CNN-LSTM HYBRID MODEL FOR ROBUST SPEECH RECOGNITION IN HIGH-NOISE ENVIRONMENTS

Project Reference No.: *49S_BE_0514*

College : CMR Institute of Technology, Bengaluru
Branch : Department of Information Science and Engineering
Guide(s) : Prof. Shilpa Mangesh Pande
Dr. Susheelamma Kh
Student(S) : Mr. Abhimanyu Tiwari
Mr. Bs Tushar
Mr. Atiksh V Jain

Keywords:
Speech-to-Text, CNN–LSTM, Automatic Speech Recognition, MFCC, Noise Robustness

Introduction:

1. Speech-to-text conversion, also known as Automatic Speech Recognition (ASR), enables machines to convert spoken language into text. It is widely used in applications such as virtual assistants, transcription systems, and voice-based interfaces. With the growth of voice-enabled technologies, the need for accurate ASR systems has increased significantly.
2. However, real-world speech recognition remains challenging due to variations in accents, pronunciation, and speaking styles. Environmental noise and interference further degrade performance. Many existing ASR systems perform well in controlled conditions but fail in noisy environments.
3. To address this, the project proposes a hybrid CNN–LSTM-based speech-to-text system. CNNs extract spectral features, while LSTMs capture temporal dependencies in speech. MFCC features are used as input due to their effectiveness in representing speech signals.
4. Noise augmentation is applied during training to improve robustness. The model uses CTC loss for end-to-end learning without alignment. The aim is to develop an efficient and robust ASR system suitable for real-world noisy environments.

Objectives:

The project aims to develop a highly robust Automatic Speech Recognition (ASR) system specifically optimized for environments with high ambient noise (e.g., factories, traffic, crowded places), where standard ASR models typically fail.

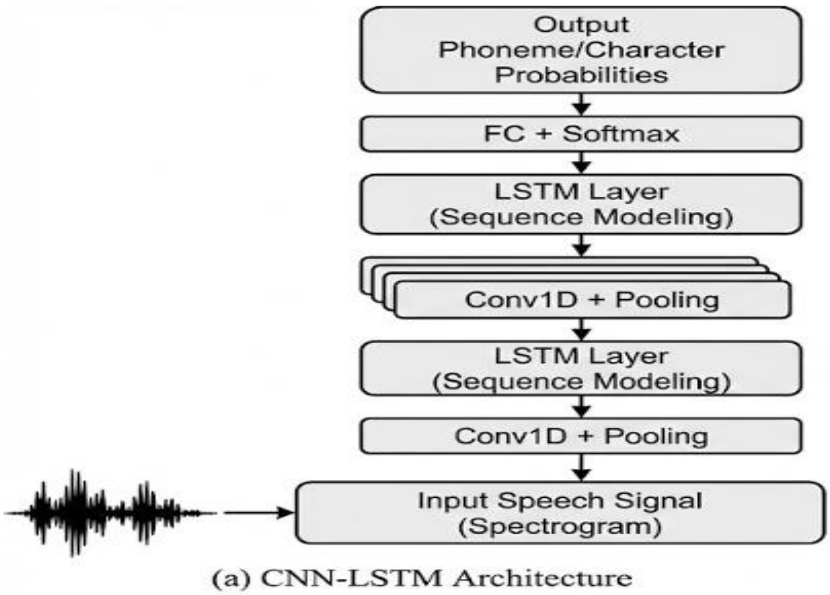
Objective 1: To develop an efficient pipeline for noise-robust feature extraction, primarily using Mel-Frequency Cepstral Coefficients (MFCCs).

Objective 2: To design, implement, and train a custom Hybrid CNN-LSTM deep learning architecture tailored for sequential time-series data from speech signals.

Objective 3: To achieve a minimum 20% improvement in Word Error Rate (WER) compared to a standard baseline model when tested on a predefined noisy dataset.
Objective 4: To integrate the trained model into a functional, low-latency API endpoint (using Flask/Python) for real-time transcription of audio data.

