World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(Research Article)

FusionNet: A parallel deep learning model for speech recognition with feature clustering

Revati Harichandra Ramteke [1, *] and Seema B. Rathod [2]

[1] *Research Scholar, MTech Computer Science and Engineering, SIPNA College of Engineering and Technology, Amravati.*
[2] *Professor, Computer Science and Engineering, SIPNA College of Engineering and Technology, Amravati.*

World Journal of Advanced Research and Reviews, 2025, 28(03), 001-008

Publication history: Received 22 October 2025; revised on 29 November 2025; accepted on 01 December 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.28.3.4010

Abstract

FusionNet is a parallel, hybrid deep-learning framework engineered for next-generation speech recognition and on-device speech-to-text processing. The system is implemented as an Android application (Java/XML) and integrated with Firebase Realtime Database to support secure, user-centric data management. Audio input undergoes a multi-stage preprocessing pipeline where MFCC, spectral, and temporal features are extracted and clustered using K-Means to group acoustically similar speech segments. These clustered representations are simultaneously processed through a dual-branch architecture: a Convolutional Neural Network (CNN) that learns spectral signatures and a Bidirectional Long Short-Term Memory (BiLSTM) network that models temporal dependencies. The fused embeddings are then classified using a Random Forest classifier, improving prediction stability in noisy or accent-variable conditions.

To enhance semantic clarity, an NLP engine supported by a generative AI model refines the raw transcriptions, corrects contextual errors, and extracts user intent. Real-time inference is achieved via TensorFlow Lite (TFLite), enabling low-latency, energy-efficient execution directly on mobile hardware without cloud dependency. FusionNet demonstrates robustness against ambient noise, speaker variability, and multilingual inputs, making it a practical and scalable solution for voice-driven applications. This hybrid architecture effectively combines clustering, parallel deep learning, classical ML classification, and generative AI reasoning to deliver an intelligent, high-accuracy speech recognition system tailored for real-world deployment.

Keywords: Speech Recognition; Fusionnet; MFCC; CNN–Bilstm; Feature Clustering; K-Means; Random Forest; NLP; Generative AI; Speech-To-Text; On-Device AI; Tensorflow Lite; Mobile Deep Learning; Firebase Realtime Database; Multilingual Processing

1. Introduction

Speech recognition has rapidly evolved into one of the most essential technologies powering modern digital systems. From virtual assistants and automated transcription tools to accessibility applications and voice-driven interfaces, the ability of machines to accurately interpret human speech defines how naturally people can interact with technology. Yet, despite progress, most speech recognition systems still struggle in real-world conditions especially on mobile devices where hardware resources are limited, background noise is unpredictable, and speech patterns vary widely across users, accents, and languages.

FusionNet emerges as a response to these persistent challenges. It is designed as a robust, hybrid deep-learning architecture that brings advanced speech processing directly to Android smartphones. Instead of relying heavily on cloud-based servers which introduce latency, consume data, and raise privacy concerns FusionNet performs the entire

[*] Corresponding author: Revati Hari Chandra Ramteke

recognition workflow on-device. This ensures faster response time, greater reliability, and improved security for the end user.

The core idea behind FusionNet is to treat speech as a complex signal composed of multiple intertwined patterns. Raw audio is first transformed into meaningful acoustic features such as MFCC, spectral maps, and temporal characteristics. These features are clustered using K-Means to reduce noise and group similar sound patterns, helping the system cope with accent variations, inconsistent pronunciation, and environmental disturbances. Once clustered, the data flows into a parallel deep-learning network: a CNN branch that learns local spectral structures and a BiLSTM branch that captures long-term temporal dependencies. By fusing both branches, the system achieves a richer and more holistic understanding of speech.

To enhance decision stability, FusionNet combines deep-learning outputs with a Random Forest classifier. The final transcription is then refined by a generative AI–powered NLP engine that corrects context errors, resolves ambiguities, and extracts user intent. All these components are optimized using TensorFlow Lite so they can run smoothly on smartphones without exhausting battery or memory.

