2025 2nd International Conference on Research Methodologies in Knowledge Management, Artificial Intelligence and Telecommunication Engineering (RMKMATE) | 979-8-3315-9848-8/25/$31.00 ©2025 IEEE | DOI: 10.1109/RMKMATE64874.2025.11042606

# A Survey on Speech Enhancement Techniques for Improved Translation in Multilingual Acoustic Environments

Geethika Pula
*Department of CSE*
*Amrita School of Computing*
Amrita Vishwa Vidyapeetham
Chennai, India
geethikapula@gmail.com

Manikumar V S S S R
*Department of CSE*
*Amrita School of Computing*
Amrita Vishwa Vidyapeetham
Chennai, India
manikumarvalavala672003@gmail.com

Sasikala D
*Department of CSE*
*Amrita School of Computing*
Amrita Vishwa Vidyapeetham
Chennai, India
d_sasikala@ch.amrita.edu

***Abstract*—Speech-to-Speech Translation (S2ST) systems are inevitable tools in real-time communication to break language barriers. The design and development of an S2ST system is specifically for public meetings, where participation in the public meeting is typically multilingual. This paper discusses the latest developments in the S2ST system, specifically language translation and speech recognition, with a focus on various accents, dialectal variations, and contextual nuances, making translation accurate and natural. The work includes how domain-specific datasets improve quality in translation, neural models that can process more efficiently, and aims to make people accessible, provide access, and encourage multilingual collaboration in public forums. Future work will focus on expanding language support, continuing to lower latency even further, and introducing emotion-aware synthesis that is increasingly meant to capture an audience.**

***Index Terms*—Speech-to-Speech Translation (S2ST), Real-Time Communication, Multilingual Accessibility, Speech Recognition, Language Translation, Speech Synthesis, Public Meetings, Neural Networks.**

## I. Introduction

### A. Background

The world has become globalised,thus public meetings and conferences are essential to exchange ideas, collaborate, and innovate. However, one of the main problems is language, which in many events causes misunderstanding, disengagement, and missed opportunities for effective collaboration for people who speak different languages. This issue is even more highlighted in large public meetings, international forums, or governmental discussions with a wide variety of participants. With the increasingly interconnected world, the need for multilingual communication solutions has increased exponentially. The key in this is Speech-to-Speech Translation (S2ST) systems, which directly translate spoken language from one language to another. This solution is provided for the ability of people speaking in their mother tongue, but at the same time offering real-time translation to the other party involved in the conversation. This eliminates the need for human translation and interpretation services, which are costly, time-consuming, and prone to errors. Everyone, regardless of linguistic background, could be able to participate in public meetings and communicate more easily with the help of S2ST technology. In the context of parliamentary meetings, effective communication is necessary because speakers will discuss and debate their statements. Also, the setting consists of multilingual speakers. In this scenario, it is very important to translate spoken words accurately so that all participants can understand the context being conveyed, irrespective of language barriers. However, achieving accurate speech-to-speech translation is challenging here because the input will contain background noise and overlapping talks of multiple speakers, which could affect the quality of the audio signal.

Speech enhancement techniques are important in improving the quality of the input audio. It can reduce the background noise and segregate speakers, which can make the speech effective and suitable for translation systems to process. A major challenge in this parliamentary meeting is that the environment is unpredictable because speakers often interrupt each other and speak over each other. Also, when the meetings are being held in large spaces, the use of multiple microphones will lead to more noise unintentionally, which complicates maintaining clear audio. Furthermore, in a country like India with different dialects, people will have their accent, tone, and pronunciation of words, which makes it hard to understand equally by everyone. To deal with such situations, there is a need for techniques like noise reduction, echo cancellation, and beamforming so that the speaker's voice will be given priority rather than the accent.

Noise reduction will help filter out the unwanted signals from the audio, and echo cancellation will remove echoes that might distort the speech. Beamforming is another technique that focuses on only relevant sounds coming from a certain direction, like the speaker's voice, rather than focusing on sounds coming from all directions. These methods have been effective to some extent; however, in settings like parliamentary meetings, where the environment is dynamic, they might get complex.

