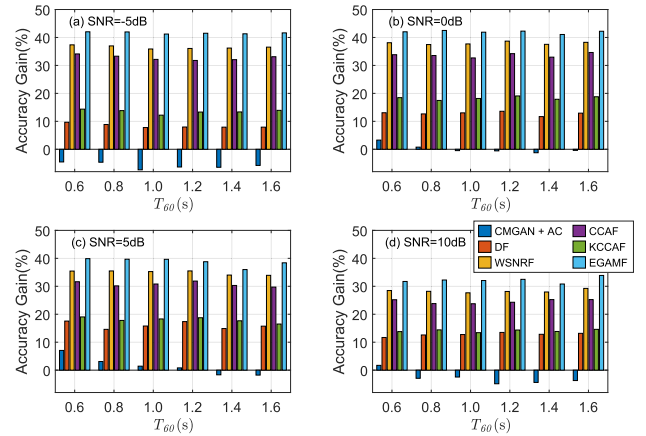


Fig. 6. Comparison of recognition performance between conventional methods and the proposed EGAMF approach. (a) SNR = -5 dB. (b) SNR = 0 dB. (c) SNR = 5 dB. (d) SNR = 10 dB.

and multispeaker in nature—as the background interference to emulate realistic and challenging acoustic conditions. EGAMF consistently outperforms all baselines across a wide range of SNRs and T_{60} . By incorporating channel-wise modulation based on emotional context, EGAMF effectively suppresses irrelevant noise while enhancing emotion-relevant features. These results validate the method's superior noise robustness and adaptive feature selection in complex acoustic environments.

V. CONCLUSION

This work presents a robust SER framework that fuses BC and AC speech to leverage their complementary strengths. The proposed method effectively preserves emotional cues while suppressing noise and reverberation by reconstructing clean AC features from BC inputs and applying emotion-guided attention to fused representations. Extensive experiments in both large and small enclosed environments demonstrate that our proposed model significantly outperforms traditional linear and nonlinear fusion strategies in both recognition accuracy and noise robustness. These results highlight the framework's strong potential for real-world deployment in emotion-aware HMI systems operating under complex acoustic conditions.

REFERENCES

- [1] S. Abdulatif, R. Cao, and B. Yang, "CMGAN: Conformer-based metric-GAN for monaural speech enhancement," *IEEE Trans. Audio, Speech Language Process.*, vol. 32, pp. 2477–2493, 2024.
- [2] Y.-J. Lu, Z.-Q. Wang, S. Watanabe, A. Richard, C. Yu, and Y. Tsao, "Conditional diffusion probabilistic model for speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Singapore, May 2022, pp. 7402–7406.
- [3] H. Yan, J. Zhang, C. Fan, Y. Zhou, and P. Liu, "LiSenNet: Lightweight sub-band and dual-path modeling for real-time speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Hyderabad, India, Apr. 2025, pp. 1–5.
- [4] X. Hao, X. Su, R. Horaud, and X. Li, "FullSubNet: A full-band and sub-band fusion model for real-time single-channel speech enhancement," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 6633–6637.
- [5] S. Kakuba, A. Poulouse, and D. S. Han, "Attention-based multi-learning approach for speech emotion recognition with dilated convolution," *IEEE Access*, vol. 10, pp. 122302–122313, 2022.

