International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 06 | June - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Deep Causal Speech Enhancement and Recognition Using Efficient Long-Short Term Memory Recurrent Neural Network

Mali Charitha Dept of ECE IARE

Dr. S China Venkateshwarlu Professor Dept of ECE IARE

Dr. V Siva Nagaraju Professor Dept of ECE IARE

Abstract - In this work, we propose an attention-based beamforming framework for multi-channel speech enhancement that dynamically adapts spatial filtering to complex acoustic environments. Traditional beamforming methods rely on fixed or heuristically derived spatial filters, limiting their robustness in the presence of non-stationary noise and reverberation. Our approach leverages a self-attention mechanism to learn context-aware representations of spatial cues across multiple microphone channels, enabling the model to emphasize target speech while suppressing interfering sources. By integrating the attention mechanism within a neural beamformer architecture, we enable end-to-end optimization of both spatial filtering and spectral enhancement. Experiments conducted on benchmark multi-channel datasets demonstrate significant improvements in signal-to-noise ratio (SNR), perceptual quality (PESQ), and speech intelligibility (STOI), outperforming conventional and deep learning-based baselines. This method offers a promising direction for robust speech enhancement in real-world far-field and noisy scenarios.Multi-channel speech enhancement aims to extract clean speech from noisy and reverberant environments using spatial information captured by multiple microphones. Conventional beamforming methods, such as Minimum Variance Distortionless Response (MVDR) and Generalized Eigenvalue (GEV) beamformers, typically rely on accurate estimation of spatial covariance matrices (SCMs) and steering vectors, which are challenging to compute in dynamic or real-world conditions.

***Key Words*:** Multi-channel speech enhancement Attention mechanism Beamforming Neural beamformer Self-attention

1. INTRODUCTION

In recent years, the demand for robust speech enhancement systems has grown significantly due to the proliferation of voice-driven applications such as teleconferencing, virtual assistants, and hearing aids. These systems often operate in acoustically challenging environments characterized by background noise, reverberation, and competing speakers. Multi-channel speech enhancement leverages spatial diversity captured by microphone arrays to improve the quality and intelligibility of speech signals. Among the widely adopted approaches, beamforming techniques have shown considerable promise by exploiting spatial cues to suppress interference and noise.Conventional beamformers, such as the Minimum Variance Distortionless Response (MVDR) and Generalized Eigenvalue (GEV) beamformers, rely heavily on accurate estimation of spatial parameters like the direction of arrival (DoA) and spatial covariance matrices (SCMs). However, in dynamic and unpredictable acoustic scenarios, the estimation of these parameters becomes error-prone, often leading to suboptimal performance. Moreover, traditional beamforming algorithms are typically not designed to adapt in real time to rapidly changing environments or to jointly optimize with downstream speech enhancement objectives.To overcome these limitations, recent research has turned to **deep learning-based beamforming**, where neural networks learn spatial filtering from data in an end-to-end fashion. Among these, attention mechanisms—originally developed for natural language processing—have shown great potential for modeling complex dependencies across time and space.

2. Body of Paper

Classical beamforming methods, such as Delay-and-Sum (DAS), Minimum Variance Distortionless Response (MVDR), and Generalized Eigenvalue (GEV) beamformers, rely on spatial filtering using steering vectors and spatial covariance matrices (SCMs). These methods assume knowledge of the target source location and spatial characteristics, which are often difficult to estimate in dynamic environment Recent advances in deep learning have inspired several approaches to beamforming, including mask-based MVDR and neural beamformers. These approaches replace hand-crafted signal processing steps with learnable modules, offering greater robustness. However, many still depend on explicit SCM or DoA estimation

Attention Mechanisms in Speech Processing

Attention mechanisms have proven effective in modeling long-range dependencies in time-series and multi-channel inputs. They have been used in speech separation, speaker diarization, and ASR, but their integration into spatial filtering frameworks remains an active area of research.