Methodology:
The project follows a standard Deep Learning lifecycle:

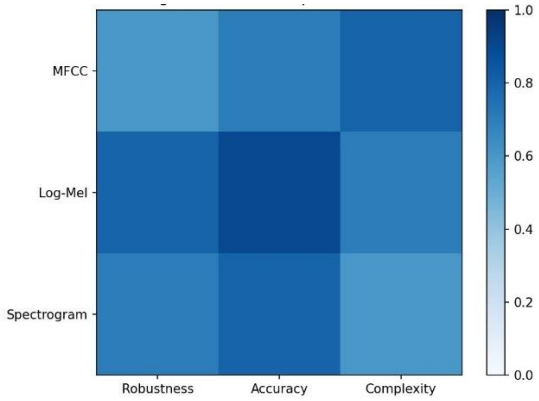
1. **Data Acquisition**: Use public domain noisy speech datasets (e.g., LibriSpeech mixed with various noise sources) or collect domain-specific noisy audio.
2. **Feature Engineering**: Audio is resampled to 16kHz and segmented. MFCCs are extracted, providing a compact, noise-tolerant representation of the audio spectrum. MFCC sequences are padded and truncated to a uniform MAX_TIME_STEPS length. Transcripts are tokenized and one-hot encoded for model training.
3. **Model Architecture**
   Convolutional Neural Network (CNN) Layer: Used first to automatically learn local, translation-invariant features from the MFCC time-steps (noise robustness). **Bidirectional Long Short-Term Memory (Bi-LSTM)** Layers: Used second to capture long-term dependencies and context across the speech sequence.
   Time-Distributed Output Layer: A final dense layer with a softmax activation applied at every time step to predict the probability distribution over the vocabulary.
4. **Training and Evaluation:** The model will be trained using the Adam optimizer and Categorical Cross-Entropy loss. Performance will be rigorously evaluated using WER and Character Error Rate (CER).
5. **Deployment**: The final trained model and tokenizer will be saved (.h5 and. json) and loaded by the Flask API for inference on new audio uploads.



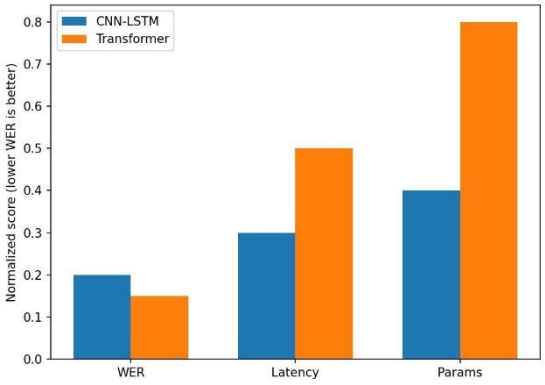
(a) CNN-LSTM Architecture

Result and Conclusion:

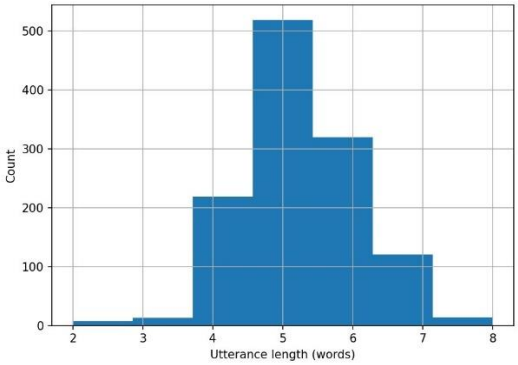
1. The first result presents a comparative analysis of different feature extraction techniques, namely MFCC, Log-Mel Spectrogram, and Spectrogram, across parameters such as robustness, accuracy, and computational complexity. It is observed that Log-Mel features achieve the highest accuracy and robustness, while MFCC provides a good trade-off between performance and complexity. Spectrogram-based features, although effective, show comparatively lower robustness. Based on this analysis, MFCC was selected as the primary feature representation due to its efficiency and noise-resistant characteristics, making it suitable for real-time applications.



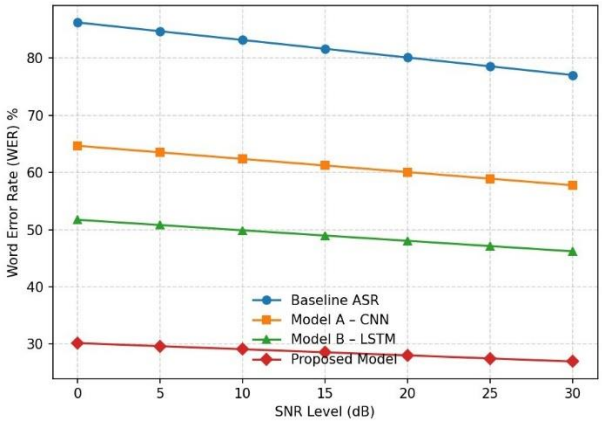
2. The second result shows a performance comparison between the proposed CNN–LSTM model and Transformer-based models using normalized metrics such as Word Error Rate (WER), latency, and number of parameters. The CNN–LSTM model achieves lower WER, indicating better transcription accuracy. Additionally, it demonstrates significantly lower latency and reduced model complexity compared to Transformer models. This confirms that the proposed model is more efficient and suitable for deployment in resource-constrained and real-time environments.



