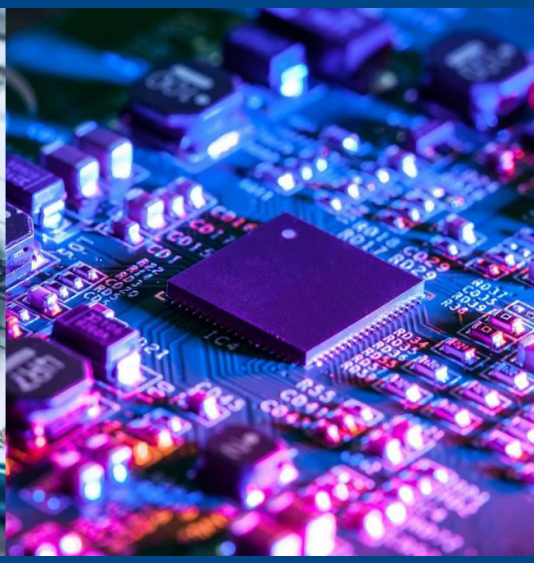
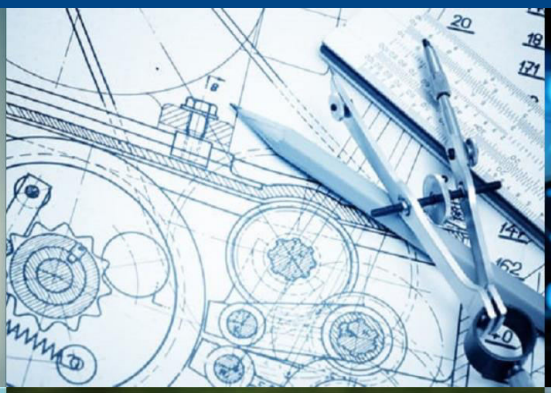
ISSN: 2582-7219

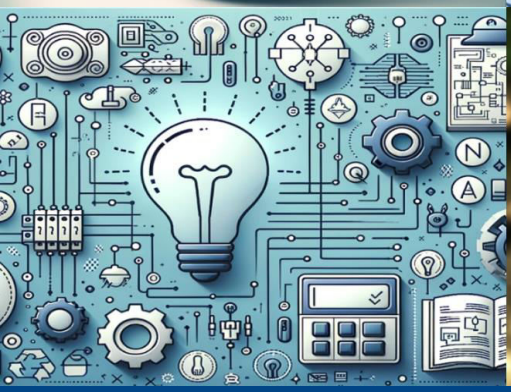




International Journal of Multidisciplinary Research in Science, Engineering and Technology

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.206 **Volume 9, Issue 3, March 2026**

© 2026 IJMRSET | Volume 9, Issue 3, March 2026| DOI: 10.15680/IJMRSET.2026.0903231

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Hybrid Deep Learning Architecture Using Attention-Driven Analysis for Optimal Multi-Scale Speech Processing

K.Vidhya[1], B.C.Keerthi[2], K.Nagarjuna[3], M.N.S.Lokesh [4], G.N.P.Jyothi[5]

U.G. Student, Department of ECE, SVIET Engineering College, Nandamuru, Pedana, Andhra Pradesh, India[1,2,3,4]

Assistant Professor, Department of ECE, SVIET Engineering College, Nandamuru, Pedana, Andhra Pradesh, India[5]

ABSTRACT: Speech processing is essential in modern communication but is often affected by noise, reverberation, and speaker interference, reducing performance in real-world applications. Traditional methods like Spectral Subtraction and Wiener Filtering struggle in complex conditions, while deep learning models such as CNNs and LSTMs have limitations in capturing long-term dependencies and multi-scale features. To overcome these issues, this project proposes the Omega-Scale architecture, a hybrid model combining CNN-Transformer (Conformer), Spiking Neural Networks, and quantum-inspired attention for efficient multi-scale speech processing, along with a multi-task framework for enhancement, emotion recognition, and deepfake detection. Implemented using PyTorch with GPU acceleration and supported by tools like Librosa and torchaudio, the system aims to improve speech quality metrics like PESQ and STOI while providing a scalable and real-time capable solution.