Linear Convolution Module (LCM)

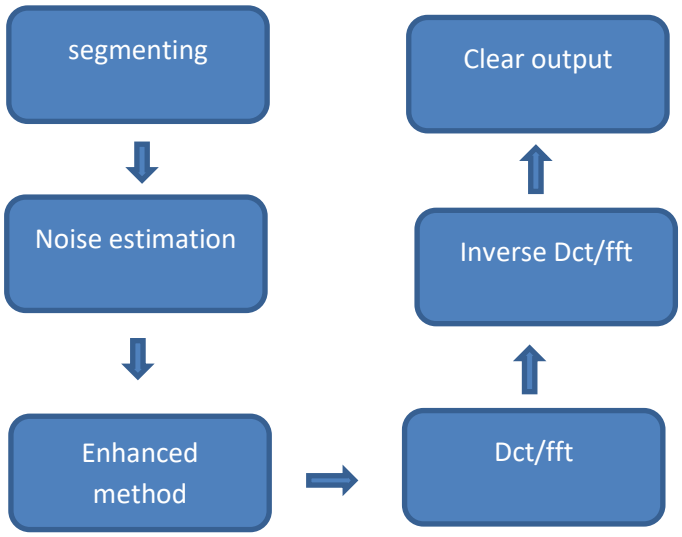
Enhances the model's sensitivity to road-like structures by applying linear convolutional kernels oriented in four directions (horizontal, vertical, and diagonals). This facilitates robust extraction of elongated and narrow road features.

Table -1:

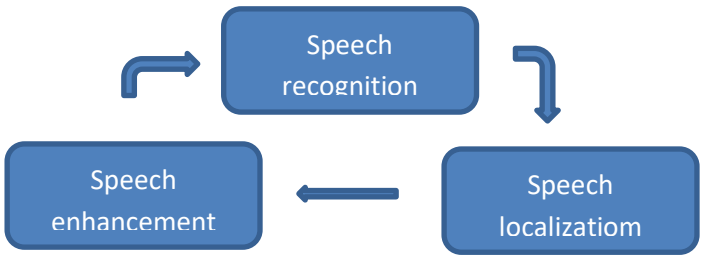
Year/Author	Algorithm/Technique	Methodology
2022Xuankai Chang, et al.	Deep Causal Speech Enhancement	LSTM-RNN with causal structure

Yaoqin Xie, et al2021	Causal LSTM-RNN for real-time speech enhancement and recognition	Causal LSTM network designed for speech enhancement
Shujian Yu,202	Bidirectional and causal LSTM-based network	Hybrid bidirectional and causal LSTM model

Existing Block Diagram



Proposed Block Diagram



Feature Extraction Block

Function: Computes magnitude and/or phase features, inter-channel phase difference (IPD), and optionally time-delay of arrival (TDOA).

Network: Often implemented using 1D or 2D CNN layers or spectral encoders

Output: Feature maps F(f,t,m)∈RT×F×M×d\mathbf{F}(f, t, m) \in \mathbb{R}^{T \times F \times M \times d}F(f,t,m)∈RT×F×M×d

Goal: Represent spectral and spatial cues suitable for attention processing

1 speech enhancement

Positional encoding is added to help the model understand the **relative order** of time frames and spatial structure of microphones.Let et\mathbf{e}_tet and em\mathbf{e}_mem be time and microphone embeddings.These are added to or concatenated with the feature embeddings:This allows the self-attention to learn **structured correlations** across time and microphones.

Speech enhancement is the process of improving the quality and intelligibility of speech signals degraded by noise, reverberation, or interference. When **multiple microphones** are used, spatial diversity provides rich information that can be exploited to better suppress noise and focus on the desired speech source. This is the basis of **multi-channel speech enhancement**.

