INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

ORIGINAL ARTICLE

Development of a Robust Foreground Speech Enhancement Module for Sub-Optimal Data

Debabrata Gogoi[1]*, **Sushanta Kabir Dutta**[1]

1 Department of Electronics and Communication Engineering, North Eastern Hill University, Meghalaya, 793022, India

OPEN ACCESS

Received: 17/05/2025

Accepted: 22/08/2025

Published: 19/09/2025

Citation: Gogoi D, Dutta SK (2025) Development of a Robust Foreground Speech Enhancement Module for Sub-Optimal Data. Indian Journal of Science and Technology 18(35): 2837-2844. https://doi.org/10.17485/IJST/v18i35.2678

* **Corresponding author**.

debabratagogoi@gmail.com

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Objectives: This work aims to enhance foreground speech by effectively removing unwanted background noise and recovering the desired signal, utilizing deep learning approaches with limited training data. **Methods:** This study addresses the above issue using a transfer learning-based technique that uses mel-spectrograms. Specifically, it proposes a transfer learning approach that builds on a pre-trained residual network (based on wav2vec2 model) that includes a statistics pooling layer as used in speaker recognition. The model is then trained using a limited amount of clean and noisy datasets. In addition, we adopt a log mel-spectrogram feature extraction technique to improve the generalization of speech enhancement models. The database used here is from the Noisy Speech Database curated by Valentini-Botinhao, Cassia (2017) and the LibriSpeech corpus. **Findings:** Using the same dataset, the performances of the baseline model of an autoencoder and a multilayer autoencoder were compared with the proposed model. The proposed approach with an STOI score of 0.88 and an SNR improvement of 3.27 dB, outperforms both the baseline models in subjective and objective evaluation. **Novelty:** This work eliminates signal truncation, a constraint observed in conventional speech enhancement pipelines, by integrating a statistics pooling layer with a pre-trained wav2vec2-based residual network for variable-length input handling. Furthermore, the model's robustness and flexibility are enhanced by the use of log mel-spectrograms in this context, allowing it to produce state-of-the-art results even with sparse supervised training data.

Keywords: Denoise; Mel-spectrogram; Signal Processing; Transfer Learning; Wav2Vec2

1 Introduction

Foreground speech enhancement is the process of increasing clarity and intelligibility of the desired speech signal by removing the background noises. The process is useful in environments in which multiple sound sources are present, and we need to concentrate only required speech. Some key aspects of foreground speech enhancement is the reduction of any stationary or non- stationary noises, speech separation, reverberation

in closed environments, speech intelligibility improvement etc. In order to perform the speech enhancement tasks many algorithms have been developed over past several decades. These algorithms were developed based on the principle of spectral subtraction, statistical analysis, different filtering methods being used widely, until the recent development of neural network-based approach [1].

Challenges faced in the traditional machine learning method are the manual selection of features and poor performance during removal of unseen noises, especially when the model is not familiar with relevant features [2]. In recent years, with the neural network architecture available, researchers proposed concepts like attention mechanisms with commendable performance in tasks related to speech enhancement. These mechanisms allow the models to learn more from the input that helps to collect more relevant features. In [3], [4], [5] proposes attention mechanism to capture deep multiple instances learning to capture foreground speech. In [6] proposed a model that works with the real noisy speech by developing a bridging module using perceptual quality metrics and checking the word error rate in order to analyse the impact of bridging model. Model training learns the optimal observation, addition of coefficients without relying on the clean reference speech. In [7] Proposed a real-time speech enhancement deep learning framework based on Deep Filter Net to remove unwanted signals and reverberation. By integrating traditional signal processing concepts and a decoder, the approach effectively reduces noise. The two-step and simultaneous inference strategies significantly enhance speech quality, especially in reverberant environments. Recent research also has explored the Multi-Modal speech enhancement network that involves combining information from various microphones, such as visual cues or contextual content, to improve the robustness and accuracy of enhancement models. In [8] uses fully convolutional wave to wave model that reduces the losses in synthesis adding the accelerometer data further improves the performance. Researchers also use additional information like visual data of lip movements to improve the performance in overlapping speech [9], [10], [11]. In [12] proposed a model built on a ResNet architecture, demonstrated that emphasizing background-dominant regions and using ensemble methods significantly improved classification accuracy across varying signal-to-background ratios, offering valuable techniques for noise-resilient audio processing.

