



Lightweight Speech Enhancement Model Based on Harmonic Attention and Phase Estimation with Skin-Attachable Accelerometer

Yonghun Song¹, Yeeun Kim¹, Yoonyoung Chung^{1,2,3*}

¹Department of Electrical Engineering, ²Department of Semiconductor Engineering, ³Center for Semiconductor Technology Convergence, Pohang University of Science and Technology (POSTECH), Pohang, Korea

{yhsong, kkkkkkye02, ychung}@postech.ac.kr

Abstract

Skin-attachable accelerometers (ACCs) capture speech vibrations through the skin, providing a noise-robust complement to microphone (MIC) signals. However, prior multi-modal models combining these signals face trade-offs between processing overhead and performance. This study proposes a lightweight ACC-assisted U-Net (LAU-Net) for real-time speech enhancement. The LAU-Net employs a harmonic attention module to enhance spectral clarity by emphasizing speech harmonics predicted from ACC signals while only increasing the parameter count from 92.29k to 92.98k. The phase estimation block of LAU-Net adaptively integrates ACC and MIC phases based on noise levels, eliminating the need for phase data training. The LAU-Net achieves a PESQ of 2.92 with 39M MACs/s on the TAPS dataset, demonstrating a balance between speech quality and computational efficiency. These results highlight the LAU-Net as a practical solution for robust and efficient speech processing with real-time edge deployment.

Index Terms: multi-modal speech enhancement, accelerometer, lightweight, on-device

1. Introduction

Microphones (MICs) are commonly used in digital communication to capture speech. However, MICs are susceptible to ambient sounds, causing significant degradation of speech quality and intelligibility in noisy environments such as battlefields and construction sites. Numerous deep neural network (DNN) speech enhancement methods [1, 2, 3, 4] have been developed to suppress background noise; however, they struggle to maintain robust performance under extremely low signal-to-noise ratio (SNR) conditions where noise dominates the acoustic signal. As an alternative, bone-conduction (BC) sensors, including skin-attachable accelerometers (ACCs) [5, 6] and throat MICs [7], are inherently resistant to background noise. The BC sensors capture bone-conducted vibrations from the neck, ears, or jaw, which contain voice components [8]. These vibrations are widely used in applications such as speech recognition [9], voice activity detection [10], and SNR estimation [11]. However, BC sensors measure only low frequencies below 2 kHz [8], requiring MIC integration to restore high-frequency spectral information essential for speech intelligibility.

Several studies have developed multi-modal speech enhancement models [12, 13, 14] combining BC sensor and MIC signals. For example, a U-Net-based generator [15] utilized BC signals in the time domain to capture speaker-specific characteristics, improving speech quality in noisy environments. Another study [16] processed BC sensor and MIC signals separately using different models in the time-frequency (TF) domain and later fused them. This TF-domain method outperformed time-

domain approaches [12, 13, 15] by leveraging complex spectrogram representations, which facilitate spatial feature learning and phase estimation. However, the increased complexity of spectral representations requires significant computational resources due to the substantial amount of feature extraction [17]. In practical applications, BC sensors are often embedded in wearable devices such as earbuds. Given their limited computing power and memory constraints, these devices require compact models with minimal parameters for real-time operation. Several lightweight multi-modal models [18, 19] with fewer than 4M parameters have been proposed for edge computing platforms. However, these approaches still struggle with a trade-off between computational efficiency and speech enhancement quality, which limits their practical applicability.

This study proposes a lightweight ACC-assisted U-Net (LAU-Net) that enables real-time speech enhancement even in low-resource environments, such as low-power microcontroller units (MCUs). The model was developed and evaluated using the TAPS dataset [20], which contains speech recordings captured with a skin-attachable ACC as the BC sensor. Our LAU-Net adopts a U-Net backbone [21] with two low-complexity components: harmonic attention module and phase estimation block. The harmonic attention module emphasizes harmonic structures estimated from the ACC signal within speech feature maps, allowing the model to focus on speaker-specific frequency components. This module enhances spectral clarity while maintaining a lightweight architecture with minimal parameter overhead. Additionally, the phase estimation block dynamically integrates ACC and noisy MIC phase information based on SNR variations, ensuring efficient phase reconstruction while eliminating the computational complexity of phase data training. As a result, the LAU-Net achieved a perceptual evaluation of speech quality (PESQ) of 2.92 with 93k parameters, outperforming previous models [18, 22, 23]. We further applied 8-bit quantization to optimize computational efficiency, reducing the model size to 0.21 MB. This architecture enables the LAU-Net to achieve real-time speech enhancement on resource-limited hardware, demonstrating its feasibility for low-power wearable and embedded systems.