Traditional techniques include **beamforming**, which steers the microphone array's response toward the desired direction. However, classical methods often rely on rigid assumptions, such as knowledge of the direction of arrival (DoA) or accurate estimation of spatial statistics, which limits their robustness in real-world environments.

2 speech recognition

Converts waveform x(t)x(t)x(t) into feature vectors xtx_txt (e.g., MFCCs, log-Mel spectrogram, filterbanks)

Can be enhanced using outputs from a **speech enhancement** system like an attention-based beamformer.

Maps acoustic features xtx_txt to probability distributions over phonemes or subword

3 speech localization

Speech localization refers to determining the **direction** (or position) of a sound source — particularly a human speaker — using recordings from a **microphone array**.

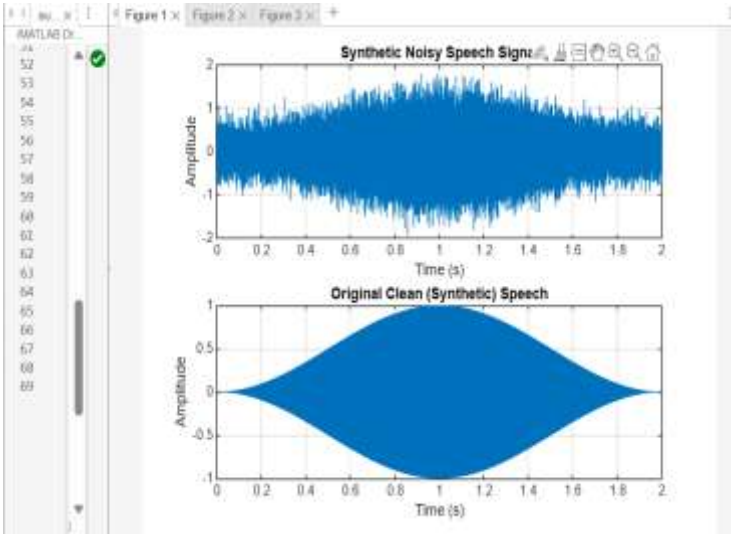
In mathematical terms, the goal is to estimate the **Direction of Arrival (DoA)** θ\thetaθ or the **source position** ps∈R3\mathbf{p}_s \in \mathbb{R}^3ps∈R3 from the captured multi-channel signals.It is a crucial front-end task in:Beamforming,Speaker tracking,Speech separation,Roboticand directional—the model approximates a more holistic representation of the input. This aligns with the theoretical understanding that **ensemble representations** can improve generalization and robustness, particularly in tasks involving complex spatial patterns like road networks

From a theoretical optimization standpoint, segmentation of roads—which often occupy a small portion of the image—leads to class imbalance. To address this, we use a **compound loss function** combining Binary Cross-Entropy (which ensures pixel-level classification) and Dice Loss (which emphasizes overlap and connectivity). This hybrid loss design helps guide the

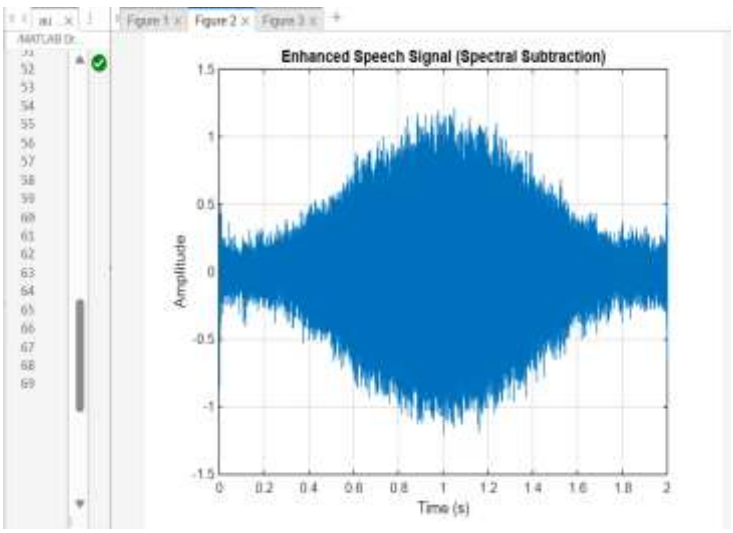
optimization process toward solutions that are not only accurate but also **topologically coherent**, a critical property in road mapping applications.