In recent years, deep learning has emerged as an effective technique for speech enhancement. Traditional methods struggle when noise conditions vary, but deep learning models like convolutional neural networks (CNNs) and recurrent neural networks (RNNs) perform better in handling these complexities. This is because they can be trained on large datasets, enabling them to distinguish between speech and noise more effectively. So, these models can be utilised for diverse noise environments like parliamentary meetings, making them more reliable in real-time applications. This paper describes and analyses various approaches to speech enhancement, both traditional and deep learning technologies, in the context of parliamentary debates. It also investigates how these approaches work against a variety of datasets ranging from laboratory recordings to field audio transcripts. The analysis of each method contributes to the development of more effective speech-to-speech translation devices designed to assist in parliamentary meetings. By improving audio speech recognition systems, enhanced translations can be achieved, breaking down communication barriers and promoting better interactions in a multilingual society. This is particularly important in parliamentary democracies, where all relevant information must be accessible to members of parliament, ensuring no group is disadvantaged due to a lack of information.

In summary, this paper aims to explore and compare various speech enhancement techniques used in speech-to-speech translation systems, especially in the case of parliamentary meetings. By examining different methods and their applications, this survey seeks to contribute to the development of more effective and reliable translation systems that can handle the unique challenges of these environments.

## II. Related Works

In the current era, there has been a significant advancement in methods that are designed for improving speech clarity, especially in noisy environments, like parliamentary sessions or any public gatherings. The approaches reviewed in this paper come under two main categories: traditional signal processing techniques and deep learning methods. Each of these has its pros and cons, reliant on the context and data used. This section describes various techniques reviewed, how they differ in performance, and recaps the findings from several studies.

### *A. Traditional Signal Processing Methods*

The earlier approaches used for speech enhancement involved conventional signal processing methods such as noise reduction, linear predictive coding (LPC), and spectral subtraction. One of the first approaches proposed in speech enhancement by selective spectral filtering by Yecchuri et al. [1] involves using LPC and selective spectral subtraction to enhance the corrupted audio by eliminating a significant amount of environmental noise. A dataset that consists of 30 English speakers (10 male speakers, 10 female speakers, and 10 child speakers) of 200 ms of 60 CVC syllables was used. Complete elimination of signal energy in frequency ranges where speech energy is weakest was done by the employed approach. This approach limited its efficiency of speech enhancement to stationary noise. A similar multi-window spectral estimation was used in the work by Sekiguchi et al. [2], which focuses on classical and improved spectral subtractions and multi-window spectrum estimation algorithms in vehicle environments. The multi-window spectrum estimation algorithm significantly improved the quality of received speech and could be extended to handle heavy traffic noise. However, the operability of this approach is confined to the frequency domain with spatial information. Another proposed work involving spectral methods for speech enhancement is discussed by Kim et al. [3] in the subjective comparison and evaluation of speech enhancement algorithms.

A standardised approach involving spectral subtractive subspace and a statistical model-based Wiener filter was used on NOIZEUS data. The work proposed provided a common noisy speech corpus, and standardised subjective testing methodologies enable reliable comparison of speech enhancement algorithms. Subjective evaluation is only performed on a subset of the corpus (16 sentences) in four noise types. One of the recent approaches proposed in spectral analysis for automatic speech recognition and enhancement involves using the Short Time Fourier Transform (STFT) filtering technique and Adaptive Window Width based on the Chirp Rate (ASTFT) on the LibriSpeech ASR corpus in spectral analysis for automatic speech recognition and enhancement [4]. This work provides an approach to enhance and recognise speech elements in noisy environments by incorporating ASTFT, which is combined with various spectrogram features to optimise speech enhancement.

### *B. Deep Learning-Based Speech Enhancement*

Deep learning has driven forward all the traditional approaches in various fields. One such deep learning work was proposed in semi-supervised multichannel speech enhancement with deep speech prior [5]. The authors integrated a generative models of DNN for speech spectra with a NMF-model of noise spectra, and a full-rank spatial model in a unified probabilistic model. A dataset with 1320 noisy speech signals emulated to be uttered in four types of noisy environments, such as a bus, a cafe, a pedestrian area, and a street junction, was used. The proposed semi-supervised multichannel speech enhancement method replaces the low-rank assumption in existing models with a deep generative speech model. This model estimates spatial characteristics of speech and noise. An acoustic and adversarial supervision (AAS) method was proposed in unpaired speech enhancement by acoustic and adversarial supervision for speech recognition [6]. The author used a combination of LibriSpeech, DEMAND, and CHiME-4 for the work proposed and evaluated on WER and DCE. This method does not perform well under varying noise conditions and requires specific training or adaptation to different environments. A Convolutional Encode-Decoder (CED) and Recurrent Neural Networks (RNNs) deconvolu-

tional layer applied for waveform generation were proposed in the work titled End-to-End Deep Convolutional Recurrent Models for Noise Robust Waveform Speech Enhancement [7].