KEYWORDS: CNN, RNN, Conformer, Spiking Neural Networks, LSTM, GUI, CNN- Transformer, Multiscale processing, Attention Mechanisms.

I. INTRODUCTION

A speech signal is the sound produced when a person speaks, carrying information such as words, emotions, and tone. When we speak, our vocal cords create vibrations that travel through the air as sound waves. These sound waves are captured by a microphone and converted into electrical or digital signals that can be processed by a computer. This process enables machines to analyze, interpret, and respond to human speech effectively.

Speech processing plays a crucial role in modern Artificial Intelligence (AI) systems by enabling natural interaction between humans and machines through voice. It allows hands-free operation, making systems more convenient and accessible, especially for people with disabilities. Additionally, it supports real-time communication and enhances user experience in various applications. Overall, speech processing makes AI systems more intelligent, user-friendly, and capable of understanding human language in a natural way.

Speech processing is widely used in many real-life applications. It powers voice assistants, supports healthcare systems through voice-based monitoring, improves communication systems, enhances security through voice recognition, and aids in education through interactive learning tools. By converting human speech into a format that computers can understand, speech processing plays a significant role in making modern technology more efficient and accessible.

II. RELATED WORK

Many researchers have developed advanced speech processing systems to improve the quality and understanding of speech signals in real-world environments. In earlier systems, traditional signal processing techniques such as Spectral Subtraction and Wiener Filtering were widely used to remove noise from speech signals. These methods were simple and mathematically based, but they struggled to perform effectively in complex and dynamic environments with varying noise conditions.

Modern systems also integrate multiple tasks such as speech enhancement, emotion recognition, and deepfake detection into a single framework using multi-task learning. These systems provide more accurate and comprehensive analysis of

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

speech signals and are widely used in applications like communication systems, healthcare, security, and smart assistants.

However, challenges still remain in terms of computational complexity, real-time performance, and efficient handling of multi-scale features. Therefore, further research is required to develop hybrid architectures that are accurate, efficient, and suitable for real-world applications.

III. METHODOLOGY

The proposed methodology is based on a hybrid deep learning system that performs real-time speech processing and analysis.

[1] 3.1 System Operation

The system continuously processes speech signals through different stages:

- Input Stage → Captures audio from microphone or files.
- Preprocessing Stage → Noise normalization and enhancement.
- Feature Extraction → Generates spectrograms and MFCCs.
- Processing Stage → Hybrid model extracts features.
- Multi-task Output → Enhancement, emotion detection, deepfake detection.

[2] 3.2 Functional Modules

The architecture consists of the following modules:

- Audio Input Module: Captures speech signals.
- Preprocessing Module: Removes noise and normalizes audio.
- Feature Extraction Module: Generates MFCC and spectrogram features.
- Conformer Module: Extracts local and global features.
- SNN Module: Captures temporal patterns efficiently.
- Attention Module: Focuses on important signal features.
- Multi-task Module: Performs enhancement and classification.
- Output Module: Displays results and visualizations.

[3] 3.3 System Flow

The system is developed using a structured processing flow:

- Audio input and preprocessing
- Feature extraction
- Hybrid model processing
- Multi-task prediction
- Result visualization

[4] 3.4 Implementation

The system is implemented using a deep learning methodology:

- Data preprocessing and feature engineering
- Model design using hybrid architecture
- Training and validation
- Performance evaluation
- Real-time deployment

© 2026 IJMRSET | Volume 9, Issue 3, March 2026| DOI: 10.15680/IJMRSET.2026.0903231

