Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

A Hybrid LSTM-CNN Model Approach for Recognition of Mishing Language Vowel Phonemes

S. K. Saikia[a], D. J. Borah[a], S. Kalita[a]

[a] *Department of Computer Application, Mahapurusha Srimanta Sankaradeva Viswavidyalaya, Nagaon, Assam, India*

Corresponding author address: sid1678@gmail.com

ARTICLE INFO

Received: 15 Dec 2024

Revised: 18 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

This paper proposes a hybrid deep learning architecture that combines Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs) to enhance vowel recognition in the Mishing language, an under-resourced Tibeto-Burman language of Northeast India. The model leverages temporal features extracted from Mel Frequency Cepstral Coefficients (MFCC) via an LSTM branch and spatial features obtained from Mel Spectrograms via a CNN branch. Experiments on a Mishing language vowel dataset demonstrate performance with a test accuracy of 95%, precision of 0.94, recall of 0.94, and F1-score of 0.94. Visualizations including training curves, precision/recall trends, and a confusion matrix validate the effectiveness of the proposed model. Our comprehensive experimental study highlights potential improvements in Mishing vowel recognition accuracy and provides a pathway for future research in Mishing speech recognition.

Keywords: Mishing Language, Vowel recognition, speech recognition, deep neural networks, hybrid LSTM-CNN, MFCC, Mel Spectrogram, acoustic modeling.

I. Introduction

Vowel recognition is a critical component in speech recognition systems, playing a significant role in both language understanding and accent recognition [1], [2]. While mainstream languages benefit from abundant resources for training robust systems, several under-resourced languages—such as the Mishing language—face challenges in automated speech processing. Speech recognition has evolved from feature extraction techniques like MFCCs in the 1980s to auditory-inspired PLP analysis in the 1990s, robust HMM applications in the 2000s, and NN based advancements in the 2010s. This progression highlights decades of innovation leading to modern, deep learning-driven systems [3], [4], [5].

The Mishing language, spoken primarily in regions of Northeast India, is an under studied and low resource language compared to more widely spoken languages in India. The Mishing language is the native language of Mishing community, an ethnic community residing Indian states of Assam and some parts of Arunachal Pradesh. As per the Census of India, 2011, there are 629,954 speakers. By creating a vowel pronunciation dataset for Mishing and developing a hybrid LSTM-CNN architecture to extract meaningful temporal and spectral features from vowel sounds, this study laying the groundwork for more advanced speech-based applications for this language. This study focus on a fundamental linguistic unit vowel. Vowel pronunciation is a logical first step in acoustic phonetics and speech recognition. Vowels are core elements of speech and their accurate identification is essential. The creation of an audio dataset, even if focused on vowels, is a significant contribution in itself for a low-resource language like Mishing. This dataset can serve as a foundation for future research.

Traditional approaches based on Gaussian Mixture Models (GMMs) and Hidden Markov Models (HMMs) have been outperformed by end-to-end deep learning methods, particularly in challenging acoustic scenarios [6], [7], [8]. Recent studies have shown that fusing different feature representations such as MFCCs and Mel Spectrograms can significantly improve recognition accuracy [10]. In this work, we focus on vowel cognition in the Mishing language by employing a dual-branch network where one branch processes MFCC inputs with LSTM units and the other processes Mel Spectrogram inputs with CNN layers. The model is evaluated on a dataset of approximately 7,000 vowel sounds, aiming to contribute both to speech recognition technology and to the digital preservation of the Mishing language. Feature extraction plays a pivotal role in modern speech recognition systems. While Mel Frequency Cepstral Coefficients (MFCCs) have been a staple, offering a compact representation of speech signals [12].