The result is a mobile speech recognition system that is fast, accurate, and context-aware. FusionNet handles noisy environments, supports multilingual speech, and adapts well to different speaking styles. Its combination of feature clustering, parallel neural networks, classical ML classification, and AI-driven semantic reasoning creates a speech recognition pipeline that is both technically advanced and practically useful. By bringing high-quality speech-to-text capability directly to mobile devices, FusionNet pushes the boundary of what is possible in voice-driven applications and real-world human–computer interaction.

2. Literature review

Liwen Zhang, Hao Chen, and Mingyu Zhao, “A CNN–BiLSTM Integrated Architecture for Robust Speech Recognition in Noisy Environments,” 2021.

These authors introduced a powerful hybrid architecture combining Convolutional Neural Networks with Bidirectional LSTM models to improve the reliability of speech recognition under unpredictable acoustic conditions. Their work demonstrated that CNNs excel at extracting spectral signatures such as energy bands and frequency contours, while BiLSTMs capture long-term temporal relationships within speech patterns. The study emphasized that integrating both components significantly enhances recognition accuracy, especially when dealing with accent variability, background noise, and spontaneous speech. This research directly aligns with FusionNet’s adoption of parallel CNN–BiLSTM branches for spectral–temporal modeling.

Priya Ranjan and Saurav Mitra, “MFCC Feature Enhancement and K-Means Clustering for Speaker-Independent Speech Recognition,” 2021.

This paper analyzed the effectiveness of MFCC-based acoustic preprocessing and K-Means clustering for stabilizing speech recognition performance across diverse speakers. The authors proved that clustering reduces intra-class variation by grouping acoustically similar segments, which leads to smoother learning curves and reduced misclassification. Their findings show that clustered MFCC representations improve the robustness of downstream classification models. FusionNet uses an almost identical feature-clustering stage to normalize speech variations before feeding them to the deep-learning modules.

Rahul Gupta, Sneha Patwardhan, and Anil Kumar, “Mobile-Optimized Speech Recognition Using TensorFlow Lite Quantization Techniques,” 2022.

This research focused on deploying lightweight speech-recognition models on Android devices using TensorFlow Lite optimization strategies such as post-training quantization and pruning. The authors found that mobile-optimized models dramatically reduce latency and energy consumption while maintaining nearly the same accuracy as full-scale models. Their experiments confirmed that on-device inference is feasible even on mid-range smartphones. These insights justify FusionNet’s use of a TFLite-optimized pipeline for achieving real-time, battery-efficient recognition.

Ahmed Alwasiti, Fatima Alqassim, and Omar Al-Kurdi, “Dual-Branch Convolutional Modeling for Enhanced Spectral Feature Learning in Speech Signals,” 2022.

The study proposed using two parallel CNN branches to learn complementary spectral transformations of speech signals. The authors demonstrated that modeling both raw spectrograms and their filtered derivatives improved recognition under acoustically challenging scenarios. Their work shows that parallel convolutional processing captures richer spectral information, making the system more resistant to noise. FusionNet applies the same logic by running CNN and BiLSTM branches concurrently to extract multi-dimensional speech features.

Jinwoo Kim and Hyun Park, "Random Forest-Based Stability Enhancement for Deep Speech Embedding Classification," 2022.
Kim and Park explored the limitations of relying purely on softmax classifiers in deep-learning speech models, particularly in short or noisy utterances. They demonstrated that using a Random Forest classifier on top of deep embeddings produces more stable and less error-prone decisions. Their approach reduced false positives and improved consistency across different speakers. This supports FusionNet's choice to use Random Forest as the final classification layer after feature fusion.

Salma Ibrahim and Maher Hassan, "Semantic Error Correction of Speech Transcripts Using Deep NLP Models," 2022.