Most of the models, while trying to capture more information to get better performance, often become more computationally expensive which leads to challenges in the realistic deployment, such as real-world environmental nose cancellation in the real time. Our study addresses this challenge of getting a better speech enhancement model by using an optimal dataset. The model is built upon the wav2vec2, a light weight pretrained architecture, added with custom layers [13]. Building upon this setup, our study aims to the extraction of meaningful representations for speech enhancement, especially in challenging acoustic environments while reducing the need for extensive training data and computational resources. By addressing these objectives, the findings of our study contribute to provide valuable insights into the application of transfer learning in speech enhancement and to the development of more robust and efficient speech enhancement systems.

The rest of the paper is organized as follows: section 2 details the methodology, section 3 deals with results and discussion and section 4 concludes the manuscript with a future direction.

2 Methodology

2.1 Database Description

The collection of works utilized in this study comprises of standard repositories from both clean and noisy parallel speech datasets. The Noisy Speech Database curated by Valentini-Botinhao, Cassia (2017) and the LibriSpeech corpus, featuring audio recordings sourced from LibriVox audiobooks, collectively constitute 1000 hours of speech recordings sampled at a rate of 16 kHz [14], [15]. These recordings were augmented with random environmental noises which were sourced from the ESC dataset as meticulously curated by Piczak (2015) [16]. This dataset has been instrumental in facilitating research in sound classification and has contributed significantly to advancements in audio processing and machine learning techniques. Additionally, the urban soundscape dataset, introduced by Salamon and Bello in 2015, provides a comprehensive resource for the study of urban acoustic environments [17]. This dataset offers a well-defined taxonomy and is a rich dataset that has played a crucial role in propelling research in this domain. The development of this database was tailored specifically to facilitate the training and evaluation of speech enhancement methods operating at a high sampling rate of 48 kHz. Before training the model training, we carefully pre-processed, including thorough quality checks and data filtering, to ensure that only high-quality samples contributed to the training process. Additionally, both the clean and noise waveforms down-sampled to a standard sampling rate of 16 kHz. This down-sampling helped reducing the computational demands and training time without any compromising the datasets accuracy.

2.2 Model Architecture

The proposed Transfer learning based Multilayer Autoencoder architecture designed to enhance foreground speech quality and intelligibility using a combination of Long Short-Term Memory (LSTM) layers and linear transformations is shown in figure 1. Input containing noisy speech data and comprising features extracted from a pre-trained wav2vec2 model. Wav2Vec2.0, developed by Facebook AI team, is a self-supervised model for speech representation learning that eliminates the need for manual transcription by using large scale unlabeled raw audio. The model architecture comprises a feature encoder, a quantization module, and a transformer-based contextual encoder. The feature encoder maps the raw waveform $x \in \mathbb{R}^T$, where T represents total number of time samples in the raw waveform, into a latent representation $z \in \mathbb{R}^{T' \times d}$ where $T' < T$ and d denotes the embedding dimension. The transformer network further processes z to generate contextualized representations that capture long-range phonetic dependencies. In this work, we utilize the pretrained model *facebook/wav2vec2-base-960h*, trained on 960 hours of the LibriSpeech corpus, which enables effective speech feature extraction directly from the waveform input. Encoder Layers contain LSTM layer with input size 256 and hidden size 64, processing the noisy speech input and LSTM layer with input size 768 and hidden size 64, handling the wav2vec2 features. Concatenation of encoder outputs and wav2vec2 encoder outputs (all hidden), facilitating information fusion. Two diffusion bidirectional LSTM layer with input and hidden size of 64, processing the combined information further refining the information. Temporal truncation is applied to match the length of the input signal to avoid any mismatch error during training. For the reconstruction of enhanced speech LSTM layer with input size 128 and hidden size 64 is implemented finally linear transformation layer at the output, mapping the decoder output to the desired output dimensionality. Output estimation is computed as the noisy input subtracted by the linear transformation of the decoder output and ultimately reconstructing the enhanced speech.