2. Methods

2.1. Overall architecture

The overall architecture of LAU-Net is shown in Figure 1. Two input signals from a skin-attachable ACC and an acoustic MIC are preprocessed in the TF domain using the short-time Fourier transform (STFT). The noisy MIC signal, which includes ambient noise (n_{noise}), is represented as follows:

$$X_{\text{Noisy}} = \text{STFT}(x_{\text{MIC}} + n_{\text{noise}}), X_{\text{Noisy}} \in \mathbb{R}^{2 \times K \times L}, \quad (1)$$

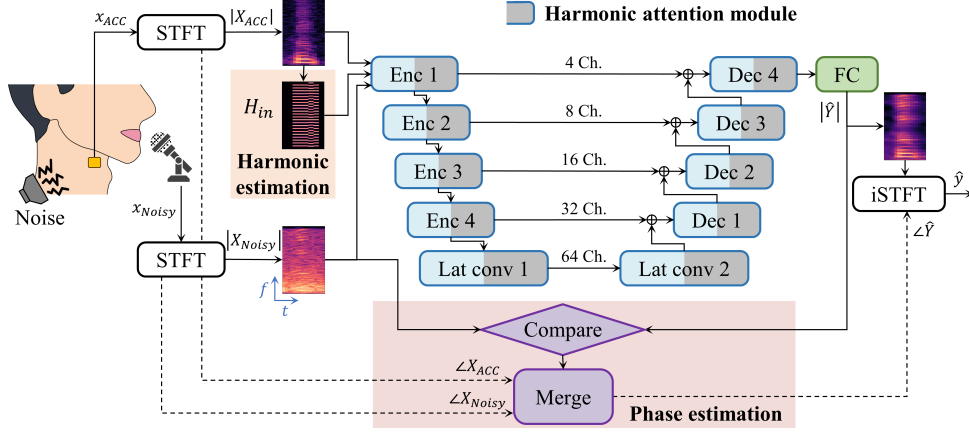


Figure 1: The overall architecture of the lightweight accelerometer-assisted U-Net (LAU-Net). The LAU-Net processes both microphone (MIC) and accelerometer (ACC) signals, utilizing a harmonic attention module and a phase estimation block.

$$X_{ACC} = \text{STFT}(x_{ACC}), X_{ACC} \in \mathbb{R}^{2 \times K \times L}, \quad (2)$$

where K denotes the number of frequency bins, L represents the number of time frames, and the first dimension corresponds to the magnitude and phase. Two magnitude components ($|X_{Noisy}|$, $|X_{ACC}|$) are concatenated into two channels and used as input to the LAU-Net. This model comprises two computationally efficient stages: a harmonic attention module and a phase estimation block. The details of each stage are explained in Sections 2.2 and 2.3. Finally, the magnitude output of LAU-Net is combined with the estimated phase and reconstructed into a time-domain waveform using the inverse STFT (iSTFT). The source codes are available in a public repository¹.

2.2. Harmonic estimation and attention module

We developed the U-Net-based model by incorporating harmonic attention modules into its encoders and decoders to train the harmonic structure of the spectrogram. Figure 2(a) shows the process of extracting harmonic components from ACC signals. The ACC captures low-frequency bands below 2 kHz, excluding ambient noise. We leveraged this advantage to identify frequencies exceeding a predefined threshold in each time frame. The threshold was set to five times the magnitude of silent segments. Among the detected frequencies, the lowest one is selected as the fundamental frequency (f_0). Subsequently, we estimate harmonic components by identifying integer multiples of f_0 , based on the intrinsic harmonic properties of voiced sounds. These harmonic frequencies are assigned an attention score of 1, which linearly decreases with increasing distance from the harmonics on the frequency axis. We set a minimum attention score of 0.5 to avoid excessive sparsity. The harmonic attention score (H_{in}) is defined as follows:

$$H_{in}(K) = \max \left(1 - \alpha \frac{|K - i \cdot f_0|}{M}, 0.5 \right), \text{ for } i \geq 1, \quad (3)$$

where α indicates the gradient factor, $i \cdot f_0$ represents the i -th harmonic frequency, and M denotes the frequency margin.

