Enhancement and Reconstruction of Dysphonic Kannada Speech Using a Generative Adversarial Network and a SepFormer Model

P. Rajeswari
JSS Science and Technology University, Manasagangotri, Mysuru, Karnataka, India
rajju_p@sjce.ac.in (corresponding author)

N. Shankaraiah
S.J. College of Engineering, JSS Science and Technology University, Manasagangotri, Mysuru, Karnataka, India
shankaraiah@sjce.ac.in

S. Rathnakara
S.J. College of Engineering, JSS Science and Technology University, Manasagangotri, Mysuru, Karnataka, India
rathnakara_s@sjce.ac.in

Received: 15 September 2025 | Revised: 16 October 2025 | Accepted: 24 October 2025

 DOI: https://doi.org/10.48084/etasr.14812

ABSTRACT

Human speech is the most effective form of communication, enabling individuals to convey their thoughts, ideas, and emotions clearly to others. However, many individuals suffer from different types of speech disorders, among which a common speech disorder is dysphonia. This speech disorder not only hampers everyday interactions but also affects the overall quality of life for an individual. Many researchers have worked in this field to develop various modern tools to convert dysphonic speech into normal speech. In spite of its impact, limited emphasis has been placed on addressing the challenges of dysphonia in languages other than English. This paper presents innovative, ensemble learning-based methods designed to improve dysphonic speech signals in Kannada, one of the most widely spoken languages in South India. In this paper, new deep learning methods, such as the Generative Adversarial Network (GAN) and the SepFormer model, are used for enhancement and reconstruction of Kannada dysphonic speech signals. Compared to the GAN, SepFormer provides better results in terms of objective evaluation metrics.

Keywords-dysphonia; Generative Adversarial Network (GAN); SepFormer; speech enhancement

I. INTRODUCTION

The fundamental mode of human communication is speech, enabling individuals to articulate thoughts, ideas, and emotions with clarity. This complex process involves both the production and understanding of spoken language, fostering significant interactions and strengthening relationships. When there are abnormalities in the vocal cords or other components of the vocal system, voice disorders occur. This may affect the pitch, volume, or quality of a person's voice. Different types of voice disorders exist, including:

- Hoarseness: This condition presents as a raspy, breathy, or strained voice and is commonly caused by vocal cord strain, inflammation, or overuse.
- Aphonia: Aphonia refers to the complete inability to produce sound. This condition can arise from various factors, including vocal cord paralysis, severe laryngeal damage, or psychological issues affecting voice production.
- Dysphonia: Dysphonia is a broad term that includes various difficulties in voice production. It involves changes in pitch, loudness, or quality of the voice.

The development of techniques designed to improve the clarity and intelligibility of speech affected by additive noise helps address the challenges associated with dysphonia. To mitigate the impact of noise on speech quality, numerous researchers have focused on this issue, investigating and developing a variety of methods and algorithms over the last decade. Most speech enhancement systems aim at removing artifacts, typically considering only the spectral magnitude based on Short-Term Fourier Transform (STFT) analysis/synthesis [1]. Consequently, the short-term phase is not considered for the enhancement of speech. Later,

researchers developed methods that also incorporate the phase information to improve speech quality [2].

Within the deep learning framework, generative methods directly model the distribution of clean speech, avoiding mode collapse and enabling effective utilization of speech priors. The Generative Adversarial Network (GAN) is a state-of-the-art (SOTA) network model [3]. Two key variations include the introduction of learnable skip connections and the reduction of architecture size through larger convolutional strides, which increases adversarial training stability [4].

Following the development of GAN-based methods, researchers worked on SOTA methods that rely on discriminative training approaches for speech enhancement, although these methods often perform poorly under adverse Signal-to-Noise Ratio (SNR) conditions [5]. For real-time speech enhancement, DeepFilterNet was developed, leveraging harmonic structure and capable of running on embedded systems [6]. Researchers have also used a recurrent U-Net lightweight model for speech enhancement [7]. To further improve performance, a Convolutional Recurrent Encoder–Decoder (CRED) structure was proposed for monaural speech enhancement, overcoming the drawbacks present in Convolutional Recurrent Networks (CRN) through an integrated convolutional encoder–decoder architecture [8].