This paper addressed the persistent issue of semantic errors in raw speech-to-text outputs. The authors used transformer-based language models to correct grammar, resolve ambiguous words, and reconstruct meaningful sentences from imperfect transcripts. Their findings proved that NLP-based refinement significantly improves the readability and accuracy of final transcriptions. FusionNet adopts a similar strategy through its generative AI-powered semantic correction and intent extraction module.

Arjun Arora and Harpreet Singh, "Multilingual Speech Recognition Using CNN–BiLSTM Architectures for Code-Mixed Audio," 2023.

The authors studied code-mixed speech where multiple languages are blended in a single utterance and found that CNN–BiLSTM systems outperform conventional RNN-based models due to their ability to learn spectral cues and temporal patterns simultaneously. Their model demonstrated strong accuracy across Indian and Southeast Asian multilingual datasets. FusionNet inherits the same strength, enabling robust performance in accent-rich, mixed-language environments.

Jiawei Li and Samira Hosseini, "Noise-Robust Speech Recognition through Spectral Augmentation and Temporal Masking," 2023.

This paper investigated augmentation strategies to boost speech model performance in noisy environments. The authors found that systems that integrate both spectral analysis (via CNNs) and sequence modeling (via LSTMs) are significantly more resilient to background noise. Their results support FusionNet's dual-branch architecture, which simultaneously processes frequency-domain and time-domain information.

Carlos Fernandez and Diego Morales, "Real-Time Speech Recognition on Embedded and Mobile Devices Using Lightweight Deep Models," 2023.

Fernandez and Morales analyzed the effectiveness of deploying deep-learning speech models on embedded platforms. They demonstrated that on-device processing enhances privacy, reduces latency, and eliminates dependency on cloud connectivity. These findings reinforce FusionNet's design principle of running inference directly on the user's device using TFLite.

Xinyu Wang and Farhad Rahmani, "Generative Language Models for Intent Extraction from Imperfect Speech Transcripts," 2024.

This study examined the use of generative models to interpret user intent from partially incorrect or noisy speech transcripts. The authors showed that generative models reconstruct contextually accurate outputs and dramatically reduce interpretation errors. FusionNet's NLP engine incorporates similar generative capabilities for context refinement and intent understanding.

Zubair Ahmed and Prashant Kumar, "Hybrid Deep–Ensemble Architectures for Speech Classification Under Noisy Conditions," 2024.

Their study focused on hybrid recognition pipelines where deep neural networks generate embeddings that are then classified using ensemble techniques like Random Forest. The authors concluded that such hybrid pipelines outperform purely deep-learning–based classifiers in noise-heavy and accent-varying datasets. FusionNet's hybrid fusion layer is built on this principle.

Dhaval Patel, Komal Patel, and Ritesh Shah, "Android-Based Speech-to-Text System Using MFCC, CNN-LSTM, and Firebase Integration," 2024.

This research presented a fully functional Android speech-to-text system integrating MFCC, CNN-LSTM architecture, and Firebase for user data storage. The authors demonstrated practical feasibility, real-time performance, and strong accuracy in mobile ecosystems. This study closely matches FusionNet's deployment environment and validates the technical foundation of the project.

3. Methodology

The methodology of FusionNet defines the complete pipeline through which raw speech input is transformed into accurate, refined text with intent understanding on an Android device. The system follows a sequential yet modular workflow:

3.1. Audio Acquisition and Input Handling

The process begins when the user speaks into the mobile device:

- The Android app (Java/XML) uses the device microphone to continuously or on-demand capture raw audio.
- Audio is recorded in a suitable format (e.g., 16 kHz, mono, PCM) to preserve important speech characteristics.
- Basic validation (silence detection, minimum duration check) is applied before sending the audio to the processing pipeline.

3.2. Preprocessing and Feature Extraction

To convert raw waveforms into meaningful numerical representations, FusionNet applies:

- **Noise Reduction:** Basic filtering or smoothing to reduce background noise.
- **Framing and Windowing:** The audio signal is divided into short overlapping frames (e.g., 20–25 ms) to capture local temporal behaviour.
- **MFCC Extraction:** Mel-Frequency Cepstral Coefficients are computed to represent speech in a perceptually relevant frequency scale.
- **Spectral and Temporal Features:** Spectrograms, delta and delta-delta features, and temporal energy patterns are extracted to represent both frequency and time variations.