- [6] S. Kakuba, A. Poulou, and D. S. Han, "Deep learning approaches for bimodal speech emotion recognition: Advancements, challenges, and a multi-learning model," *IEEE Access*, vol. 11, pp. 113769–113789, 2023.
- [7] C. Li, F. Yang, and J. Yang, "Bone conduction-aided speech enhancement with two-tower network and contrastive learning," *IEEE Trans. Audio, Speech Language Process.*, vol. 33, pp. 163–174, 2025.
- [8] H. Wang, X. Zhang, and D. Wang, "Fusing bone-conduction and air-conduction sensors for complex-domain speech enhancement," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 30, pp. 3134–3143, 2022.
- [9] C. Yu, K.-H. Hung, S.-S. Wang, Y. Tsao, and J.-W. Hung, "Time-domain multi-modal bone/air conducted speech enhancement," *IEEE Signal Process. Lett.*, vol. 27, pp. 1035–1039, 2020.
- [10] M. S. Hosain, Y. Sugiura, N. Yasui, and T. Shimamura, "Deep-learning-based speech emotion recognition using synthetic bone-conducted speech," *J. Signal Process.*, vol. 27, no. 6, pp. 151–163, 2023.
- [11] M. S. Hosain, Y. Sugiura, M. S. Rahman, and T. Shimamura, "EmoBone: A multinational audio dataset of emotional bone conducted speech," *IEEE Trans. Elect. Electron. Eng.*, vol. 19, no. 9, pp. 1492–1506, 2024.
- [12] J. Pohjalainen, F. Ringeval, Z. Zhang, and B. Schuller, "Spectral and cepstral audio noise reduction techniques in speech emotion recognition," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 670–674.
- [13] S. Zhao, Y. Yang, and J. Chen, "Effect of reverberation in speech-based emotion recognition," in *Proc. IEEE Int. Conf. Sci. Electr. Eng. Isr. (ICSEE)*, Eilat, Israel, Dec. 2018, pp. 1–5.
- [14] M. Pandharipande, R. Chakraborty, A. Panda, and S. K. Kopparapu, "An unsupervised frame selection technique for robust emotion recognition in noisy speech," in *Proc. 26th Eur. Signal Process. Conf. (EUSIPCO)*, Rome, Italy, Sep. 2018, pp. 2055–2059.
- [15] A. K. Alimuradov, A. Yu. Tychkov, and P. P. Churakov, "A method for noise-robust speech signal processing to assess human psycho-emotional state," in *Proc. 3rd School Dyn. Complex Netw. Their Appl. Intellectual Robot. (DCNAIR)*, Innopolis, Russia, Sep. 2019, pp. 6–8.
- [16] P. Tran, T. Letowski, and M. McBride, "Bone conduction microphone: Head sensitivity mapping for speech intelligibility and sound quality," in *Proc. Int. Conf. Audio, Lang. Image Process.*, Shanghai, China, Jul. 2008, pp. 107–111.
- [17] M. McBride, P. Tran, T. Letowski, and R. Patrick, "The effect of bone conduction microphone locations on speech intelligibility and sound quality," *Appl. Ergonom.*, vol. 42, no. 3, pp. 495–502, Mar. 2011.
- [18] M. S. Rahman, A. Saha, and T. Shimamura, "Low-frequency band noise suppression using bone conducted speech," in *Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process.*, Victoria, BC, Canada, Aug. 2011, pp. 520–525.
- [19] H. S. Shin, H.-G. Kang, and T. Fingscheidt, "Survey of speech enhancement supported by a bone conduction microphone," in *Proc. Speech Communication; 10. ITG Symp.*, Braunschweig, Germany, Sep. 2012, pp. 1–4.
- [20] M. Wang, J. Chen, X.-L. Zhang, and S. Rahardja, "End-to-end multi-modal speech recognition on an air and bone conducted speech corpus," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 31, pp. 513–524, 2023.
- [21] T. Shimamura and T. Tamiya, "A reconstruction filter for bone-conducted speech," in *Proc. 48th Midwest Symp. Circuits Syst.*, Covington, KY, USA, vol. 2, 2005, pp. 1847–1850.
- [22] K. Kondo, T. Fujita, and K. Nakagawa, "On equalization of bone conducted speech for improved speech quality," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, Vancouver, BC, Canada, Aug. 2006, pp. 426–431.
- [23] T. Dekens and W. Verhelst, "Body conducted speech enhancement by equalization and signal fusion," *IEEE Trans. Audio, Speech, Language Process.*, vol. 21, no. 12, pp. 2481–2492, Dec. 2013.
- [24] P. Singh, M. K. Mukul, and R. Prasad, "Bone conducted speech signal enhancement using LPC and MFCC," in *Proc. 10th Int. Conf. Intell. Hum. Comput. Interact. (IHCI)*, Allahabad, India: Springer, 2018, pp. 148–158.
- [25] T. Shimamura, J. Mamiya, and T. Tamiya, "Improving bone-conducted speech quality via neural network," in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol.*, Vancouver, BC, Canada, Aug. 2006, pp. 628–632.
- [26] H.-P. Liu, Y. Tsao, and C.-S. Fuh, "Bone-conducted speech enhancement using deep denoising autoencoder," *Speech Commun.*, vol. 104, pp. 106–112, Nov. 2018.
- [27] C. Zheng, X. Zhang, M. Sun, J. Yang, and Y. Xing, "A novel throat microphone speech enhancement framework based on deep BLSTM recurrent neural networks," in *Proc. IEEE 4th Int. Conf. Commun. (ICCC)*, Chengdu, China, Dec. 2018, pp. 1258–1262.
- [28] Q. Pan, J. Zhou, T. Gao, and L. Tao, "Bone-conducted speech to air-conducted speech conversion based on CycleConsistent adversarial networks," in *Proc. IEEE 3rd Int. Conf. Inf. Commun. Signal Process. (ICICSP)*, Shanghai, China, Sep. 2020, pp. 168–172.
- [29] C. Zheng, T. Cao, J. Yang, X. Zhang, and M. Sun, "Spectra restoration of bone-conducted speech via attention-based contextual information and spectro-temporal structure constraint," *IEICE Trans. Fundam. Electron., Commun. Comput. Sci.*, vol. 102, no. 12, pp. 2001–2007, 2019.
- [30] H. Q. Nguyen and M. Unoki, "Improvement in bone-conducted speech restoration using linear prediction and long short-term memory model," *J. Signal Process.*, vol. 24, no. 4, pp. 175–178, 2020.
- [31] T. Hussain, Y. Tsao, S. M. Siniscalchi, J.-C. Wang, H.-M. Wang, and W.-H. Liao, "Bone-conducted speech enhancement using hierarchical extreme learning machine," in *Proc. 10th Int. Workshop Spoken Dialogue Syst. Increasing Naturalness Flexibility Spoken Dialogue Interact.* Springer, 2021, pp. 153–162.
- [32] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Commun.*, vol. 12, no. 3, pp. 247–251, Jul. 1993.
- [33] MagicHub. *ASR-CABNOIS: A Cabin Noise Dataset*. Accessed: Mar. 24, 2025. [Online]. Available: <https://magichub.com/datasets/in-vehicle-noise-dataset/>
- [34] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor, "Canonical correlation analysis: An overview with application to learning methods," *Neural Comput.*, vol. 16, no. 12, pp. 2639–2664, Dec. 2004.
- [35] Q. Wang, M. Wang, Y. Yang, and X. Zhang, "Multi-modal emotion recognition using EEG and speech signals," *Comput. Biol. Med.*, vol. 149, Oct. 2022, Art. no. 105907.
- [36] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomed. Signal Process. Control*, vol. 59, May 2020, Art. no. 101894.
- [37] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, UT, USA, Jun. 2018, pp. 7132–7141.
- [38] J. B. Allen and D. A. Berkley, "Image method for efficiently simulating small-room acoustics," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 943–950, Apr. 1979.