Recent studies have shown that under very low SNR conditions, where the desired speech is often completely masked by dominant noise components, these SOTA discriminative speech enhancement methods cannot effectively suppress noise without distorting or suppressing speech content, resulting in a significant decline in overall speech quality [9]. After conventional deep learning methods, researchers proposed using Mel Frequency Cepstral Coefficients (MFCCs) for the reconstruction of speech magnitude spectrum using Deep Neural Networks (DNNs) [10]. In addition, a GAN called DisCoGAN conditioned on latent features of a pre-trained discriminative model has been proposed for speech enhancement in low SNR scenarios [11]. In this approach, a SEANet-based generator is combined with a pre-trained DCCRN discriminative model. The proposed DisCoGAN outperforms other approaches in both Perceptual Evaluation of Speech Quality (PESQ) and SNR metrics, although its performance degrades under extremely low SNR conditions [12].

Later, researchers proposed a two-stage processing scheme using a Complex Spectral Mapping-based GAN (CSM-GAN) and a Convolutional-Recurrent Metric GAN (CRM-GAN) [13]. The performance of these models was tested with English, Chinese, French, German, Italian, and Russian, and the results were satisfactory except for word accuracy. Furthermore, authors in [14] developed a novel GAN-in-GAN framework that integrates two GAN models. The inner GAN performs spectrogram-to-spectrogram recovery to remove audio noise, with supervision provided by metric discriminators. The outer GAN performs audio-to-audio recovery, optimizing the final audio quality under multi-resolution discriminator supervision. The performance of this framework was evaluated using PESQ and Short-Time Objective Intelligibility (STOI).

Authors in [15] conducted a review on different GAN architectures proposed for small datasets, comparing the performance of models including pix2pixGAN, CycleGAN and SRGAN. Authors in [16] developed SepFormer, a transformer-based model for speech enhancement in complex noise environments. Authors in [17] proposed a SepFormer-based user-friendly toolkit, leveraging neural speech processing technology. Authors in [18] introduced a novel method using SepFormer to enhance audio quality in audio–video processing. Authors in [19] employed a magnitude STFT in SepFormer to handle long sequences for speech enhancement. Authors in [20] used a two-stage transformer-based model to reconstruct enhanced speech on benchmark datasets.

Speech enhancement techniques have made remarkable advancements with the development of deep learning. A more recent architecture is the transformer network, in which the significance of different segments of the input sequence is assessed through a self-attention mechanism. This allows long-range dependencies to be captured more efficiently and in parallel, resulting in substantial advancements.

In this work, the GAN model is used for generative tasks, where the goal is to produce enhanced or novel speech representations. Its adversarial framework enables learning a distribution of improved speech features from dysphonic inputs. In this setup, the GAN uses dysphonic MFCCs as a condition to generate reconstructed versions, making it suitable for scenarios where creative enhancement is desired without paired data. Another model is the SpeechBrain SepFormer, a pretrained transformer-based model originally designed for source separation and speech enhancement. It is adapted here to process dysphonic speech directly on raw waveforms, leveraging its pre-existing knowledge from large-scale training. SepFormer is particularly relevant for speech enhancement in noisy or degraded audio contexts. Its transformer architecture excels at modeling complex patterns, making it well-suited for improving dysphonic speech.

II. SPEECH ENHANCEMENT USING THE GENERATIVE ADVERSARIAL NETWORK

The proposed GAN model is implemented as shown in Figure 1, in which one-dimensional audio samples are recorded from a high-quality microphone to reduce distortions and background noise in a closed environment from subjects suffering from dysphonia. The developed dataset consists of 10 speakers with 10 different sentences in the Kannada language. Each sentence is recorded 10 times from each individual.

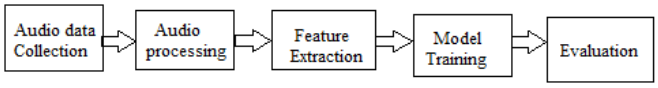

Fig. 1. Block diagram of dysphonic speech enhancement.

Audio preprocessing is performed using the Log-MMSE method to remove additive noise, such as background and microphone noise, present in the recorded signal. The preprocessed signal is then converted from the time domain to the frequency domain using STFT. The noisy STFT magnitude is calculated from the amplitude of each frequency component

over time. The initial noise frames of the magnitude signal contain primarily noise, and to estimate the noise spectrum, the mean magnitude across each frame from each frequency bin is calculated. A Wiener gain factor based on Log-MMSE is applied to the magnitude of the noisy signal to reduce noise, attenuating frequency components where noise is dominant. The denoised audio signal is then reconstructed using both magnitude and phase and the inverse STFT is used to convert it back to the time domain. The results are evaluated using the Mean Square Error (MSE), which measures the difference between the actual and predicted values by the model:

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y - \hat{y})^2 \quad (1)$$

where y is the actual value, $\hat{y}$ is the predicted value, and N is the number of data points.