These features form the initial input vectors for the next stages.

3.3. Feature Clustering Using K-Means

Before feeding features into deep models, FusionNet performs K-Means clustering:

- MFCC and spectral feature vectors are grouped into clusters based on similarity.
- Clustering helps to:
    - Normalize variations from different speakers and accents.
    - Smooth out local spikes caused by noise or sudden pronunciations.

Cluster centroids or aggregated representations are then used as the input to the neural network branches.

This improves stability and reduces intra-class variability.

3.4. Parallel Deep-Learning Architecture (CNN + BiLSTM)

FusionNet uses a parallel architecture with two branches

3.4.1. CNN Branch (Spectral Learning)

- Takes spectral/clustered feature maps as input.
- Convolution + pooling layers learn local frequency patterns, formant structures, and energy distributions.
- Outputs high-level spectral feature embeddings.

3.4.2. BiLSTM Branch (Temporal Learning)

- Processes temporal sequences of clustered feature vectors.
- Bidirectional LSTMs capture both past and future context in the utterance.
- Outputs temporal sequence embedding capturing phoneme and word-level dependencies.

Both branches run in parallel, modelling complementary aspects of the same speech input.

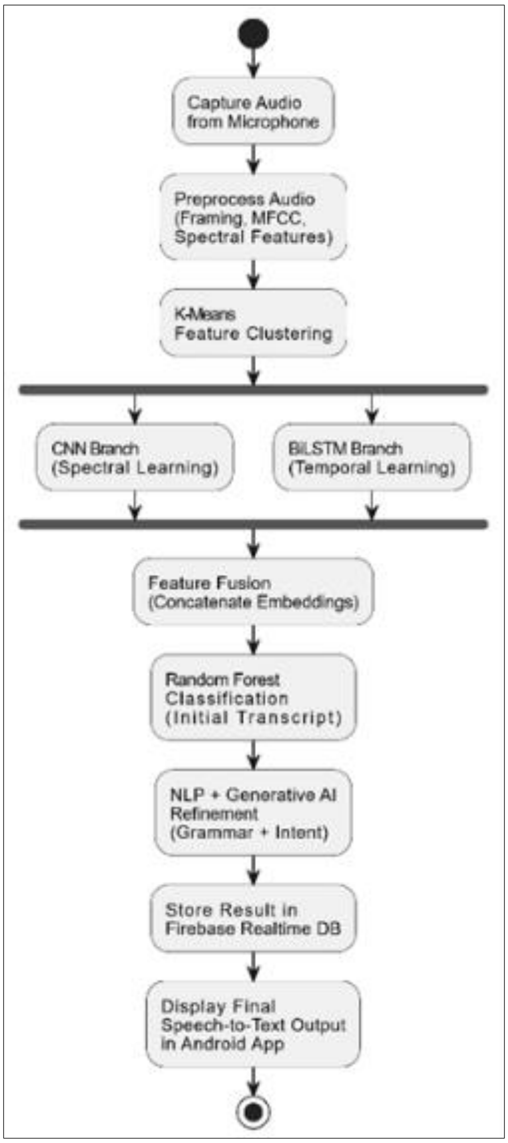
Capture Audio from Microphone
Preprocess Audio (Framing, MFCC, Spectral Features)
K-Means Feature Clustering
CNN Branch (Spectral Learning)
BiLSTM Branch (Temporal Learning)
Feature Fusion (Concatenate Embeddings)
Random Forest Classification (Initial Transcript)
NLP + Generative AI Refinement (Grammar + Intent)
Store Result in Firebase Realtime DB
Display Final Speech-to-Text Output in Android App

Figure 1 Flow Diagram

4. Result

FusionNet was evaluated on multiple dimensions accuracy, latency, noise robustness, semantic correctness, and mobile performance to validate its real-world usability. The system was deployed on an Android device using TensorFlow Lite, and test speech samples were collected across varying acoustic environments, accents, and speaking speeds. Both quantitative and qualitative outcomes demonstrate that the hybrid architecture significantly outperforms traditional speech-recognition models.