The proposed system begins with the raw waveform input for noisy mel spectrogram as $X^{(n)} \in \mathbb{R}^{T \times 256}$and for Wav2Vec2 Features $X^{(w)} \in \mathbb{R}^{T \times 768}$, which is processed through the Wav2Vec2 encoder to extract high-level, context-aware speech features. This step can be represented as:

$$H^{(w)} = LSTM_w(X^w) \in \mathbb{R}^{T \times 64} \tag{1}$$

In parallel, a secondary feature encoder represented as:

$$H^{(n)} = LSTM_{noisy}(X^n) \in \mathbb{R}^{T \times 64} \tag{2}$$

where T is the total number of temporal frames and 64 denotes the feature vector dimensionality per frame. These temporal feature sequences are then subjected to statistics pooling to produce utterance-level representations by computing the mean and standard deviation along the time axis. Traditional speech enhancement pipelines typically require input utterances to be truncated or padded to a fixed length L, particularly when using batch training or convolutional networks. This process often leads to information loss or computational inefficiency. In contrast, our proposed model integrates a statistics pooling layer that enables processing of variable-length inputs without truncation. Let the sequence of latent features from either the log Mel-spectrogram or the Wav2Vec2 encoder be denoted by:

$$H = \{h_1, h_2, \ldots\ldots, h_T\}, \;\; h_t \in \mathbb{R}^d \tag{3}$$

$$\mu(H) = \frac{1}{T}\sum_{t=1}^{T} h_t \tag{4}$$

$$\sigma(H) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}(h_t - \mu)^2} \tag{5}$$

Where,

$\mu(H)$ is the mean vector.

$\sigma(H)$ is the standard deviation vector.

Here, h_t Feature vector at time frame t ranging from 1 to T in d-dimensional real-valued vector space. To obtain a fixed-size utterance-level representation from a variable-length sequence, here apply a statistics pooling function, which can then

be passed to fully connected or residual layers without requiring truncation or padding to a fixed length. The pooled vector is obtained by concatenating both statistics:

$$Z^{(w)} = StatisticalPool\,(H^w) = [\mu(H)||\sigma(H)] \in \mathbb{R}^{2d} \tag{6}$$

Where,

|| denotes concatenation.

$Z^{(w)}$ is the final statistical pooling output.

After feature fusion, the representation is passed through two parallel Bidirectional LSTM branches. The first branch captures dominant bidirectional temporal dependencies, while the second produces modeling complementary residual temporal variations. This dual-path recurrent encoding enriches temporal context, enhancing the model's ability to separate speech from noise. These embeddings are then aligned with $Z^{(w)}$ through a fusion mechanism such as feature concatenation:

$$F = \left[\overline{Z}^{(w)} \parallel \overline{H}^{(n)}\right] \in \mathbb{R}^{T \times d_f} \tag{7}$$

where $\overline{Z}^{(w)} and\, \overline{H}^{(n)}$ denotes the representations obtained after the pooling and and Bidirectional LSTM processing steps, respectively and $d_f$represents the final fused dimensionality i.e. 128 dimensions. The fused representation is passed to the output prediction layer:

$$\widehat{Y} = \sigma(WF + b) \tag{8}$$

Where W and b are learnable parameters, $\sigma(\bullet)$ is the activation function, and $\widehat{Y}$ is the enhanced speech representation. This sequential transformation ensures that both contextual $Wav2Vec2$ and $spectral\,(Mel)$ information's are used for robust noise suppression and feature discrimination.

2.3 Training and Validation Analysis

The Figure 1 shows the LSTM-Autoencoder that builds on a pre-trained wav2vec2 model. The results of the same are available in Table 1.