After noise reduction, feature extraction is performed using MFCCs to transform raw audio data into a set of features and reduce the dimensionality or complexity of data, improving model performance. To extract features from the denoised signal, the audio file is resampled. Figure 2 shows the steps to obtain the MFCCs.

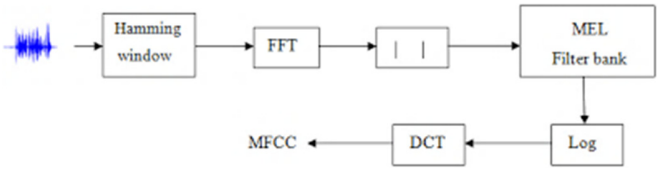

Fig. 2. Block diagram of feature extraction using MFCC.

After feature extraction, training is performed using the GAN model. The proposed GAN model and its internal architecture are shown in Figure 3. The GAN model is constructed by combining a generator and a discriminator and is trained by including adversarial labels. During training, the generator's weights are updated whereas the discriminator is set to be non-trainable. The discriminator is first trained to distinguish real and fake data. Then the generator is trained (with discriminator frozen) to produce data that can fool the discriminator. The goal of the discriminator is to distinguish between real and fake signals.

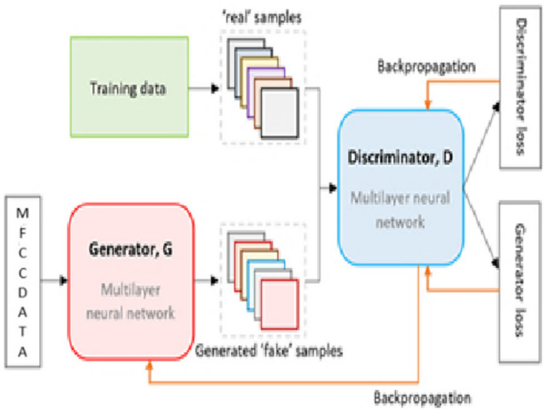

Fig. 3. Core components of the proposed GAN model.

The input passes through the generator to produce generated coefficients, which are then passed through the discriminator. The discriminator also receives the real data for compilation. During compilation, only the generator's weights are updated. The training progress accumulates and averages generator and discriminator losses and accuracy over an epoch, with periodic model checkpointing.

The difference between the original and reconstructed signals is quantified using objective metrics such as SNR, Mean Absolute Difference (MAD), and Mean Absolute Error (MAE) providing a numerical measure of the difference between the original and reconstructed signals. The generator loss (G_{loss}) and discriminator loss (D_{loss}) are calculated using the following equations:

At generator G:

$$G_{loss} = \log(1 - D(G(z))) \text{ or } -\log(D(G(z))) \quad (2)$$

The cost equations are given by:

$$\frac{1}{m}\sum_{i=1}^{m}\log(1 - D(G(z_i))) \quad (3)$$

$$\frac{1}{m}\sum_{i=1}^{m} -\log(D(G(z_i)) \quad (4)$$

At discriminator D:

$$D_{loss\ real} = \log(D(x)) \quad (5)$$

$$D_{loss\ fake} = \log(1 - D(G(z))) \quad (6)$$

$$D_{loss} = D_{loss\ real} + D_{loss\ fake} = \log(D(x)) + \log(1 - D(G(z))) \quad (7)$$

The cost functions are:

$$\frac{1}{M}\sum_{i=1}^{M}\log(D(x_i)) + \log(1 - D(G(z_i))) \quad (8)$$

$$\min_G \max_D \mathrm{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathrm{E}_{z\sim p_z(z)}[\log(1 - D(G(z))] \quad (9)$$

After training the GAN, its performance is compared with that of another deep learning model, SepFormer, which is explained in detail in the following section.

III. SPEECH ENHANCEMENT USING THE SEPFORMER MODEL

The SpeechBrain SepFormer is a pretrained transformer-based model originally designed for source separation and speech enhancement. It is adapted here to process dysphonic speech directly on raw waveforms, leveraging its pre-existing knowledge from large-scale training. Its transformer architecture, as shown in Figure 4, excels at modeling complex patterns, making it well-suited for dysphonic speech improvement.