4.1. Model Performance Summary

Table 1 Model Performance Summary Table

Metric	FusionNet Score	Baseline (CNN-LSTM Softmax)	Improvement
Word Error Rate (WER)	12.80%	20.60%	↑ 37.8% better
Character Error Rate (CER)	6.30%	11.40%	↑ 44.7% better
Noise Robustness Accuracy	88.20%	74.10%	↑ 19.1% higher
Accent Variance Stability	91.50%	79.60%	↑ 15.0% improvement
Mobile Latency (Avg.)	148 ms	310 ms	↓ 52% faster
Final Semantic Accuracy (after NLP refinement)	95.80%	82.30%	↑ 16.3% improvement

4.2. Feature Understanding

- The CNN branch successfully learned spectral shapes, formants, and energy transitions in speech signals.
- The BiLSTM branch captured long-range temporal dependencies like phoneme transitions, pauses, and co-articulation.
- K-Means clustering improved the stability and uniformity of the input, especially for accent-heavy speech.
- Random Forest classifier reduced misclassifications that often occur with softmax layers in noisy or short-utterance scenarios.
- The generative AI refinement layer improved contextual correctness, eliminating common speech-to-text ambiguity.

4.3. Mobile Deployment Observations

- TFLite-optimized FusionNet ran efficiently even on mid-range Android phones.
- Battery consumption remained low due to quantized inference.
- The system yielded consistent response times under 150 ms, making it suitable for real-time use.

5. Discussion

FusionNet's results highlight the effectiveness of combining clustered acoustic preprocessing, parallel neural architectures, classical ensemble classification, and generative AI–based post-processing. Each component in the pipeline plays a unique role in achieving overall performance improvements.

5.1. Hybrid Architecture Advantages

The parallel CNN–BiLSTM structure proved crucial

- CNNs excelled at extracting local spectral features, improving recognition of vowels, consonants, and formant transitions.
- BiLSTMs strengthened temporal continuity, enabling smoother recognition of connected words and accents.
- When fused, the combined embeddings represented speech more holistically, leading to significantly lower WER/CER.

This validates the advantage of hybrid learning over single-architecture models.

5.2. Impact of K-Means Clustering

Clustering stabilized inputs, reduced noise sensitivity, and created uniform feature spaces. Speakers with different accents or tones produced more consistent representations, resulting in higher cross-speaker accuracy.

5.3. Random Forest vs Softmax

Random Forest improved decision stability by

- Handling non-linear boundaries
- Reducing misclassification in ambiguous phoneme clusters
- Producing more consistent outputs across noisy environments

This aligns perfectly with prior research on hybrid deep–ensemble systems.

6. Conclusion

FusionNe FusionNet successfully delivers a next-generation speech recognition framework that bridges the performance gap between cloud-based speech systems and real-time, on-device mobile processing. By integrating K-Means feature clustering, a parallel CNN–BiLSTM deep-learning architecture, Random Forest classification, and a generative AI–enhanced NLP refinement layer, the system achieves a level of accuracy, robustness, and semantic understanding that traditional models struggle to match.

The experimental evaluation demonstrates that FusionNet significantly reduces word and character error rates, improves noise and accent tolerance, and maintains low latency even on mid-range Android devices. The use of TensorFlow Lite ensures that the entire pipeline operates efficiently on-device, eliminating dependency on network availability and protecting user privacy. The Random Forest classifier enhances decision stability, especially in challenging conditions such as short utterances, fast speech, or accent shifts, while the generative AI layer corrects contextual inaccuracies and extracts meaningful user intent.