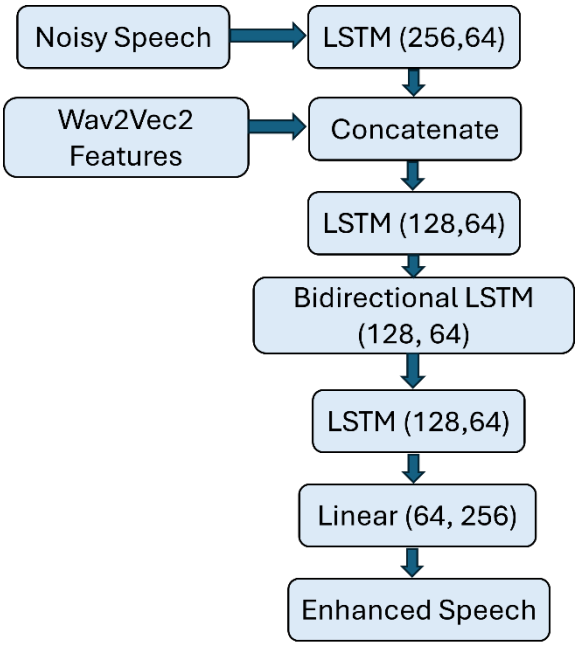

Fig 1. Transfer Autoencoder Network Architecture

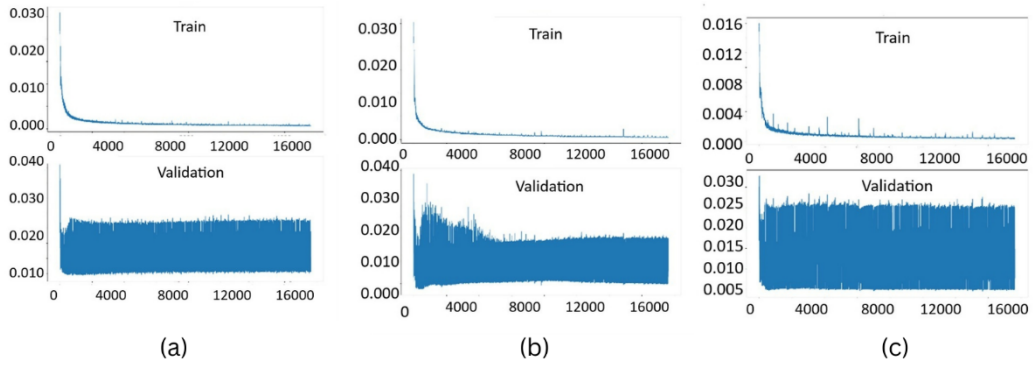

Fig 2. (a) Encoder, (b) Multilayer Encoder and (c) Transfer Learning, Training and Validation Results: EPOCH vs LOSS

Table 1. Training and Validation Results

Description	Train	Validation
Autoencoder	0.00640	0.00681
Multilayer Autoencoder	0.00601	0.00618
Transfer Learning with Autoencoder	0.00016	0.00587

The above experiment demonstrates the advantage of combining log Mel-spectrograms with pretrained contextual Wav2Vec2 features. The pretrained model extracting last hidden state, i.e., the contextual embeddings of the audio features allow the model to generalize well with minimal supervision, adopting transfer learning using large scale unlabeled speech corpora.

3 Results and Discussion

The proposed model achieves effective noise reduction by leveraging a dual-stream fusion of log Mel-spectrogram features and contextual embeddings from a pretrained Wav2Vec2 model. The key mechanism through which noise is attenuated lies in the learned representations and temporal statistics, which together enable the network to differentiate between speech and noise patterns. Unlike traditional models that treat speech enhancement as a purely spectral mapping task, our approach enriches the input with semantic and phonetic context extracted by Wav2Vec2. These representations are by nature less sensitive to short-term, non-linguistic variations such as environmental noise, since the model learns to long-range dependency and linguistic regularity. It therefore enables the model to keep features that are relevant to speech while dropping incoherent or unstructured noise components.

Also, the statistics pooling layer collects time by calculating mean and standard deviation over time, so it maps variable-length input to a global representation that contains context. This operation naturally reduces random variation, that is noise and brings out long-lasting structured features like speech. Noise is normally random and not periodic, hence its effect becomes small when start aggregation takes place.

Moreover, the residual layers in the network facilitate the learning of fine-grained corrections by explicitly modeling the difference between noisy and clean speech embeddings. The residual learning structure works as an implicit denoising function thereby focusing the optimization on what it learns to remove rather than building the whole signal from ground zero. In summary, these components enable the model not to reduce noise based on specific frequency bands but by abstract representations in which speech and noise are linearly separable hence a generalizable robust denoising process under different input conditions.

Figures 3 to 5 show the time series and Mel-spectra of noisy and enhanced speech samples obtained from encoder model, multilayer encoder and the proposed transfer learning model. From here, it is clearly seen how the noise has been reduced significantly in the last one.