The LAU-Net takes H_{in} and X_{in} as inputs, where X_{in} is the channel-wise concatenation of $|X_{Noisy}|$ and $|X_{ACC}|$, and follows the U-Net architecture [21]. The encoder blocks repeat a se-

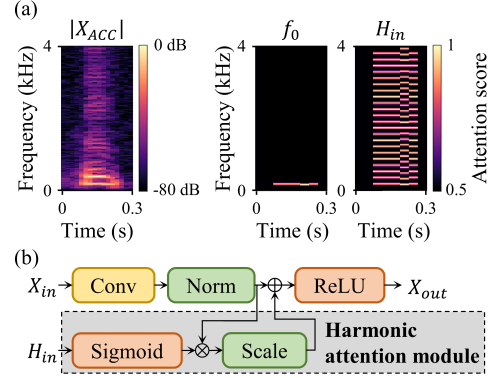


Figure 2: Encoder and decoder blocks with the harmonic attention module. (a) Extraction of harmonic components using the fundamental frequency estimated from accelerometer (ACC) signals. (b) Integration of the harmonic attention module in U-Net to emphasize harmonic structures in the spectrogram. $\oplus$ represents the element-wise summation operation.

quence of two 3×3 convolutions (stride 1, padding 1), batch normalization, ReLU, and 2×2 max pooling, while the decoders mirror this sequence but replace max pooling with up-sampling operations. Skip connections link encoders to corresponding decoder blocks to preserve spatial information. The output tensor of the final decoder is processed by a 1×1 convolution (stride 1) that acts as the fully connected layer, mapping the output from $1 \times 4 \times 256 \times 16$ to $1 \times 1 \times 256 \times 16$. As shown in Figure 2(b), we incorporate harmonic attention modules after every batch normalization layer in the encoder and decoder blocks. In the first encoder block, X_{in} is the initial two-channel input; in all subsequent blocks, it denotes the intermediate activation produced by the preceding block, and the corresponding H_{in} is max-pooled to match the spatial dimensions of X_{in} . The aligned H_{in} is passed through a sigmoid function and then multiplied by the normalized X_{in} . The resulting feature maps are scaled channel-wise by learnable factors, allowing the network to train different harmonic frequencies. These scaled harmonic attention values are added to the normalized X_{in} , effectively emphasizing harmonic components while preserving the original spectral content.

¹<https://github.com/yhsong06/LAU-Net.git>

2.3. Phase estimation block

Phase prediction is particularly challenging for lightweight models due to their limited computational capacity [24]. Instead of training the phase, we propose a novel approach that dynamically integrates the noisy MIC phase ($\angle X_{\text{Noisy}}$) and ACC phase ($\angle X_{\text{ACC}}$). Figure 3(a) illustrates spectrums of MIC, noisy MIC, and ACC signals at -5 dB SNR. The yellow region indicates frequency bands primarily affected by noise, whereas the orange region represents noise-free bands. Figure 3(b) presents the corresponding phase information, showing the phase differences between $\angle X_{\text{MIC}}$ and $\angle X_{\text{Noisy}}$, as well as between $\angle X_{\text{MIC}}$ and $\angle X_{\text{ACC}}$. The phase difference $|\angle X_{\text{Noisy}} - \angle X_{\text{MIC}}|$ is close to zero in low-noise regions, indicating strong phase coherence. However, under high-noise conditions, $\angle X_{\text{ACC}}$ exhibits a closer resemblance to $\angle X_{\text{MIC}}$ than $\angle X_{\text{Noisy}}$. Based on these results, we merge the phase information depending on the ambient noise level. The noise component is estimated as follows:

$$\hat{N}(K, L) = |X_{\text{Noisy}}(K, L)| - |\hat{Y}(K, L)|, \quad (4)$$

where $\hat{N}(K, L)$ represents the noise magnitude, $X_{\text{Noisy}}(K, L)$ denotes the noisy MIC magnitude, and $\hat{Y}(K, L)$ is the magnitude predicted by our model. The phase ($\angle \hat{Y}$) is then determined by comparing $\hat{Y}(K, L)$ and $\hat{N}(K, L)$ as follows:

$$\angle \hat{Y}(K, L) = \begin{cases} \angle X_{\text{Noisy}}(K, L), & \text{if } |\hat{Y}(K, L)| > |\hat{N}(K, L)| \\ \angle X_{\text{ACC}}(K, L), & \text{otherwise} \end{cases} \quad (5)$$