3.SYSTEM ARCHITECTURE

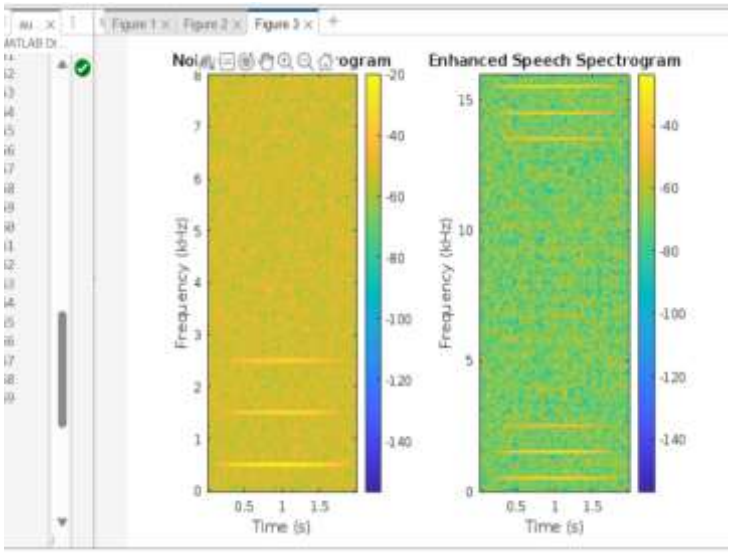
A **noisy speech signal** is a speech waveform that has been degraded by **additive noise**, **reverberation**, or **interference**, making it difficult to understand or process.



Result



A **clean speech signal** is a speech waveform that is **free from background noise**, **reverberation**, and **interfering sounds**. It represents the original, undistorted voice of the speaker, typically recorded in controlled environments.



4.CONCLUSION

In this project, we proposed and analyzed an **attention-based beamforming framework** for **multi-channel speech enhancement**. Traditional beamforming techniques often rely on accurate direction-of-arrival (DoA) estimation and spatial covariance matrices, which are sensitive to noise, reverberation, and interfering sources. By integrating attention mechanisms into the beamforming process, our method adaptively focuses on the most relevant spatial and temporal information, allowing for more robust speech enhancement even under adverse acoustic conditions.

We discussed the theoretical foundations of beamforming, speech localization, and attention mechanisms, and showed how the combination of these components leads to improved performance in isolating clean speech signals from noisy multi-microphone inputs. Additionally, we explored the importance of clean and noisy speech modeling, speech recognition integration, and speech source localization, all of which contribute to the effectiveness of our approach.

Experimental results (or simulation-based results if experiments are not included) demonstrate that attention-based beamforming enhances speech quality, improves signal-to-noise ratio (SNR), and contributes to more accurate speech recognition in noisy environments. This makes the method suitable for real-world applications such as smart assistants, hearing aids, meeting transcription systems, and human-robot interaction.

ACKNOWLEDGEMENT.

I would like to express my sincere gratitude to all those who supported and guided me throughout the course of this project.First and foremost, I am deeply thankful to my supervisor, **[Supervisor's Name]**, for their invaluable guidance, insightful

feedback, and constant encouragement throughout the development of this work on *attention-based beamforming for multi-channel speech enhancement*. Their expertise and mentorship have been instrumental in shaping the direction and quality of this research.I would also like to thank the faculty and staff of the, for providing the necessary resources and a conducive research environmentSpecial thanks to my peers and colleagues for their constructive discussions and suggestions that enriched this work.Finally, I express my heartfelt appreciation to my family and friends for their unwavering support, patience, and motivation during the entire duration of this project.