The proposed DL models for noise-robust waveform speech enhancement aim to improve the clarity and quality of speech signals by addressing challenges related to background noise. This approach integrates Complex Embedding Domain (CED) and Recurrent Neural Networks (RNNs) within a Complex Residual Network (CRN) framework. By combining these techniques, the model enhances speech quality and robustness against noise, delivering more accurate and intelligible audio outputs. This model poses critical challenges for real-time processing and usage due to model complexity and inference time reduction.

In a similar manner, various deep neural networks (DNN) were evaluated and compared with spectral methods and transformers in the work titled Deep Neural Network Techniques for Monaural Speech Enhancement and Separation: State-of-the-Art Analysis [8]. Various datasets with noisy speech mixtures, including reverberated and denoised speech, were used for training and evaluation. This work focuses primarily on monaural speech enhancement, excluding multichannel approaches. Various unsupervised learning techniques and domain adaptation for improved speech enhancement were explored. A unique combination of Deep Complex Convolution Recurrent Network (DCCRN), CNN, and LSTM was proposed and used in the work titled DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement [9]. The work introduced DCCRN to effectively handle complex-valued operations for improved speech enhancement. DCCRN achieves competitive performance with fewer parameters compared to existing models. This DCCRN model is limited to specific noise types and real-time processing constraints and fails to generalise to all types of noise. In the research work on Cross-Corpus Generalisation of Deep Learning Based Speech Enhancement [10], various techniques such as channel normalisation, better training corpus, and smaller frame shift are used to improve cross-corpus generalisation of DNN-based speech enhancement.

Deep learning-based speech enhancement models often struggle to generalize well to untrained speech corpora, primarily due to channel mismatch between the corpora. This issue arises when models trained on one dataset fail to perform optimally on a different dataset, largely because of differences in the acoustic conditions or recording channels. To address this challenge, the research titled ”Self-attending RNN for Speech Enhancement to Improve Cross-corpus Generalisation” [11] explores the use of self-attending recurrent neural networks (ARN) for enhancing audio signals.

The key innovation in this study is the integration of a self-attention mechanism within RNNs, which improves the model’s ability to capture long-range dependencies and contextual information within the speech signal. By incorporating self-attention, ARN enables RNNs to better focus on relevant parts of the speech signal, thereby improving speech enhancement performance, especially in cross-corpus scenarios. This approach represents a significant step forward in addressing the limitations of conventional deep neural networks (DNNs), which often struggle with such generalization issues.

### *C. Recurrent Neural Networks (RNNs) and Variants*

RNNs and its variants like Long Short-Term Memory (LSTM), Bi-Directional Long Short-Term Memory, and Gated Recurrent Units were widely applied in the field of speech processing because these networks can handle sequential data and long-term characteristics.Some methodologies inlcude the utilization of Convolutional Encode-Decoder (CED) and Recurrent Neural Networks (RNNs) Deconvolutional layer used for waveform generation as presented in CED and RNNs used for waveform generation. citesivaraman2022. In the research titled Gated Residual Networks with Dilated Convolutions for Monaural Speech Enhancement [12], they employed Gated Residual Networks (GRN) and Dilated Convolutions for enhanced speech intelligibility and quality. It is effective in generalising to untrained noises and speakers but is limited to specific noise types and speakers. In a similar study, long short-term memory for speaker generalisation in supervised speech separation [13], an LSTM-based model was used to improve speaker generalisation in noisy environments. LSTM effectively captures long-term dependencies for better speech separation. This LSTM model failed to specify unseen noise types and speakers but showcased its superiority over other conventional models. A unique combination of Deep Complex Convolution Recurrent Network (DCCRN), CNN, and LSTM was proposed and used in the work titled DCCRN: Deep Complex Convolution Recurrent Network for Phase-Aware Speech Enhancement [14]. The work introduced DCCRN to effectively handle complex-valued operations for improved speech enhancement. DCCRN achieves competitive performance with fewer parameters compared to existing models. This DCCRN model is limited to specific noise types and real-time processing constraints and fails to generalise to all types of noise.