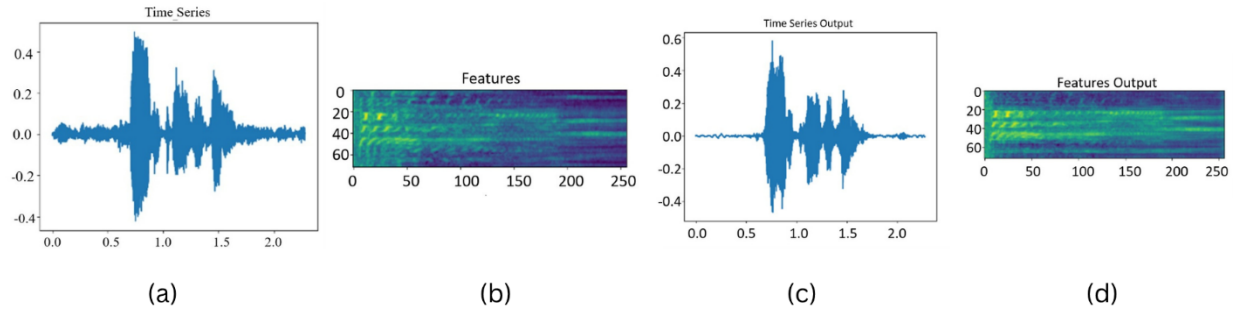

Fig 3. Time Series (Time in second vs Amplitude) and Mel-Spectra (Time in Millisecond vs Mel Frequency) using Encoder model (a) Noisy Speech Time Series (b) Noisy Speech Mel-Spectrogram (c) Enhanced Speech Time Series (d) Enhanced Speech Mel-Spectrogram

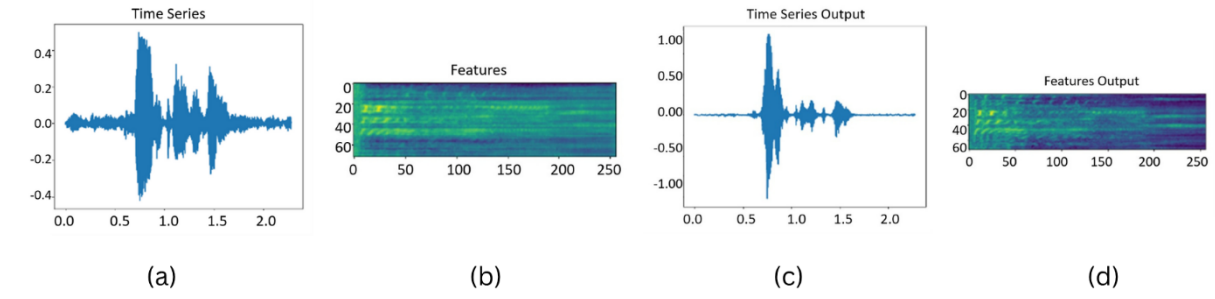

Fig 4. Time Series (Time in second vs Amplitude) and Mel-Spectra (Time in Millisecond vs Mel Frequency) using Multilayer-Encoder model (a) Noisy Speech Time Series (b) Noisy Speech Mel-Spectrogram (c) Enhanced Speech Time Series (d) Enhanced Speech Mel-Spectrogram

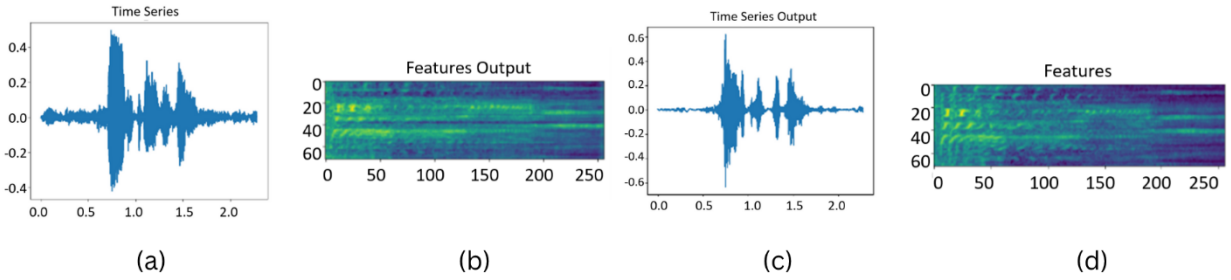

Fig 5. Time Series (Time in second vs Amplitude) and Mel-Spectra (Time in Millisecond vs Mel Frequency) using the Transfer Learning model (a) Noisy Speech Time Series (b) Noisy Speech Mel-Spectrogram (c) Enhanced Speech Time Series (d) Enhanced Speech Mel-Spectrogram