This approach adaptively selects between $\angle X_{\text{Noisy}}$ and $\angle X_{\text{ACC}}$ based on SNR, leading to a phase representation closer to the $\angle X_{\text{MIC}}$. By leveraging reliable phase information from the ACC signal, our method eliminates the need for phase data training, significantly reducing overall computational complexity.

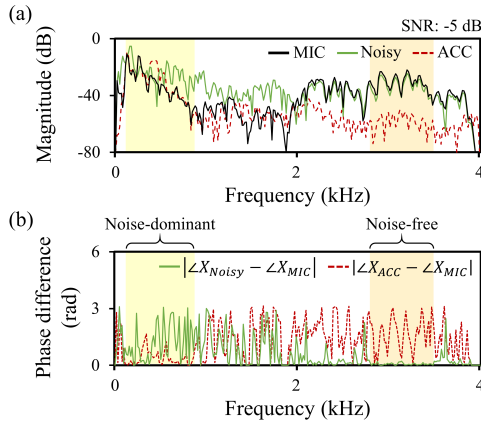


Figure 3: (a) Spectrums of microphone (MIC), noisy MIC, and accelerometer (ACC) signals at -5 dB signal-to-noise ratio (SNR). (b) Phase differences between MIC and noisy MIC signals, as well as between MIC and ACC signals. The yellow region indicates noise-dominant bands, while the orange region represents noise-free bands.

2.4. Loss function

We designed a loss function that incorporates log-magnitude information and harmonic components to guide model training. First, $L_{\text{magnitude}}$ is computed between the network output and

the ground-truth MIC signal to assess overall spectral reconstruction performance [25]. Second, L_{harmonic} is calculated as the L1 loss of the output and MIC signal, weighted by the harmonic attention score (H_j) extracted from the ACC signal. The total loss (L) is defined as an equal-weighted combination of $L_{\text{magnitude}}$ and L_{harmonic} :

$$L = 0.5 \cdot L_{\text{magnitude}} + 0.5 \cdot L_{\text{harmonic}}, \quad (6)$$

$$L_{\text{magnitude}} = \frac{1}{N} \sum_{j=1}^N |\hat{Y}_j - Y_j|, \quad (7)$$

$$L_{\text{harmonic}} = \frac{1}{N} \sum_{j=1}^N |H_j \cdot \hat{Y}_j - H_j \cdot Y_j|, \quad (8)$$

where Y_j represents the magnitude of the ground-truth MIC signal and N denotes the batch size number. This loss enables the model to preserve overall spectral characteristics while effectively enhancing harmonic component reconstruction.

3. Experiments

3.1. Dataset

We used the TAPS dataset [20], a paired corpus of ACC and MIC recordings, to train and evaluate our model. This dataset contains speech from 60 speakers, each uttering 100 different Korean sentences. The dataset is split into 40 speakers for training, 10 for validation, and 10 for testing. We combined the training and validation sets for our experiment, using 50 speakers for training. ACC and MIC signals were downsampled to 8 kHz to reduce computational complexity. Additionally, noise sources from the DNS Challenge dataset [26] were added to MIC signals at SNR levels of -5, 0, 5, and 10 dB. This dataset contains 65k noise clips across 150 classes, from which we randomly selected 6,000 samples without duplication.

3.2. Training details

The data was processed by applying STFT with a frame length of 64 ms and an overlap of 75% using a Hanning window. The logarithmic scaling was then applied to spectrograms. The processed data was divided into 256×16 units, corresponding to 304 ms segments. We implemented the LAU-Net using PyTorch and trained it on a PC equipped with an Intel i7-6800K processor, 128 GB of RAM, and an NVIDIA RTX 3080 Ti GPU. The model was trained for 200 epochs with a batch size of 64 using the Adam optimizer [27] with a learning rate of 0.001.