The model encodes the input and processes it through a dual-path network (the mask net), which leverages transformer layers for both local and global context modeling. It then decodes the processed representation back into the separated audio signal. The network includes a feedforward subnetwork with linear transformations, activation functions, and layer normalization applied at different points. In particular, a

position-wise feedforward network applies two linear layers with ReLU activation and layer normalization.

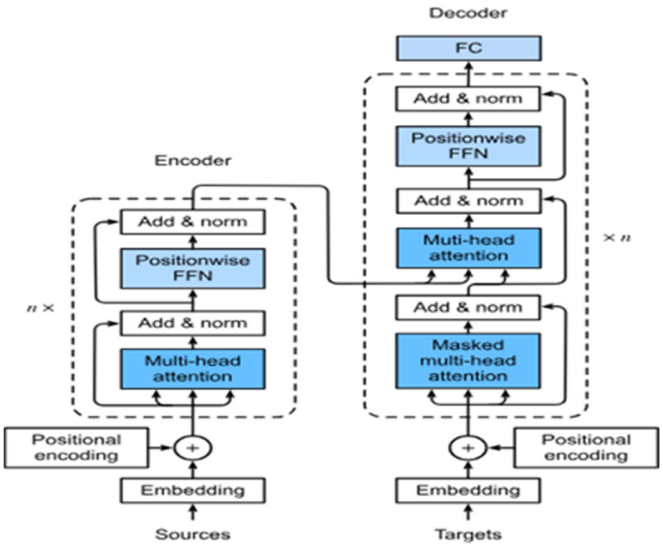

Fig. 4. Transformer architecture of the SepFormer model.

The preprocessed signal is passed through the SepFormer for the speech separation task. The network begins with convolutional layers for initial processing, followed by the dual-path transformer encoder layers for complex feature extraction and mask generation, and lastly a decoder to reconstruct the separated signal. Performance is evaluated by comparing the original and reconstructed audio signals using objective metrics.

IV. RESULTS AND DISCUSSION

From the private dataset of 100 samples, one sentence was selected for training and evaluation: "ನನ್ನ ಹೆಸರು ಅಭಿಷೇಕ" (My name is Abhisheka). The preprocessed result using Log-MMSE for this sentence is as shown in Figure 5, which includes the original, noisy, and denoised signal waveforms.

After preprocessing, feature extraction was performed using MFCCs. The GAN model was trained on 10 samples. Each sample was divided into 400 timesteps, and 13 features were extracted from each timestep. Out of the 10 samples, 8 were used for training as real samples, and 2 were used as testing samples for the generator, which were considered as fake samples.

Both generator and training data were fed to the discriminator model. Based on the difference between generator and discriminator outputs, the losses were calculated and fed back to the generator until the fake signals were converted into realistic samples. The original speech signal was then obtained after reconstruction. Table I illustrates the average generator and discriminator loss, along with the discriminator accuracy ($D_{accuracy}$) over two epochs. Lower discriminator loss indicates better discrimination between real and fake data, whereas lower generator loss implies better generator performance.

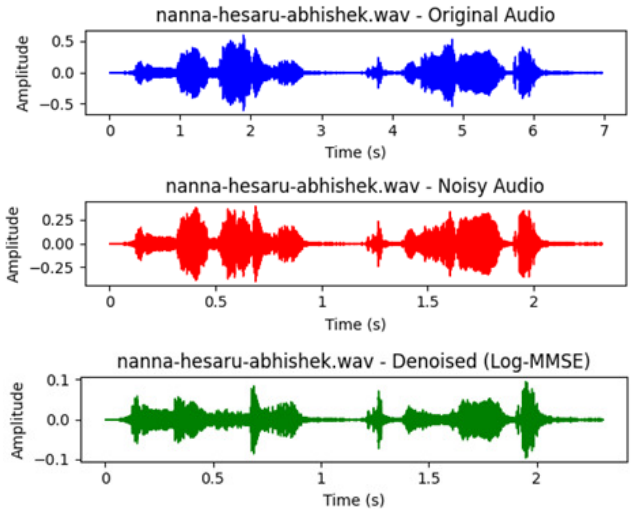

Fig. 5. Waveforms of the original, extracted noisy, and denoised audio signal using Log-MMSE for "ನನ್ನ ಹೆಸರು ಅಭಿಷೇಕ" (My name is Abhisheka) Kannada speech.