3.1 Assessment of Accuracy

The proposed model based on a transfer learning approach is a robust and efficient architecture for speech enhancement. The model is simple enough for implementation as it is not computationally expensive. Yet, it can provide a fair amount of noise reduction. The performance of the proposed model is compared with similar models like Deep Denoising Autoencoder (DDAE) [18] and Denoising Autoencoder with Masking Estimation (DAEME) [19], having a low resource utilization yet capable enough to provide the required noise reduction. Upon such comparison as shown in Table 2, it is seen that our model outperforms all the contemporaries. The measurement parameters used are STOI and SNR (dB). For assessment of accuracy of the proposed model, the enhanced speech samples obtained from output are compared with reference original clean speech. While objective metrics, such as signal-to-noise ratio (SNR) and speech quality measures, provide valuable insights, they may not always accurately reflect the perceived quality of the enhanced speech [20], [21]. Subjective listening tests, on the other hand, offer a more direct and holistic approach to evaluating the effectiveness of speech enhancement techniques.

Table 2. Accuracy Assessment

Description	Analysis Parameter	DDAE	DAEME	Auto encoder	Multilayer Autoencoder	Transfer Learning
Non Stationary Noise	SNR Before Enhancement	2.27 dB				
	SNR After Enhancement	NA	NA	4.69 dB	4.91 dB	5.54 dB
	STOI Before Enhancement	0.65				
	STOI After Enhancement	0.83	0.84	0.72	0.75	**0.88**

Here subjective listening tests included a group of 20 human subjects who assess the quality and speech intelligibility for enhanced speech by all the models. Listeners were asked to grade the enhanced speech quality, using a scale from one to five. In total, 14 speech samples were analysed, and resulted output of our encoder model is an average rating of 2.5, while those samples processed through the multilayer encoder attained an average rating of 3.6. Notably, employing transfer learning yielded the highest average rating of 3.8, this represents a significant improvement over the baseline in terms of speech clarity and comprehension. Through the summarized subjective listening test, it is noted that transfer learning outperforms the existing approaches in speech enhancement.

4 Conclusion

This study presented a novel approach for speech enhancement that uses the power of transfer learning. By adopting a pre-trained deep learning model, we are able to achieve better performance as well as minimizing the requirement for extensive datasets commonly needed for training such complex models from scratch. Additionally, our method demonstrates a significant reduction in computational costs, making it a practical solution for real-world applications with limited resources available. The effectiveness of the proposed technique was validated through rigorous testing, where it showed promising performance, particularly in the context of non-stationary noise. Our proposed model achieved 0.85 dB higher SNR than the standard autoencoder and 0.63 dB higher SNR than multilayer autoencoder. For speech intelligibility (STOI) test our model jumped from 0.65 to 0.88, meaning listeners could catch nearly 90% of words versus 73% with the basic autoencoder and outperforms the results from the existing autoencoder models DDAE (0.84) and DAEME (0.83). Despite these positive outcomes, challenge remain in handling of highly non-stationary noise environments. Future work may be encouraged in two key opportunities for one is systematic parameter optimization that could enhance the model's adaptive capabilities, and another in expanding the training corpus to include more diverse, real-world noise profiles which would likely improve generalization.