Spectral masking learning with DNN and RNN combined with intelligibility improvement filter and reconstructive method was proposed in the work named On Learning Spectral Masking for Single Channel Speech Enhancement Using Feedforward and Recurrent Neural Networks [15]. Feedforward and recurrent neural networks are trained to learn spectral masking for single-channel speech enhancement, significantly improving speech quality, intelligibility, and automatic speech recognition performance.

### *D. Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs)*

A sub-convolutional U-Net (SCUNet) with a TAN mechanism for speech enhancement (TANSCUNet) is proposed and used in the research work titled Sub-convolutional U-Net with transformer attention network for end-to-end single-channel speech enhancement [16]. The TANSCUNet framework for single-channel speech enhancement utilises a sub-convolutional encoder-decoder model with varying kernel

sizes to capture multi-scale features, enhancing both local and global contextual information. It incorporates adaptive time-frequency attention modules and an adaptive hierarchical attention module to effectively manage long-term dependencies and contextual aggregation. Another approach involves using Convolutional Encode-Decoder (CED) and Recurrent Neural Networks (RNNs) with a deconvolutional layer applied for waveform generation, as proposed in Convolutional Encode-Decoder (CED) and Recurrent Neural Networks (RNNs). Deconvolutional layer applied for waveform generation [17]. The authors aimed to create a noise-robust waveform speech enhancement DL model to improve the clarity and quality of speech signals by addressing challenges related to background noise. This approach integrates Complex Embedding Domain (CED) and Recurrent Neural Networks (RNNs) within a Complex Residual Network (CRN) framework. This work includes phase estimation for speech magnitude and develops a more robust loss function for better results. A distinct neural cascade architecture with three modules and a triple-domain loss function was devised in the research work, Neural Cascade Architecture With Triple-Domain Loss for Speech Enhancement. [18].

The paper introduces a neural cascade architecture for speech enhancement, utilising three modules optimising magnitude spectrogram, time-domain signal, and complex spectrogram, showing superior performance in objective quality and intelligibility. The training ended with potential overfitting and difficulty in generalising to different types of noise that were not present in the dataset. In the research work named Speech Enhancement Using Multi-Stage Self-Attentive Temporal Convolutional Networks Temporal Convolutional Networks [19], Temporal Convolutional Networks (TCNs) were used with multistage learning and self-attention mechanisms to enhance the noisy audio files from the LibriSpeech and VCTK corpora. Multi-stage SA-TCN system with self-attention blocks and stacks of dilated TCN blocks to refine predictions. Multi-stage learning with self-attention and TCNs can effectively enhance speech quality.

### *E. Sub-Convolutional U-Net and Transformer Attention Models*

The TANSCUNet model is a combination of transformer attention with U-Net, which is used for time-frequency domain processing. This model is also proven as a powerful tool for single-channel speech enhancement. The authors evaluated this TANSCUNet on open-source datasets like Common Voice Corpus and NOIZEUS, and they proclaimed results as SDR 12.03 and STOI as 79.65%, showing dominating performance in dealing with the noisy environments. Also, this model has a few limitations when compared with state-of-the-art methods and while dealing with real-world situations. The work done by them ensured that the model would be tested on a broader set of datasets so that generalisation is improved.

### *F. Semi-Supervised and DNN-Based Approaches*

A notable study [20] introduced a new method for the speech enhancement process using deep neural networks (DNNs) in a semi-supervised manner, which will combine both traditional methods and deep learning techniques. The authors made use of the CHiME3 dataset. They were able to reduce background noise, and also, the latency of speech is higher when compared to older methods. This technique was proven to perform effectively even in diverse noise environments. Besides, there is another technique, which is acoustic and adversarial supervision (AAS), which was tested for Librispeech and DEMAND datasets. This AAS method improved speech accuracy. It was evaluated using metrics like word error rate (WER), which was low, making sure that it is suitable for complex environments.

### *G. Spectral Filtering and Traditional Enhancement Techniques*

In traditional signal processing, LPC and selective spectral subtraction are widely used for speech enhancement. So basically, these methods were aimed at improving metrics like SNR and SDR by filtering out unwanted noise signals. In one of the recent studies, the authors made use of LPC and reported improved SNR in custom datasets. However, when it comes to non-stationary noise environments, the effectiveness is limited because there will be multiple speakers speaking, and also background noise intervenes in the case of parliament.