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

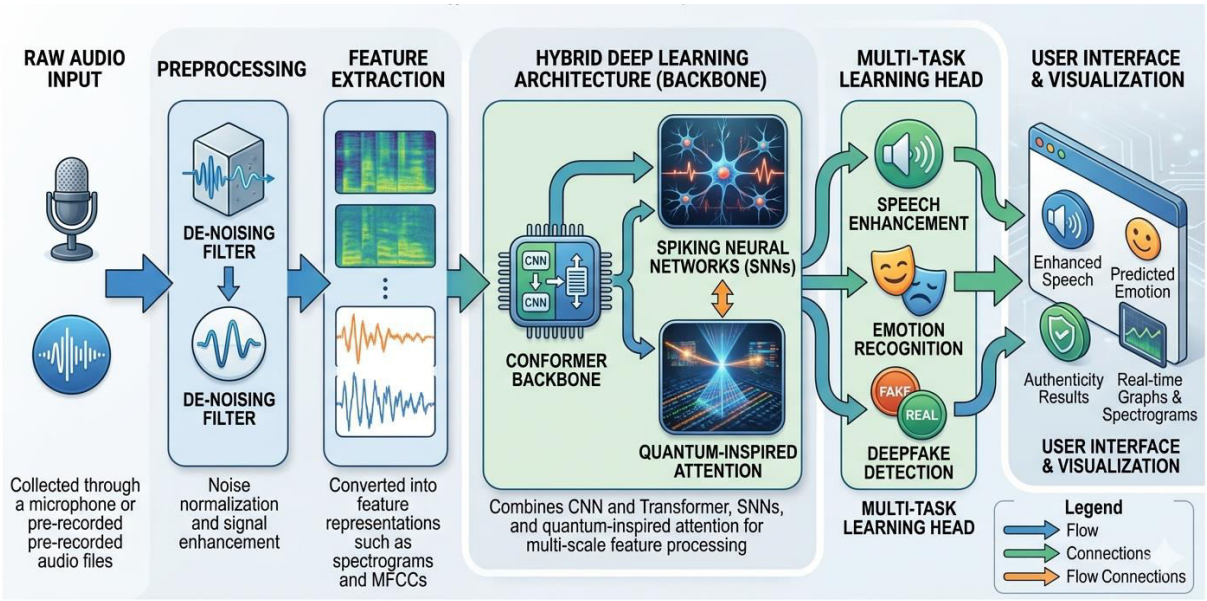


FIG 1: BLOCK DIAGRAM OF THE PROPOSED HYBRID DEEP LEARNING

The FIG 1 represents a hybrid deep learning-based speech processing system designed for real-time analysis and enhancement of speech signals. The system begins with raw audio input collected through a microphone or pre-recorded audio files. This input is passed to the preprocessing stage, where de-noising filters are applied for noise normalization and signal enhancement. The processed signal is then converted into feature representations such as spectrograms and MFCCs in the feature extraction stage. These features are fed into the hybrid deep learning architecture, which consists of a Conformer backbone that combines CNN and Transformer models. Additionally, Spiking Neural Networks (SNNs) and quantum-inspired attention mechanisms are integrated to capture temporal patterns and improve feature selection. The processed output is then passed to the multi-task learning head, where speech enhancement, emotion recognition, and deepfake detection are performed. Finally, the results are displayed in the user interface in the form of enhanced speech, predicted emotion, authenticity results, and real-time graphs.

IV. EXPERIMENTAL RESULTS

The proposed system is evaluated based on performance, accuracy, and real-time processing capability.