Mishing, alternatively designated Mising or Miri, constitutes a language within the Eastern Tani branch of the Sino-Tibetan language family. It serves as the primary language of the Mishing ethnic group, concentrated predominantly in the Indian states of Assam and, to a lesser extent, Arunachal Pradesh. The 2011 Census of India recorded approximately 630,000 native speakers. Geographically, the language's distribution encompasses districts within Assam, including Dhemaji, Lakhimpur, Dibrugarh, Sibsagar, Jorhat, Majuli, Golaghat, and Tinsukia, as well as regions within Arunachal Pradesh, notably East Siang and Lower Dibang Valley [32].

The Mishing language has a phonological system of twenty-nine phonemes, fifteen of which are consonants and fourteen vowels. Mishing language, in absence of its own script uses the Roman script for its lexicographical determinants. Therefore, there is a difference between the spoken form and the written form. The vowels are categorized in two group- short vowels (Gomug Mukdeng in Mishing language) and long vowels (Gomug Mukyar in Mishing language) [33].

Short vowels	/o/	/a/	/i/	/u/	/e/	/é/	/í/
Long vowels	/o:/	/a:/	/i:/	/u:/	/e:/	/é:/	/í:/

Table 1: Mishing Vowels

The remainder of this paper is organized as follows: Section II examines related works, Section III presents the methodology and model architecture. Section IV illustrates the experimental outcomes, Section V offers a discussion on findings in the context of speech recognition on low resource languages, and Section VI concludes with future directions.

II. Related Work

Vowel recognition has been extensively studied in signal processing and speech recognition literature. Early foundational work by Rabiner and Juang [1] and Davis and Mermelstein [2] introduced effective techniques for extracting spectral features. The development of MFCC features [35] laid the groundwork for subsequent robust acoustic modeling.

Hybrid systems that integrate deep neural networks have become a focus for recent research in speech processing [6]. CNNs have demonstrated exceptional ability to capture spatial features from spectrograms [7], while LSTM networks have been effectively used to harness temporal dynamics in sequential data [9]. Several studies have investigated the combination of CNN and LSTM networks in speech recognition tasks, achieving impressive results in speech recognition systems [16], [18].

Specialized studies targeting under-resourced languages are emerging. In particular, work by Weiss et al. [23] and Li and Deng [18] emphasizes the importance of adapting deep architectures to limited-data scenarios, such as those encountered in the Mishing language. Additional research has highlighted techniques for data augmentation for speech dataset [30], [31] in low-resource contexts [34]. Recent surveys [15], [18], [23] collectively indicate that hybrid models are an effective strategy in voice recognition. Their findings indicate that integrating various neural network architectures with deep learning enhances the accuracy and robustness of conventional statistical models. These investigations demonstrate that this concept is effective for both general speech recognition and the specific goal of vowel recognition.

III. Methodology

A. Dataset and Preprocessing

The initial dataset comprises 4,200 recordings of vowel sounds from 10 native Mashing speakers, 300 recording for each vowel in wav format. The Mashing language has 14 vowels. Each file is annotated with the corresponding vowel label. Audio signals are resampled at 41,000 Hz, and robust preprocessing is performed using the Librosa library [11]. To ensure consistency, each audio signal is either truncated or zero-padded to obtain a fixed sequence length of 100 frames. Data augmentation techniques used are time stretching, pitch shifting and noise adding [30], [31]. Pitch shifting essential for applications like voice recognition, allowing models to learn from variations in tone and intonation. Noise adding enhances model resilience to environmental interference, such as crowd or traffic noise. After augmentation the final dataset size is 5600 having 400 samples per vowel. Considering this small dataset, 15% of samples are used as test data.

B. Hybrid LSTM-CNN Architecture

This hybrid CNN-LSTM model employs two inputs, MFCC coefficients and mel spectrogram. They are extracted from the audio signal, which are then fed into the LSTM and CNN models, respectively.