3.3. Evaluation metrics and baselines

The performance of LAU-Net was evaluated using four objective metrics: the narrowband version of PESQ [28] and three predicted mean opinion scores (CSIG, CBAK, and COVL) [29]. PESQ ranges from -0.5 to 4.5 and assesses the perceptual quality of speech by comparing enhanced signals to clean signals. CSIG evaluates signal distortion, CBAK measures background noise suppression, and COVL reflects the enhanced speech's overall quality, ranging from 1 to 5. Higher values across all metrics indicate better enhancement performance. Furthermore, model complexity was assessed by the number of parameters, multiply-accumulate operations per second (MACs/s), inference time, and model size.

We deployed three baseline models for comparison: NSNet [22], FSPEN [23], and Vibvoice [18]. NSNet and FSPEN are

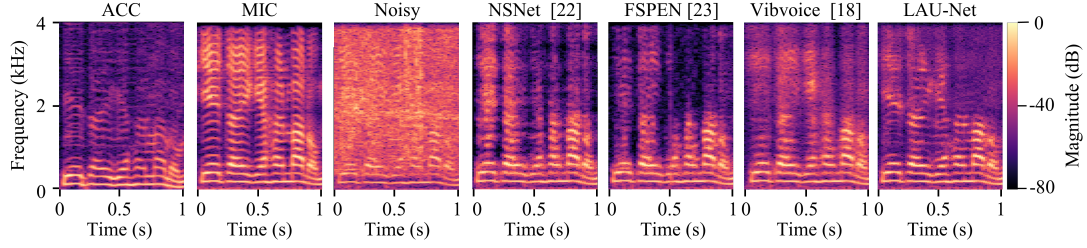


Figure 4: Spectrogram results of various models and our LAU-Net evaluated on the TAPS [20] + DNS Challenge noise dataset [26]. The LAU-Net effectively suppresses noise while reconstructing speech with distinct high-frequency components.

lightweight speech enhancement models that operate solely on audio input with small parameter sizes. In contrast, Vibvoice employs an encoder-decoder structure that integrates both audio and acceleration signals for speech enhancement. All baseline models were retrained using the same dataset (TAPS [20] with DNS Challenge noise [26]). The implementations for each baseline were obtained from publicly available repositories^{2,3,4}, with training configurations following the original setups.

4. Results

4.1. Performance comparison and ablation study

We compared the performance of our LAU-Net with various models, including NSNet, FSPEN, and Vibvoice (Table 1 and Figure 4). Among these models, the LAU-Net achieved the highest PESQ and COVL scores, indicating significant improvements in speech intelligibility and perceptual quality. Audio samples produced by our model are available at <https://yhsong06.github.io/LAU-Net/>. Additionally, the LAU-Net operated with only 93k parameters and 39M MACs/s, significantly lowering computational costs compared to Vibvoice. These results demonstrate LAU-Net’s effectiveness in speech enhancement with low computational complexity.

Table 2 presents the impact of each component in the LAU-Net on speech enhancement performance. The baseline model is a U-Net [21], and incorporating the ACC signal as an additional input improved PESQ and CSIG by +0.42 and +0.52, respectively. This improvement results from the ACC signal providing reliable low-frequency information, which remains unaffected by noise. The harmonic attention module further enhanced all evaluation metrics by effectively extracting key harmonic structures. While the phase estimation block contributed to phase reconstruction, its performance gain was smaller than that of the harmonic attention module. Integrating all components, the LAU-Net achieved the highest PESQ, CBAK, and COVL scores. Although the LAU-Net slightly increased computational cost compared to the baseline, this increase was negligible relative to its performance gains.