[1] 4.1 Performance Analysis

The system performance is analyzed under different conditions:

- Clean Audio → High accuracy and clear enhancement
- Noisy Audio → Effective noise reduction and stable output
- Real-time Input → Low latency and continuous processing

[2] 4.2 Accuracy Evaluation

The model achieves reliable results across multiple tasks:

- Speech Enhancement → Improved signal clarity
- Emotion Recognition → Accurate classification of emotions
- Deepfake Detection → Effective identification of fake audio

[3] 4.3 System Efficiency

The efficiency of the system is evaluated as follows:

- Reduced processing delay
- Optimized resource utilization
- Energy-efficient computation using SNN

© 2026 IJMRSET | Volume 9, Issue 3, March 2026| DOI: 10.15680/IJMRSET.2026.0903231

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

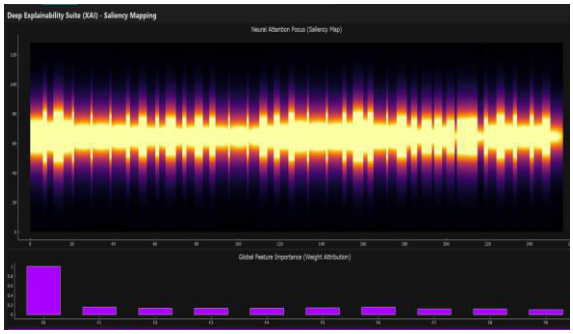
[4] 4.4 Visualization Results

The output results are visualized using:

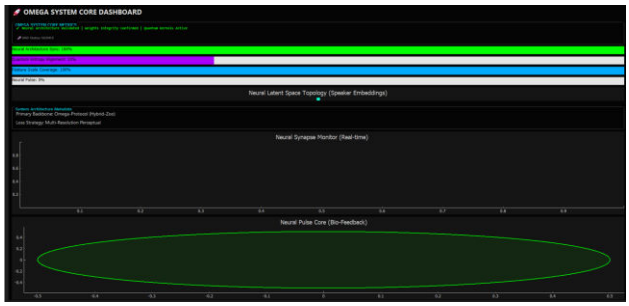
- Spectrogram representations
- Feature distribution graphs
- Prediction output displays