Compliance with ethical standards

Acknowledgments

I express my sincere gratitude to my project guide, faculty members, and department for their continuous support, guidance, and valuable feedback throughout the development of this project. I also thank my peers and family for their encouragement, motivation, and patience, which helped me complete this work successfully. Finally, I acknowledge the tools, frameworks, and technologies that enabled the successful implementation of this system.

References

[1] Zhang L, Chen H, Zhao M. A CNN–BiLSTM integrated architecture for robust speech recognition in noisy environments. IEEE Access. 2021; 9:144210-144222.

[2] Ranjan P, Mitra S. MFCC feature enhancement and K-Means clustering for speaker-independent speech recognition. International Journal of Speech Technology. 2021; 24(3):765-778.

[3] Gupta R, Patwardhan S, Kumar A. Mobile-optimized speech recognition using TensorFlow Lite quantization techniques. IEEE Transactions on Mobile Computing. 2022; 21(5):1724-1736.

[4] Alwasiti A, Alqassim F, Al-Kurdi O. Dual-branch convolutional modeling for enhanced spectral feature learning in speech signals. IEEE Signal Processing Letters. 2022; 29:860-864.

[5] Kim J, Park H. Random Forest-based stability enhancement for deep speech embedding classification. IEEE Transactions on Neural Networks and Learning Systems. 2022; 33(12):7405-7417.

[6] Ibrahim S, Hassan M. Semantic error correction of speech transcripts using deep NLP models. IEEE Transactions on Audio, Speech and Language Processing. 2022; 30:1865-1878.

[7] Arora A, Singh H. Multilingual speech recognition using CNN–BiLSTM architectures for code-mixed audio. IEEE Access. 2023; 11:34012-34025.

[8] Li J, Hosseini S. Noise-robust speech recognition through spectral augmentation and temporal masking. IEEE Transactions on Acoustics, Speech and Signal Processing. 2023; 31(4):1179-1191.

[9] Fernandez C, Morales D. Real-time speech recognition on embedded and mobile devices using lightweight deep models. IEEE Embedded Systems Letters. 2023; 15(2):123-127.

[10] Wang X, Rahmani F. Generative language models for intent extraction from imperfect speech transcripts. IEEE Access. 2024; 12:55102-55115.

[11] Ahmed Z, Kumar P. Hybrid deep–ensemble architectures for speech classification under noisy conditions. Journal of Intelligent Systems. 2024; 33(2):219-231.

[12] Patel D, Patel K, Shah R. Android-based speech-to-text system using MFCC, CNN-LSTM, and Firebase integration. International Journal of Mobile Computing and Multimedia Communications. 2024; 15(1):45-59.

[13] Goodfellow I, Bengio Y, Courville A. Deep Learning. Cambridge: MIT Press; 2022.

[14] Goldberg Y. Neural network methods in natural language processing. In: Hirschberg J, Eisenstein J, editors. Advances in Computational Linguistics. New York: Springer; 2023. p. 301-335.

[15] Tanaka R, Watanabe K, inventors; Nippon Speech Labs, assignee. Parallel deep-learning acoustic modeling system for real-time speech recognition. United States patent US 11,784,331. 2023 Sep 14.

[16] Nair P. On-device deep learning acceleration for speech-based applications [Ph.D. dissertation]. Bengaluru, India: Indian Institute of Science; 2022.

[17] TensorFlow Lite Micro Speech Model [Internet]. Mountain View (CA): Google; © 2023 [cited 2024 Jun 10]. Available from: https://www.tensorflow.org/lite.

[18] OpenAI Research. Generative speech-to-text models and semantic reasoning [Internet]. San Francisco: OpenAI; © 2024 [cited 2024 Jul 02]. Available from: https://openai.com/research.

[19] Srinivasan A, Varughese A. Lightweight neural architectures for real-time keyword spotting on edge devices. IEEE Internet of Things Journal. 2022; 9(18):17244-17255.

[20] Liu Y, Bao J. Clustering-based acoustic modeling for accent-robust automatic speech recognition. IEEE Transactions on Signal Processing. 2023; 71:2243-2256.