4.2. Real-time demonstration

We conducted an embedding test on the MCU (STMicroelectronics, STM32H753ZI) to validate the real-time processing capabilities of LAU-Net, as shown in Table 3. After 8-bit quantization, the model size was decreased approximately 15-fold from 3.17 to 0.21 MB, while the parameter size was reduced from 0.35 to 0.09 MB. Despite these reductions, the quantized

Table 1: Comparison of performance and model complexity

Model	PESQ	CSIG	CBAK	COVL	Para.	MACs/s
Noisy	1.75±0.5	2.74±0.9	2.22±0.7	2.17±0.8	-	-
NSNet [22]	2.18±0.5	3.11±0.7	2.56±0.5	2.58±0.7	1.3M	149M
FSPEN [23]	2.21±0.6	2.78±0.9	2.78±0.5	2.94±0.7	79k	89M
Vibvoice [18]	2.70±0.5	3.76±0.4	2.70±0.4	3.18±0.5	3.1M	2.5G
LAU-Net	2.92±0.5	4.20±0.5	3.42±0.4	3.58±0.5	93k	39M

Table 2: Results of the ablation study

Model	PESQ	CSIG	CBAK	COVL	Para.	MACs/s
U-Net [21]	2.35	3.62	3.03	2.97	92.29k	36.97M
+ ACC signal input	2.77	4.14	3.35	3.47	92.33k	37.47M
+ Harmonic attention	2.89	4.21	3.41	3.56	92.98k	38.98M
+ Phase estimation	2.83	4.16	3.38	3.51	92.33k	37.47M
LAU-Net	2.92	4.20	3.42	3.58	92.98k	38.98M

+ indicates the addition of input, module, or block.

Table 3: Performance comparison based on quantization and computing platforms

Model	PESQ	CSIG	CBAK	COVL	Para. size (MB)	Model size (MB)	Inf. time (ms)
LAU-Net (Full precision, GPU)	2.92	4.20	3.42	3.58	0.35	3.17	7.98
LAU-Net (8-bit Quant., MCU)	2.86	4.19	3.38	3.53	0.09	0.21	144

model maintained competitive performance, with PESQ dropping by 2.0% and COVL by 1.4%. Moreover, the quantized LAU-Net required 144 ms to process a 304 ms input segment on an Arm Cortex-M7 MCU, demonstrating its feasibility for real-time speech enhancement on edge devices.

5. Conclusion

This study presents the LAU-Net, a multi-modal speech enhancement model designed for real-time deployment in resource-constrained environments, such as wearable and edge devices. Our model incorporates a harmonic attention module, which enhances speech harmonics predicted from ACC signals with minimal additional parameters. The phase estimation block of LAU-Net integrates ACC and noisy MIC phase information based on noise levels, reducing the computational cost associated with phase data training. Beyond these low-complexity mechanisms, we applied 8-bit quantization to optimize the model for MCU implementation, ensuring efficient execution on embedded platforms. These advancements allow the LAU-Net to achieve high speech quality with low computational complexity, providing an effective solution for real-time edge applications.

²<https://github.com/GuillaumeVW/NSNet.git>,

³<https://github.com/CHUK-AIoT-Sensing/Eearable.git>,

⁴<https://github.com/gitwukeyi/FSPEN.git>

6. Acknowledgements

This research was supported by the National Research Foundation of Korea grants (RS-2025-00516311, RS-2020-NRF047144) and by the Institute of Information & Communications Technology Planning & Evaluation grant (RS-2019-III191906, Artificial Intelligence Graduate School Program).