FIG 2: GRAPHICAL USER INTERFACE RESULT

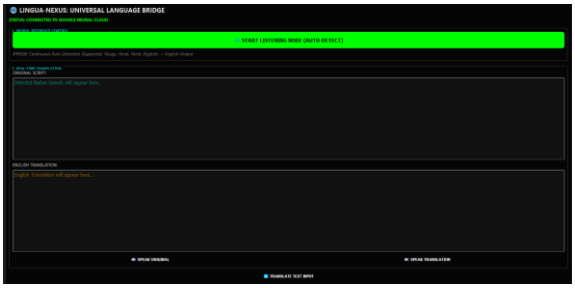
(a) GUI



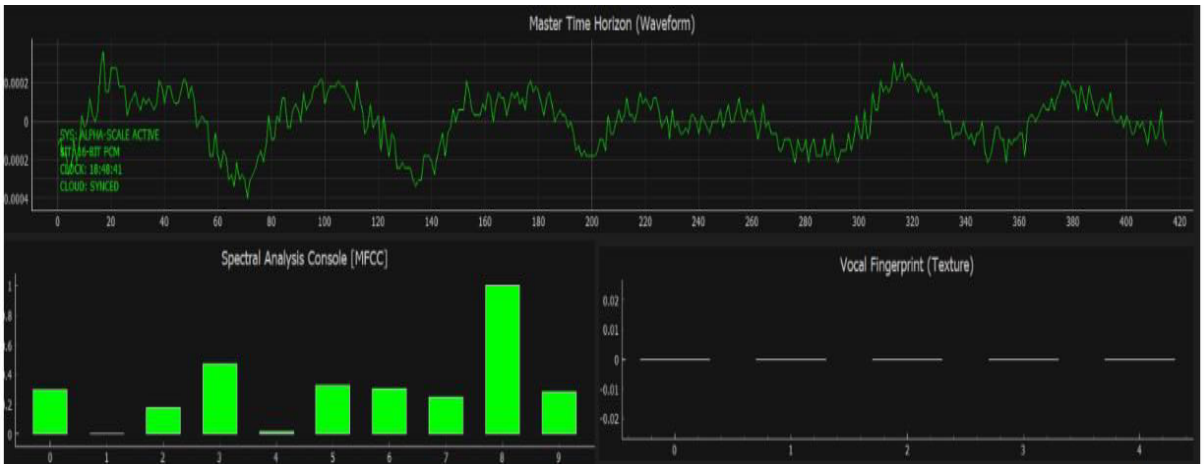
(b) Neural Attention Focus



(c) System Dashboard



(d) Output Translator, Interpreter



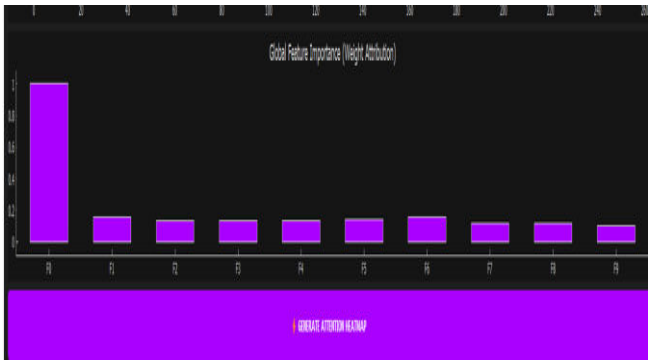
(e) Input Audio Speech And Spectral Analysis

© 2026 IJMRSET | Volume 9, Issue 3, March 2026| DOI: 10.15680/IJMRSET.2026.0903231

ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 8.206| ESTD Year: 2018|

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



(f) Attention Heatmap

(g) Emotion Detection

Fig 2 shows the results of Hybrid deep learning for speech processing (a) GUI, (b) Neural Attention Focus, (c) System Dashboard, (d) Output Translator, Interpreter, (e) Input Audio Speech And Spectral Analysis, (f) Attention Heatmap, (g) Emotion Detection.

V. CONCLUSION

The proposed hybrid deep learning-based speech processing system provides an efficient and reliable solution for real-time speech analysis and enhancement. By integrating advanced techniques such as Conformer architecture, Spiking Neural Networks (SNNs), and quantum-inspired attention, the system is able to accurately process speech signals and improve overall quality. It effectively performs multiple tasks such as speech enhancement, emotion recognition, and deepfake detection within a single framework. With the help of modern AI techniques, the processed data can be easily analyzed and visualized through a user-friendly interface, allowing users to understand speech information clearly and efficiently.

This system reduces the limitations of traditional methods and provides continuous and automated speech processing. It is scalable, efficient, and suitable for applications such as communication systems, healthcare, security, and smart assistants. Although the system shows strong performance, there is still scope for improvement in areas such as reducing computational complexity, improving real-time efficiency, and enhancing model accuracy with larger datasets. Overall, the proposed system is a powerful and effective solution for modern speech processing applications.

REFERENCES

[1] Ian Goodfellow, Yoshua Bengio, Aaron Courville (2016) Deep Learning. MIT Press
[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I (2017) Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS)
[3] Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y (2020) Conformer: Convolution-Augmented Transformer for Speech Recognition. Interspeech
[4] Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. Neural Computation 9(8):1735–1780
[5] Wang D, Chen J (2018) Supervised Speech Separation Based on Deep Learning: An Overview. IEEE/ACM Transactions on Audio, Speech, and Language Processing 26(10):1702–1726
[6] LeCun Y, Bengio Y, Hinton G (2015) Deep Learning. Nature 521:436–444
[7] Gerstner W, Kistler WM (2002) Spiking Neuron Models. Cambridge University Press
[8] Lipton ZC (2015) A Critical Review of Recurrent Neural Networks for Sequence Learning. arXiv preprint arXiv:1506.00019
[9] Ko T, Peddinti V, Povey D, Seltzer M, Khudanpur S (2015) Audio Augmentation for Speech Recognition. Interspeech

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH

IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com