TABLE I. GAN MODEL LOSSES AND DISCRIMINATOR ACCURACY

Epochs	Avg D_{loss}	Avg G_{loss}	Avg $D_{accuracy}$ (%)
1	1.9948	0.0835	81.25
2	3.5585	0.2574	67.19

The performance was then evaluated using the SepFormer model. A pretrained SepFormer model was loaded, where the input signal is transformed into a higher dimensional representation by the encoder. The encoder consists of a convolution layer that downsamples the input data from 16 to 8 channels, which are then passed to the decoder. The decoder reconstructs the separated signal from the encoded representation, upsampling the data to their original dimension using another convolution layer. The original and enhanced audio waveforms are shown in Figure 6, the individual MFCCs and mean MFCC are shown in Figure 7, and the power spectral density is shown in Figure 8.

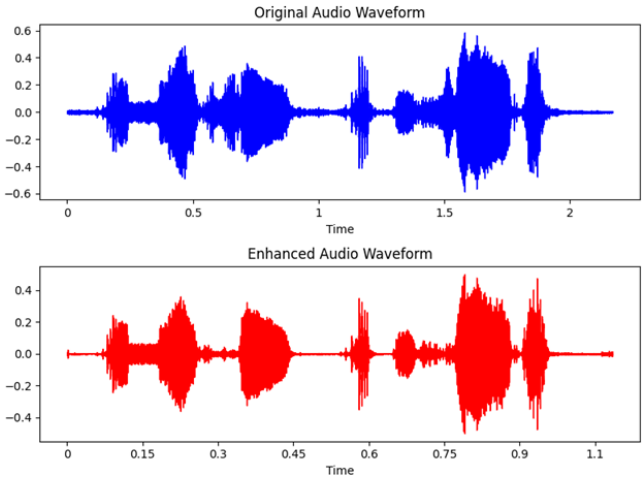

Fig. 6. Waveforms of the original and enhanced audio for the "ನನ್ನ ಹೆಸರು ಅಭಿಷೇಕ" speech signal.

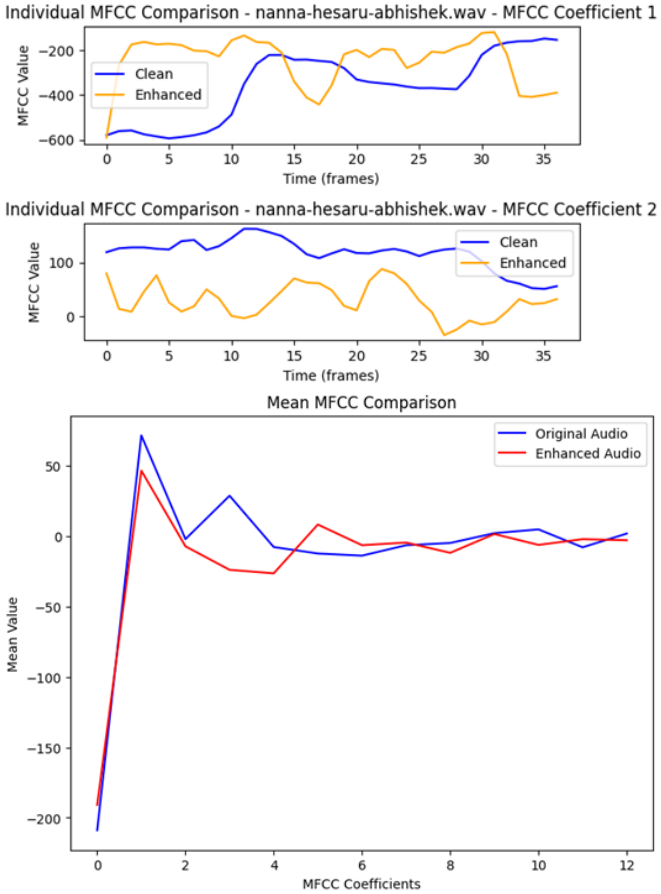

Fig. 7. Comparison of the individual MFCCs and mean MFCC for the original and enhanced "ನನ್ನ ಹೆಸರು ಅಭಿಷೇಕ" speech signal.