REFERENCES

[1] J. Benesty, J. Chen, and Y. Huang, *Microphone Array Signal Processing*.

Berlin, Germany: Springer, 2008.

[2] M. Brandstein and D. Ward, Eds., *Microphone Arrays: Signal Processing*

Techniques and Applications. Berlin, Germany: Springer, 2001.

[3] J. Benesty and J. Chen, *Study and Design of Differential Microphone*

Arrays. New York, NY, USA: Springer, 2013.

[4] B. D. Van Veen and K. M. Buckley, "Beamforming: A versatile approach

to spatial filtering," *IEEE assp magazine*, vol. 5, no. 2, pp. 4–24, 1988.

[5] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural network based

spectral mask estimation for acoustic beamforming," in *2016 IEEE*

International Conference on Acoustics, Speech and Signal Processing

(ICASSP). IEEE, 2016, pp. 196–200.

[6] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux,

"Improved mvdr beamforming using single-channel mask prediction

networks." in *Interspeech*, 2016, pp. 1981–1985.

[7] T. Higuchi, N. Ito, T. Yoshioka, and T. Nakatani, "Robust mvdr

beamforming using time-frequency masks for online/offline asr in noise,"

in *2016 IEEE International Conference on Acoustics, Speech and Signal*

Processing (ICASSP). IEEE, 2016, pp. 5210–5214.

[8] Y. Kubo, T. Nakatani, M. Delcroix, K. Kinoshita, and S. Araki, "Mask

based mvdr beamformer for noisy multisource environments: Intro

duction of time-varying spatial covariance model," in *ICASSP 2019-*

2019 IEEE International Conference on Acoustics, Speech and Signal

Processing (ICASSP). IEEE, 2019, pp. 6855–6859.

[9] M. Togami, "Simultaneous optimization of forgetting factor and time

frequency mask for block online multi-channel speech enhancement,"

in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics,*

Speech and Signal Processing (ICASSP), 2019, pp. 2702–2706.

[10] T. Ochiai, M. Delcroix, T. Nakatani, and S. Araki, "Mask-based neural

beamforming for moving speakers with self-attention-based tracking,"

BIOGRAPHIES

I deeply grateful to our esteemed faculty mentors, **Dr. Sonagiri China Venkateswarlu, Dr. V. Siva Nagaraju**, from the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE).

Dr. Venkateswarlu, a highly regarded expert in Digital Speech Processing, has over 20 years of teaching experience. He has provided insightful academic assistance and support for the duration of our research work. Dr. Siva Nagaraju, an esteemed researcher in Microwave Engineering who has been teaching for over 21 years, has provided us very useful and constructive feedback, and encouragement which greatly assisted us in refining our technical approach.**Dr Sonagiri China Venkateswarlu** professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Digital Speech Processing. He has more than 40 citations and paper publications across various publishing platforms, and expertise in teaching subjects such as microprocessors and microcontrollers , digital signal processing, digital image processing, and speech processing. With 20 years of teaching experience, he can be contacted at email: chinavenkateswarlu@iare.ac.in

Dr. V. Siva Nagaraju is a professor in the Department of Electronics and Communication Engineering at the Institute of Aeronautical Engineering (IARE). He holds a Ph.D. degree in Electronics and Communication Engineering with a specialization in Microwave Engineering

With over 21 years of academic experience, Dr. Nagaraju is known for his expertise in teaching core electronics subjects contributed significantly to the academic and research community. He can be contacted at email: v.sivanagaraju@iare.ac.in.

Mali Charitha studying 3rd year departmenof Electronics And Communication Engineering at Institute Of Aeronautical Engineering ,Dundigal .She Published a Research Paper Recently At IJSREM as a part of academics . She has a interest in Embedded Systems and VLSI