7. References

- [1] Y. Hu, Y. Liu, S. Lv, M. Xing, S. Zhang, Y. Fu, J. Wu, B. Zhang, and L. Xie, "Dccrn: Deep complex convolution recurrent network for phase-aware speech enhancement," in *Interspeech*, 2020, pp. 2472–2476.
- [2] H. Zhao, S. Zarar, I. Tashev, and C.-H. Lee, "Convolutional-recurrent neural networks for speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 2401–2405.
- [3] Y. Luo, Z. Chen, and T. Yoshioka, "Dual-path rnn: Efficient long sequence modeling for time-domain single-channel speech separation," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 46–50.
- [4] A. Défossez, G. Synnaeve, and Y. Adi, "Real time speech enhancement in the waveform domain," in *Interspeech*, 2020, pp. 3291–3295.
- [5] K. Lee, X. Ni, J. Y. Lee, H. Arafa, D. J. Pe, S. Xu, R. Avila, M. Irie, J. H. Lee, R. L. Easterlin *et al.*, "Mechano-acoustic sensing of physiological processes and body motions via a soft wireless device placed at the suprasternal notch," *Nature biomedical engineering*, vol. 4, pp. 148–158, 2020.
- [6] Y. Song, I. Yun, S. Giovanoli, C. A. Easthope, and Y. Chung, "Multimodal deep ensemble classification system with wearable vibration sensor for detecting throat-related events," *npj Digital Medicine*, vol. 8, no. 14, 2025.
- [7] J. Cai, S. Marini, P. Malarne, F. Grenez, and J. Schoentgen, "An analytic modeling approach to enhancing throat microphone speech commands for keyword spotting," in *Interspeech*, 2010, pp. 989–992.
- [8] J. Hauret, T. Joubaud, V. Zimpfer, and É. 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 (ICASSP)*, 2023, pp. 1–5.
- [9] M. T. Turan and E. Erzin, "Improving phoneme recognition of throat microphone speech recordings using transfer learning," *Speech Communication*, vol. 129, pp. 25–32, 2021.
- [10] C.-H. Lee, B. D. Rao, and H. Garudadri, "Bone-conduction sensor assisted noise estimation for improved speech enhancement," in *Interspeech*, 2018, pp. 1180–1184.
- [11] H. S. Shin, T. Fingscheidt, and H.-G. Kang, "A priori snr estimation using air- and bone-conduction microphones," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 2015–2025, 2015.
- [12] M. Wang, J. Chen, X. Zhang, Z. Huang, and S. Rahardja, "Multi-modal speech enhancement with bone-conducted speech in time domain," *Applied Acoustics*, vol. 200, p. 109058, 2022.
- [13] 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 Processing Letters*, vol. 27, pp. 1035–1039, 2020.
- [14] Y. Zhou, Y. Chen, Y. Ma, and H. Liu, "A real-time dual-microphone speech enhancement algorithm assisted by bone conduction sensor," *Sensors*, vol. 20, no. 18, p. 5050, 2020.
- [15] M. Tagliasacchi, Y. Li, K. Misiunas, and D. Roblek, "Seanet: A multi-modal speech enhancement network," in *Interspeech*, 2020, pp. 1126–1130.
- [16] H. Wang, X. Zhang, and D. Wang, "Attention-based fusion for bone-conducted and air-conducted speech enhancement in the complex domain," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 7757–7761.
- [17] A. Mehrish, N. Majumder, R. Bharadwaj, R. Mihalcea, and S. Poria, "A review of deep learning techniques for speech processing," *Information Fusion*, vol. 99, p. 101869, 2023.
- [18] L. He, H. Hou, S. Shi, X. Shuai, and Z. 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*, 2023, pp. 14–27.
- [19] M. Ohlenbusch, C. Rollwage, and S. Doclo, "Low-complexity own voice reconstruction for hearables with an in-ear microphone," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.
- [20] Y. Kim, Y. Song, and Y. Chung, "Taps: Throat and acoustic paired speech dataset for deep learning-based speech enhancement," *arXiv preprint arXiv:2502.11478*, 2025.
- [21] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention*, 2015, pp. 234–241.
- [22] Y. Xia, S. Braun, C. K. Reddy, H. Dubey, R. Cutler, and I. Tashev, "Weighted speech distortion losses for neural-network-based real-time speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 871–875.
- [23] L. Yang, W. Liu, R. Meng, G. Lee, S. Baek, and H.-G. Moon, "Fspen: An ultra-lightweight network for real time speech enhancement," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 10 671–10 675.
- [24] H. Wu, K. Tan, B. Xu, A. Kumar, and D. Wong, "Rethinking complex-valued deep neural networks for monaural speech enhancement," in *Interspeech*, 2023, pp. 3889–3893.
- [25] S. Ö. Arık, H. Jun, and G. Diamos, "Fast spectrogram inversion using multi-head convolutional neural networks," *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 94–98, 2018.
- [26] H. Dubey, A. Aazami, V. Gopal, B. Naderi, S. Braun, R. Cutler, A. Ju, M. Zohourian, M. Tang, M. Golestaneh *et al.*, "Icassp 2023 deep noise suppression challenge," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 725–737, 2024.
- [27] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.
- [28] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. 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 (ICASSP)*, 2001, pp. 749–752.
- [29] Y. Hu and P. C. Loizou, "Evaluation of objective quality measures for speech enhancement," *IEEE Transactions on audio, speech, and language processing*, vol. 16, no. 1, pp. 229–238, 2007.