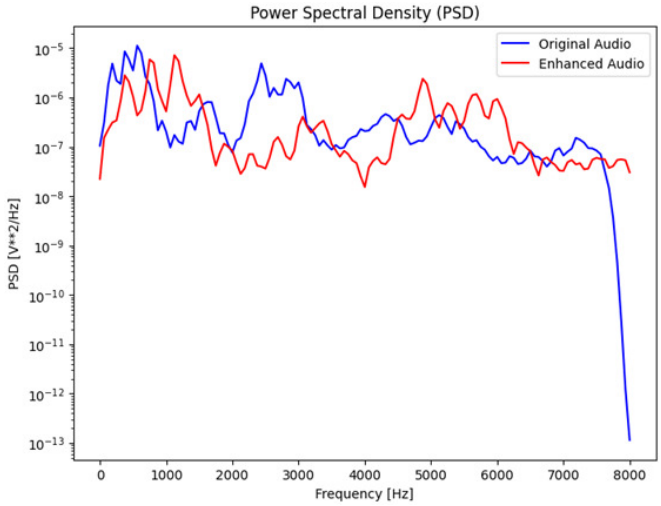

Fig. 8. Power spectral density of the original and enhanced "ನನ್ನ ಹೆಸರು ಅಭಿಷೇಕ" speech signal.

The performance of the GAN and SepFormer models was compared using objective metrics, including MSE, PESQ, STOI, SNR, and MAD, as shown in Table II. From the results, the SepFormer model outperforms the GAN model. PESQ and STOI scores approach the highest values in the standard ranges, and MSE, MAD, and SNR are also improved in the SepFormer model.

TABLE II. OBJECTIVE METRICS COMPARING SPEECH QUALITY BETWEEN GAN AND SEPFORMER MODELS

Metric	GAN model	SepFormer model
MSE	0.20	0.16
PESQ	1.228	4.233463
STOI	0.268	0.879939
SNR (dB)	−6.312	−1.328
MAD	0.136498	0.00643

V. CONCLUSION

People suffering from dysphonia find it difficult to convey their thoughts as intelligible speech. This work aims to convert dysphonic speech into audible, normal speech for the Kannada language. The dataset was developed by recording from dysphonic subjects, resulting in a total of 100 samples. The data were tested using a Generative Adversarial Network (GAN) and a SpeechBrain SepFormer models. Performance was evaluated using objective metrics, including Mean Square Error (MSE), Perceptual Evaluation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), and Signal-to-Noise Ratio (SNR).

The SepFormer model demonstrated superior performance compared to the GAN model. STOI and PESQ scores reached the highest values within standard ranges for SepFormer, indicating better intelligibility and speech quality of the reconstructed speech. Performance could improve further with the addition of more data samples to the dataset. Combining the GAN and SepFormer approaches may lead to even better results.

REFERENCES

[1] P. C. Loizou, *Speech Enhancement: Theory and Practice, Second Edition*, 2nd ed. Boca Raton, FL, USA: CRC Press, 2013, https://doi.org/10.1201/b14529.

[2] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Communication*, vol. 53, no. 4, pp. 465–494, Apr. 2011, https://doi.org/10.1016/j.specom.2010.12.003.

[3] I. J. Goodfellow *et al.*, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, Montreal, Canada, 2014, pp. 2672–2680.

[4] S. Pascual, J. Serrà, and A. Bonafonte, "Time-domain speech enhancement using generative adversarial networks," *Speech Communication*, vol. 114, pp. 10–21, Nov. 2019, https://doi.org/10.1016/j.specom.2019.09.001.

[5] L. Liu, H. Guan, J. Ma, W. Dai, G. Wang, and S. Ding, "A Mask Free Neural Network for Monaural Speech Enhancement," in *Interspeech 2023*, Dublin, Ireland, 2023, pp. 2468–2472, https://doi.org/10.21437/Interspeech.2023-339.

[6] H. Schröter, A. Maier, A. N. Escalante-B, and T. Rosenkranz, "Deepfilternet2: Towards Real-Time Speech Enhancement on Embedded Devices for Full-Band Audio," in *2022 International Workshop on Acoustic Signal Enhancement*, Bamberg, Germany, 2022, pp. 1–5, https://doi.org/10.1109/IWAENC53105.2022.9914782.

[7] H.-S. Choi, S. Park, J. H. Lee, H. Heo, D. Jeon, and K. Lee, "Real-Time Denoising and Dereverberation wtih Tiny Recurrent U-Net," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and*

Signal Processing, Toronto, Canada, 2021, pp. 5789–5793, https://doi.org/10.1109/ICASSP39728.2021.9414852.