### *H. Deep Autoencoders and Hybrid Architectures*

In the year 2020, a study made use of deep autoencoder (DAE) models for speech enhancement over the CHiME3 dataset, and it achieved an improvement of 3.17% in accuracy over traditional approaches. It showcased a promising noise reduction capability, which makes it fit for real-time speech enhancement-based applications. Also, some hybrid models integrate GRU-based classifiers with feedforward classifier architectures, which work under various noise conditions with greater capability. A recent study made use of this method and reported a hike in graphical patterns and SNR for the CHiME3 dataset. These models play a vital role in real-time speech-to-speech translation systems.

### *I. Convolutional Neural Networks and Temporal Envelope Reconstruction*

A newer method, which was published in 2023, has applied temporal envelope reconstruction, which makes use of 1-D Convolutional Neural Networks (CNNs). This paper focuses on reconstructing the temporal envelope of speech signals so that clarity can be improved. They worked on the WSJ0 dataset, from which they reported a 27% improvement in STOI, ensuring that CNN-based models can be effective in real-time speech enhancement. However, in conditions with low SNR, these CNNs often struggle. Future works are being focused on robust architectures like LSTM and GRU so that even in acoustic environments, the performance would be better.

## III. Thematic Framework

While looking for optimal ways to translate speech at parliamentary meetings, researchers use various methods to improve the quality of audio signals. These methods can be segregated into two broad categories: traditional signal processing techniques & advanced deep learning models.

### *A. Traditional Signal Processing Techniques*

Traditional techniques are great for boosting speech quality. They focus on noise reduction, echo cancellation, and beamforming. These tools work well in quiet spaces but lack in performance in environments like parliamentary meetings packed with voices, echoes, & noise everywhere.

**Noise Reduction** Background sounds can distort what a speaker is trying to convey. Techniques such as spectral subtraction and Wiener filtering help enhance speech. Studies depict these techniques work accurately in static environments, but they often become challenging in real-life situations, like public meetings where noise levels vary.

**Echo Cancellation** In parliamentary meetings, people usually gather in large numbers that might echo a lot. Echo cancellation algorithms step in to get rid of the pesky sound reflections. Traditional methods, like adaptive filters, work significantly, but they might lose their effectiveness when echoes start changing quickly.

**Beamforming** Beamforming is about grasping speech from one direction, usually where the speaker stands. This method helps focus on the person talking while silencing out background noise & other voices interlacing at the same time. It can achieve considerable success when using multiple microphones; however, it still encounters challenges when speakers move around or when multiple people are speaking simultaneously.

So, while these traditional methods are helpful, they do face some real challenges in lively places. Understanding & improving how they function helps for better translations.

### *B. Deep Learning Techniques*

As traditional signal processing techniques face complexities in real-world acoustic environments, there is a need for robust methods in speech enhancement. And for that, deep learning models would be the right choice.

**Convolutional neural networks** CNN models can identify the pattern of spectrograms and isolate the background noise from the original audio much more effectively. For instance, the U-Net ResBiLSTM Transformer-net has been applied to the WSJ0 dataset, achieving a 0.6 dB increase in SI-SDR, demonstrating the effectiveness of CNN-based architectures for speech enhancement.

**Recurrent Neural Networks (RNNs) and Variants** Apart from convolutional models, RNNs, along with Long Short-Term Memory (LSTM) and Bidirectional LSTM (BiLSTM), have performed optimally in improving audio signals by capturing temporal dependencies. For example, applying the BiLSTM model to the LibriSpeech dataset results in a high SNR, indicating that the speech enhancement remains accurate even in real-time environments such as parliament sessions.

**Transformers and Attention Mechanisms** The use of transformer models in combination with RNNs and CNNs has produced improvements in speech enhancement. The Dual-Path Transformer Network (DPTNet), for example, demonstrated the model's potential in multi-speaker environments by achieving a 20.6 dB SDR when applied to the WSJ0-2mix dataset.

**Generative Models** For speech enhancement, researchers applied Variational Autoencoders (VAEs) or Generative Adversarial Networks (GANs), and they mentioned that these models can generate clean speech from noisy input signals. As proof, the Conditional Generative Framework Neural Speech Codec (TF-Codec) demonstrated decent scores with SIG, BAK, and OVR scores of 4.27, 4.63, and 4.09, respectively.