References

1) Vary P, Martin R. Digital Speech Transmission and Enhancement. John Wiley & Sons. 2023. https://doi.org/10.1002/9781119060970.
2) Yuliani AR, et al. Speech enhancement using deep learning methods: A review. *Jurnal Elektronika dan Telekomunikasi*. 2021;21(1):19–26. https://doi.org/10.14203/jet.v21.19-26.
3) Parisae V, Bhavanam SN. Adaptive attention mechanism for single channel speech enhancement. *Multimedia Tools and Applications*. 2025;84(2):831–856. https://doi.org/10.1007/s11042-024-19076-0.
4) Hung JW, Li TJ, Su BY. Enhancing Subband Speech Processing: Integrating Multi-View Attention Module into Inter-SubNet for Superior Speech Enhancement. *Electronics*. 2025;14(8):1640. https://doi.org/10.3390/electronics14081640.
5) Hebbar R, et al. Deep multiple instance learning for foreground speech localization in ambient audio from wearable devices. *EURASIP Journal on Audio, Speech, and Music Processing*. 2021;2021(1):7. https://doi.org/10.1186/s13636-020-00194-0.
6) Liu P, et al. Fusion Attention Mechanism for Foreground Detection Based on Multiscale U-Net Architecture. *Computational Intelligence and Neuroscience*. 2022;2022(1):7432615. https://doi.org/10.1155/2022/7432615.
7) Cui Z, et al. Reducing the Gap Between Pretrained Speech Enhancement and Recognition Models Using a Real Speech-Trained Bridging Module. *arXiv preprint arXiv:250102452*. 2025. https://doi.org/10.1109/icassp49660.2025.10889737.
8) Rosenbaum T, et al. Deep-Learning Framework for Efficient Real-Time Speech Enhancement and Dereverberation. *Sensors*. 2025;25(3):630. https://doi.org/10.3390/s25030630.
9) Tagliasacchi M, et al. SEANet: A multi-modal speech enhancement network. *arXiv preprint arXiv:200902095*. 2020. https://doi.org/10.48550/arXiv.2009.02095.
10) Xiong J, et al. Look&listen: Multi-modal correlation learning for active speaker detection and speech enhancement. *IEEE Transactions on Multimedia*. 2022;25:5800–5812. https://doi.org/10.1109/tmm.2022.3199109.
11) Simone GD, Greco A, Rosa F, et al. Context-aware data augmentation for enhanced speech command recognition in industrial environments. *Scientific Reports*. 2025;15:17445. https://doi.org/10.1038/s41598-025-01886-3.
12) Song S, Song Y, Madhu N. Robust Detection of Background Acoustic Scene in the Presence of Foreground Speech. *Applied Sciences*. 2024;14(2):609. https://doi.org/10.3390/app14020609.
13) Yerramreddy DR, et al. Speech Recognition Paradigms: A Comparative Evaluation of SpeechBrain, Whisper and Wav2Vec2 Models. IEEE. 2024. https://doi.org/10.1109/i2ct61223.2024.10544133.

14) Valentini-Botinhao C. Noisy speech database for training speech enhancement algorithms and tts models. *University of Edinburgh School of Informatics Centre for Speech Technology Research (CSTR)*. 2017. https://doi.org/10.7488/ds/2117.
15) Panayotov V, et al. Librispeech: an asr corpus based on public domain audio books. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2015. https://doi.org/10.1109/icassp.2015.7178964.
16) Piczak KJ. ESC. *Proceedings of the 23rd ACM International Conference on Multimedia*. 2015. https://doi.org/10.1145/2733373.2806390.
17) Salamon J, Jacoby C, Bello JP. A Dataset and Taxonomy for Urban Sound Research. *Proceedings of the 22nd ACM International Conference on Multimedia*. 2014. https://doi.org/10.1145/2647868.2655045.
18) Liu HP, Tsao Y, Fuh CS. Bone-conducted speech enhancement using deep denoising autoencoder. *Speech Communication*. 2018;104:106–112. https://doi.org/10.1016/j.specom.2018.06.002.
19) Yu C, et al. Speech enhancement based on denoising autoencoder with multi-branched encoders. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*. 2020;28:2756–2769. https://doi.org/10.1109/taslp.2020.3025638.
20) Lin Z, Zhou L, Qiu X. A composite objective measure on subjective evaluation of speech enhancement algorithms. *Applied Acoustics*. 2019;145:144–148. https://doi.org/10.1016/j.apacoust.2018.10.002.
21) Benesty J. Fundamentals of speech enhancement. Berlin: Springer. 2018. https://doi.org/10.1007/978-3-319-74524-4_4.