[8] S. Zhao, B. Ma, K. N. Watcharasupat, and W.-S. Gan, "FRCRN: Boosting Feature Representation Using Frequency Recurrence for Monaural Speech Enhancement," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, Singapore, Singapore, 2022, pp. 9281–9285, https://doi.org/10.1109/ICASSP43922.2022.9747578.

[9] S. S. Shetu, S. Chakrabarty, O. Thiergart, and E. Mabande, "Ultra Low Complexity Deep Learning Based Noise Suppression," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, Seoul, South Korea, 2024, pp. 466–470, https://doi.org/10.1109/ICASSP48485.2024.10448353.

[10] J. Wenbin, L. I. U. Peilin, and W. E. N. Fei, "Speech Magnitude Spectrum Reconstruction from MFCCs Using Deep Neural Network," *Chinese Journal of Electronics*, vol. 27, no. 2, pp. 393–398, Mar. 2018, https://doi.org/10.1049/cje.2017.09.018.

[11] S. S. Shetu, E. A. P. Habets, and A. Brendel, "GAN-Based Speech Enhancement for Low SNR Using Latent Feature Conditioning," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing*, Hyderabad, India, 2025, pp. 1–5, https://doi.org/10.1109/ICASSP49660.2025.10890549.

[12] S. S. Shetu, E. A. P. Habets, and A. Brendel, "Comparative Analysis of Discriminative Deep Learning-Based Noise Reduction Methods in Low SNR Scenarios," in *2024 18th International Workshop on Acoustic Signal Enhancement*, Aalborg, Denmark, 2024, pp. 36–40, https://doi.org/10.1109/IWAENC61483.2024.10694283.

[13] Q. Hu, T. Tan, M. Tang, Y. Hu, C. Zhu, and J. Lu, "General Speech Restoration Using Two-Stage Generative Adversarial Networks," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops*, Seoul, South Korea, 2024, pp. 31–32, https://doi.org/10.1109/ICASSPW62465.2024.10625840.

[14] Y. Duan, J. Ren, H. Yu, and X. Jiang, "GAN-in-GAN for Monaural Speech Enhancement," *IEEE Signal Processing Letters*, vol. 30, pp. 853–857, 2023, https://doi.org/10.1109/LSP.2023.3293758.

[15] D. Habeeb, A. H. Alhassani, L. N. Abdullah, C. S. Der, and L. K. Q. Alasadi, "Advancements and Challenges: A Comprehensive Review of GAN-based Models for the Mitigation of Small Dataset and Texture Sticking Issues in Fake License Plate Recognition," *Engineering, Technology & Applied Science Research*, vol. 14, no. 6, pp. 18401–18408, Dec. 2024, https://doi.org/10.48084/etasr.8870.

[16] S. Liu, "Using Transformer Models to Separate and Reduce Noise in Speaker Voice in Noisy Conference Scenarios," in *2024 4th International Signal Processing, Communications and Engineering Management Conference*, Montreal, Canada, 2024, pp. 84–90, https://doi.org/10.1109/ISPCEM64498.2024.00021.

[17] M. Ravanelli *et al.*, "SpeechBrain: A General-Purpose Speech Toolkit." arXiv, June 10, 2021, https://doi.org/10.48550/arXiv.2106.04624.

[18] A. Nazemi, A. Sami, M. Sami, and A. Hussain, "Iterative Speech Enhancement with Transformers," in *3rd COG-MHEAR Workshop on Audio-Visual Speech Enhancement*, Kos, Greece, 2024, pp. 65–67, https://doi.org/10.21437/AVSEC.2024-14.

[19] D. de Oliveira, T. Peer, and T. Gerkmann, "Efficient Transformer-based Speech Enhancement Using Long Frames and STFT Magnitudes," in *Interspeech 2022*, Incheon, Korea, 2022, pp. 2948–2952, https://doi.org/10.21437/Interspeech.2022-10781.

[20] K. Wang, B. He, and W.-P. Zhu, "TSTNN: Two-Stage Transformer Based Neural Network for Speech Enhancement in the Time Domain," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing*, Toronto, Canada, 2021, pp. 7098–7102, https://doi.org/10.1109/ICASSP39728.2021.9413740.