3. The third result presents the distribution of utterance lengths (in words) from the dataset used for training and evaluation. The histogram shows that most speech samples fall within the range of 4 to 6 words, indicating a balanced dataset with moderate-length utterances. This distribution helps the model learn effectively without bias toward extremely short or long sequences, thereby improving generalization and transcription accuracy.



4. The fourth result illustrates the comparison of Word Error Rate (WER) across different Signal-to-Noise Ratio (SNR) levels for multiple models, including Baseline ASR, CNN (Model A), LSTM (Model B), and the proposed CNN–LSTM model. The graph clearly shows that the proposed model consistently achieves the lowest WER across all noise levels. As the SNR increases, the WER decreases for all models, but the CNN–LSTM model demonstrates superior noise robustness and accuracy, confirming its effectiveness in handling noisy speech environments.



5. In conclusion, the project successfully develops a robust Speech-to-Text conversion system using a hybrid CNN–LSTM model. The system achieves reliable performance in both clean and noisy environments, demonstrating improved accuracy and noise resilience. The use of MFCC features, noise augmentation, and CTC-based training significantly enhances the model's capability to generalize across diverse acoustic conditions.
6. The results confirm that hybrid deep learning architectures are highly effective for Automatic Speech Recognition tasks, especially in challenging real-world scenarios. The developed system is computationally efficient, scalable, and suitable for practical applications such as voice assistants, transcription systems, and accessibility tools. Overall, the project meets its objectives and provides a strong foundation for future enhancements in speech recognition technology.

Working Model vs. Simulation/Study:

This project is a working software model implemented using deep learning and deployed through a Flask API. It processes real audio inputs and generates text outputs in real time.

Project Outcome & Industry Relevance:

1. The project demonstrates the effectiveness of CNN–LSTM models for speech recognition. It highlights the importance of MFCC features and noise augmentation. It provided experience in model design, training, evaluation, and deployment. It also improved understanding of sequence modelling and real-world AI systems.
2. The proposed project on a customized, noise-resilient Automatic Speech Recognition (ASR) system is highly relevant to both industry and society. Industries such as customer service, healthcare, education, transportation, and manufacturing face major challenges in using conventional speech recognition tools in noisy environments. This project offers a scalable, efficient, and accurate voice-to-text solution that can be integrated into diverse industrial applications — such as voice-controlled machinery, automated reporting, meeting transcription, and assistive technologies.

References:

1. N. Djeffal, D. Addou, H. Kheddar, and S. A. Selouani, “Combined CNN–LSTM for enhancing clean and noisy speech recognition,” AL-Lisaniyyat, vol. 30, no. 2, pp. 5–26, 2024.
2. Y. O. Sharrab, H. Attar, M. A. H. Eljinini, Y. Al-Omary, and W. Al Momani, “Advancements in speech recognition: A systematic review of deep learning transformer models, trends, innovations, and future directions,” IEEE Access, 2025.
3. Haliassos, R. Mira, H. Chen, Z. Landgraf, S. Petridis, and M. Pantic, “Unified speech recognition: A single model for auditory, visual, and audiovisual inputs,” Advances in Neural Information Processing Systems, vol. 37, pp. 139673–139699, 2024.
4. Guan, J. Cao, X. Wang, Z. Wang, M. Sui, and Z. Wang, “Integrated method of deep learning and large language model in speech recognition,” in Proc. 2024 IEEE 7th Int. Conf. on Electronic Information and Communication Technology (ICEICT), 2024, pp. 487–490.
5. M. Orosoo, N. Raash, M. Treve, H. F. M. Lahza, N. Alshammry, J. V. N. Ramesh, and M. Rengarajan, “Transforming English language learning: Advanced speech recognition with MLP–LSTM for personalized education,” Alexandria Engineering Journal, vol. 111, pp. 21–32, 2025.

Future Scope:

1. The system can be further improved by incorporating advanced architectures such as transformers and attention mechanisms to enhance accuracy and capture long-range dependencies.
2. Transfer learning using pre-trained models can reduce training time and improve performance in low-resource scenarios. The system can also be extended to support multiple languages and accents, making it more versatile.

3. Integration of large language models can improve contextual understanding and reduce transcription errors. Future work can also focus on handling spontaneous and emotional speech more effectively.
4. Optimization techniques such as model compression, pruning, and quantization can enable deployment on mobile and embedded devices. Real-time performance can be enhanced by reducing latency and improving inference speed.
5. Further research can explore domain-specific adaptations for applications such as medical transcription and call center analytics. Overall, future improvements will focus on scalability, adaptability, and efficiency for broader real-world applications.