## IV. Dataset

Generally, researchers work on multiple datasets to determine how effectively the model performs. These datasets will range from controlled environments to noisy environments. They try to simulate various real-world conditions.

**WSJ0 Dataset** The WSJ0 dataset [21] is an open-source dataset that is mostly used for isolating speech and task enhancement. It contains recordings of 123 speakers for 141 hours. Since this is a vast dataset, techniques like Deep Neural Networks (DNNs), transformers, and RNNs were utilised to assess this dataset, achieving noteworthy improvements in metrics like Signal-to-Distortion Ratio (SDR) and Short-Time Objective Intelligibility (STOI). For example, a study employing Gated Residual Networks (GRN) reported an STOI of 22.18 and a PESQ of 3.01.

**LibriSpeech Dataset** This dataset is used by researchers for training and evaluating speech enhancement models [22]. Methods like BiLSTM and convolutional encoder-decoder architectures have achieved substantial improvements in quality and intelligibility metrics, as evidenced by the high SNR achieved by these methods.

**TIMIT and NOISEX-92** These datasets are particularly useful for evaluating models in noisy environments [23], [24]. Spectral masking with DNNs and RNNs reported an improvement of 17.6% in STOI and a 5.22 dB increase in SDR when tested on TIMIT and NOISEX-92.

## V. Performance Metrics

Several key metrics are used to evaluate the performance of speech enhancement models:

**Signal-to-Distortion Ratio (SDR)** SDR measures the quality of the enhanced speech signal in terms of the ratio of the desired signal to the distortion. High SDR values indicate better performance. For instance, DPTNet achieved an SDR of 20.6 dB, setting a new benchmark for speech enhancement.

**Short-Time Objective Intelligibility (STOI)** STOI is a measure of how intelligible the enhanced speech is to human listeners. Techniques like BiLSTM have reported STOI improvements of over 17%, demonstrating their effectiveness in making speech clearer.

**Perceptual Evaluation of Speech Quality (PESQ)** PESQ is an objective metric that assesses speech quality by comparing the enhanced signal with a reference signal. Techniques like DCCRN achieved a PESQ score of 3.50, highlighting the potential of deep learning models in improving speech quality.

## VI. Discussions

Despite achieving great results in the area of speech enhancement for parliamentary meeting scenarios, there are still a few problems that remain unsolved. One of them is how to process real-time speech in the presence of changing acoustic conditions resulting from dynamic, noisy environments with multiple possibly concurrent speakers. Such scenarios introduce variability in the acoustic conditions and impede effective generalisation. Overlapping speech, that is, the case of two or more people speaking at the same time, is another significant issue with current systems. This is known as the ”cocktail party effect,” and isolating and translating the individual voices and thus speech-to-speech translation quality becomes a very challenging task. Besides this, datasets currently used do not represent the practical aspects of real acoustic environments but focus on relatively controlled situations. This may cause a mismatch of the training data to real-world conditions, which can further contribute to suboptimal performance in live applications.

Moreover, computational cost and accuracy are traded off, especially in deep learning models. Although some models provide excellent performance metrics in controlled settings, they usually consume many computational resources, hence not viable for real-time applications. It is a persistent struggle in developing these systems to maintain high accuracy while maintaining low-latency processing. Traditionally, though they are good for controlled environments, they lack power in dynamic, noisy scenarios.

Deep learning techniques hold great promise but still need improvement on the handling of different acoustic environments, real-time processing, and multiple speaker cases. The future of research work will be in hybrid models that combine the strengths of traditional signal processing with deep learning models. More significantly, larger datasets, especially those that specifically elucidate the challenges of parliamentary meetings, will be needed to train and test models appropriately. A major focus area would be low-latency models: how these can operate in real time without diminished performance to permit a deployment in live settings.

## VII. Future Scope

Hybrid models represent the future of speech enhancement research. These models combine the precision of traditional signal processing techniques with the flexibility and adaptability of modern deep learning architectures. Using the fine-tuned capabilities of traditional methods, particularly for tasks such as noise reduction and echo cancellation, alongside the strengths of deep learning, future models could overcome many of the limitations faced by current approaches. Another promising direction is the integration of visual cues, such

TABLE I
Summary of Key Speech Enhancement Techniques and Results

| Study | Technique Used | Dataset | Performance Metrics | Key Findings |
|---|---|---|---|---|
| TANSCUNet with Transformer Attention | U-Net architecture with Adaptive Time-Frequency Attention | Common Voice Corpus, NOIZEUS | SDR: 12.03, STOI: 79.65%, PESQ: 2.73 | Significant improvement in speech clarity in noisy environments, effective for single-channel enhancement |
| Semi-Supervised Multichannel Speech | DNN-based generative model, Spectral Masking | CHiME3, Librispeech | PESQ, SDR, STOI | Achieved strong generalization to noisy speech, outperformed traditional methods |
| Acoustic and Adversarial Supervision | Adversarial learning model for unpaired speech enhancement | Librispeech + DEMAND, CHiME-4 | WER, DCE | Improved generalization and reduced WER in noisy environments |
| Speech Enhancement via Spectral Filtering | LPC, Selective Spectral Subtraction | Custom Dataset | SNR | Improved SNR, though less effective in non-stationary noise conditions |
| Dual-Path Transformer Network | Transformer with RNN for speech enhancement | WSJ0-2mix | SDR: 20.6 dB | Outperformed state-of-the-art methods, especially effective for multi-speaker environments |
| Deep Autoencoder (DAE) | Deep Autoencoder for speech denoising | CHiME3 | Accuracy: +3.17%, SNR | Strong noise reduction and clarity improvements, effective in low-SNR conditions |
| Temporal Envelope Reconstruction | 1D CNN-based temporal envelope reconstruction for speech enhancement | WSJ0 | STOI: +27%, Speech Spectrum Shaping | Significant intelligibility improvement, but challenges in very low-SNR conditions |
| Conditional Generative Framework | GAN-based model with auto-regressive Transformer for speech codec enhancement | Librispeech ASR Corpus | SIG: 4.27, BAK: 4.63, OVR: 4.09 | Strong performance in subjective quality metrics, effective in multi-noises. |

as lip movements, directly into the models. This multimodal approach could significantly improve the ability to distinguish speech from background noise, making it especially valuable in noisy environments, such as parliamentary meetings where speakers often face a variety of background conditions.

Yet another area in which personalised speech enhancement models come into great promise is in multi-accent and multi-lingual environments. Such models would learn an individual's accent, speech patterns, and the environment in which the person speaks so that translation would be more accurate and speech clarity would be better. Real-time speech enhancement systems will also be a key area of attention. The spread of applications for real-time speech-to-speech translation in live settings will require improvement in latency while maintaining accuracy at high levels. Future work will include the design of more efficient architectures and optimisation techniques that should be able to balance these competing requirements much better.

## VIII. Conclusion

A comparative study with the current state-of-the-art system is necessary to understand developments in recent speech enhancement techniques. Recent deep learning models, such as TANSCUNet and Dual-Path Transformer Networks, have outperformed other techniques concerning SDR, STOI, and PESQ in comparison to traditional benchmarking. For instance, TANSCUNet outperformed classical techniques, such as LPC and Wiener filtering, with a performance that failed to do something appropriately in an environment of complexity. The success of the previous models also occurred in DPT-Net, which is known as the Dual-Path Transformer Network. DPTNet has been able to obtain an SDR of 20.6 dB on the WSJ0-2mix dataset; these values are far from surpassing existing state-of-the-art methods. The supremacy of these transformer-based models underscores shifting sands under which speech enhancement techniques come in environments such as parliamentary meetings, inherently posing several challenges in terms of multi-speaker environments and the overlap of speech.

However, their utility remains strong in very specific applications. Techniques such as spectral subtraction, echo cancellation, and beamforming may offer better computational efficiency; hence, they are apt for much simpler environments, where performance in real time becomes very important. One major trend in the field is the merging of traditional approaches and deep learning to realise hybrid systems that offer the best of both worlds.

## References


[1] S. Yecchuri and S. Vanambathina, "Sub-convolutional u-net with transformer attention network for end-to-end single-channel speech enhancement," *J AUDIO SPEECH MUSIC PROC*, vol. 8, 2024.

[2] K. Sekiguchi, Y. Bando, A. A. Nugraha, K. Yoshii, and T. Kawahara, "Semi-supervised multichannel speech enhancement with a deep speech prior," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 27, no. 12, pp. 2197–2212, 2019.

[3] G.-m. Kim *et al.*, "Unpaired speech enhancement by acoustic and adversarial supervision for speech recognition," *IEEE Signal Processing Letters*, vol. 26, pp. 159–163, 2018.

[4] D. O'Shaughnessy, "Speech enhancement by selective spectral filtering," *The Journal of the Acoustical Society of America*, vol. 87, no. S1, pp. S104–S104, 1990.

[5] B. Dendani, H. Bahi, and T. Sari, "Speech enhancement based on deep autoencoder for remote arabic speech recognition," in *Image and Signal Processing: 9th International Conference, ICISP 2020*. Marrakesh, Morocco: Springer International Publishing, 2020, pp. 221–229.

[6] T. S. Sarika, S. Sreekumar, and A. G. Hari Narayanan, "Enhancement of speech recognition (voice quest)," *International Journal of Applied Engineering Research*, vol. 10, pp. 708–711, 2015.

[7] C. Wang, Y. Li, and H. Lu, "Speech enhancement algorithms in vehicle environment," *International Journal of Performability Engineering*, vol. 15, no. 11, p. 3081, 2019.

[8] J. Benesty *et al.*, "A brief overview of speech enhancement with linear filtering," *EURASIP Journal on Advances in Signal Processing*, pp. 1–10, 2014.

[9] M. N. Pradeep and M. Suresh, "Speech enhancement-adaptive algorithms," in *2024 International Conference on Smart Systems for applications in Electrical Sciences (ICSSES)*. IEEE, 2024.

[10] R. Soleymanpour *et al.*, "Speech enhancement algorithm based on a convolutional neural network reconstruction of the temporal envelope of speech in noisy environments," *IEEE Access*, vol. 11, pp. 5328–5336, 2023.

[11] R. Ullah *et al.*, "End-to-end deep convolutional recurrent models for noise robust waveform speech enhancement," *Sensors*, vol. 22, no. 20, p. 7782, 2022.

[12] T. Grzywalski and S. Drgas, "Speech enhancement by multiple propagation through the same neural network," *Sensors*, vol. 22, no. 7, p. 2440, 2022.

[13] H. Wang and D. Wang, "Neural cascade architecture with triple-domain loss for speech enhancement," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 734–743, 2021.

[14] T.-A. Hsieh *et al.*, "Wavecrn: An efficient convolutional recurrent neural network for end-to-end speech enhancement," *IEEE Signal Processing Letters*, vol. 27, pp. 2149–2153, 2020.

[15] Y. Hu and P. C. Loizou, "Subjective comparison and evaluation of speech enhancement algorithms," *Speech Communication*, vol. 49, no. 7-8, pp. 588–601, 2007.

[16] K.A. Mohamed Junaid, T. Sethukarasi, M. Vigilson Prem, Adi Alhudhaif, Norah Alnaim, "A novel efficient Rank-Revealing QR matrix and Schur decomposition method for big data mining and clustering (RRQR-SDM)", Information Sciences, Volume 657, 2024, 119957, ISSN 0020-0255, https://doi.org/10.1016/j.ins.2023.119957.

[17] P. Ochieng, "Deep neural network techniques for monaural speech enhancement and separation: state of the art analysis," *Artificial Intelligence Review*, vol. 56, no. Suppl. 3, pp. 3651–3703, 2023.

[18] A.K. Reshmy,"Data Mining of Unstructured Big Data In Cloud Computing" International Journal of Business Intelligence and Data Mining, Vol.13 Issue 1-3 2017

[19] Praveen D S., "Generative Adversarial Networks (GAN) and HDFS-Based Realtime Traffic Forecasting System Using CCTV Surveillance." Symmetry 2023, 15, 779. https://doi.org/ 10.3390/sym15040779.

[20] S. Kim, M. Maity, and M. Kim, "Incremental binarization on recurrent neural networks for single-channel source separation," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*. IEEE, 2019, pp. 376–380.

[21] S. Maharjan, "Wsj0-2mix dataset," https://www.kaggle.com/datasets/sonishmaharjan555/wsj0-2mix, 2020, accessed: 2025-04-27.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech asr corpus," https://www.openslr.org/12, 2015, accessed: 2025-04-27.

[23] L. D. Consortium, "Timit acoustic-phonetic continuous speech corpus," https://catalog.ldc.upenn.edu/LDC93S1, 1993, accessed: 2025-04-27.

[24] S. Ancy,"Online Learning Model For Handling Different Concept Drifts Using Diverse Ensemble Classifiers," cybernetics and Systems", Vol.50 Issue 7, 2019