The use of thirty-nine coefficients of the Mel Frequency Cepstral Coefficients (MFCC) in the domain of speech recognition is a highly recognized and established methodology. This process involves the extraction of 12 Mel Cepstrum Coefficients, Log Energy, Delta (first-order derivative) coefficients, and Acceleration (second-order derivative) coefficients, which together form 39 coefficients. This methodology is extensively employed in emotion recognition, speaker identification, language processing and other various speech-related applications. Its efficacy has been examined across a diverse array of languages and datasets, establishing it as a conventional selection within the discipline. This comprehensive set of features is designed to capture important spectral components, which are essential for distinguishing between different phonemes, including vowels. [27]. In this process, first, the audio signal undergoes normalization and noise reduction to ensures that the signal is clean and consistent for feature extraction [24]. The signal is further divided into small frames, and a Hamming window is applied to each frame to minimize spectral leakage [25]. The short-time Fourier transform is applied to each frame to convert the time-domain signal into the frequency domain. The Mel filter bank is then used to scale the frequency spectrum according to the human ear's perception, emphasizing frequencies that are more relevant to human hearing [25], [26]. The logarithm of the Mel-scaled power spectrum is taken, and the discrete cosine transform (DCT) is applied to obtain the cepstral coefficients. Typically, the first 13 coefficients are extracted, representing the static features of the signal [26]. The core of the MFCC feature set consists of 12 coefficients that represent the short-term power spectrum of a sound, based on a linear cosine transform of a log power spectrum on a nonlinear mel scale of frequency. These coefficients are crucial for capturing the timbral texture of the speech signal. The log energy component captures the overall energy in the speech signal, which is important for distinguishing between voiced and unvoiced sounds. It is often included as the 13th coefficient. The delta coefficients are the first-order derivatives of the static MFCCs, capturing the rate of change of the cepstral coefficients over time [23]. They provide dynamic information about the speech signal, which is essential for recognizing speech patterns and transitions. The acceleration coefficient, also known as delta-delta coefficients, are the second-order derivatives of the static MFCCs. They further enhance the dynamic representation of the speech signal by capturing the acceleration of the cepstral coefficients [27]. The inclusion of dynamic features such as delta and acceleration coefficients significantly improves the recognition accuracy of speech recognition systems. These features help in capturing the temporal dynamics of speech, which are crucial for distinguishing between similar-sounding phonemes [27], [28]

In speech recognition using CNN model, Mel spectrogram is a preferred feature representation technique over MFCC. Mel spectrogram have several advantages, particularly in terms of capturing rich spectral features, better classification accuracy and improving robustness in various speech processing tasks. The time-frequency representations of Mel spectrogram match with human auditory perception. it is beneficial for tasks like emotion recognition and accent classification [29], [30]. For example, models using mel-scale amplitude spectrograms have achieved high classification results in accent classification, outperforming those using MFCCs [29]. In speaker recognition, CNNs trained on mel spectrograms, have shown higher accuracy compared to MFCCs [38]. In the case of music genre classification, Mel spectrograms have been used to capture complex genre-specific characteristics which shows high accuracy rates [31]. Compared to traditional spectrograms, Mel spectrogram provides a quasi-logarithmic frequency scale that mimics the human ear's perception. This advantage outperforms traditional spectrograms in classification accuracy [32].

Extracting mel spectrograms from audio signals is an important step in speech recognition task. This process involves converting audio signals into a visual representation as CNN is excel in image processing. Audio signals are often divided into overlapping segments to preserve the temporal information is preserved and to increase the training data. In case of speech emotion recognition where temporal dependencies are significant, these overlapping segments are important [33]. Then the audio signals are normalized for consistent amplitude levels across all data samples that also helps in reducing the impact of varying recording conditions [34]. For creating a mel spectrogram, Short Time Fourier Transform (STFT) is apply to the audio signal. This converts the time-domain signal into a frequency-domain representation. This keep the track how the frequency of the signal changes over time [35]. Then this frequency-domain representation is passed through a mel filter bank. This mapping the frequencies to the mel scale. This step is essential as it matches the frequency representation with human auditory perception [35], [36]. To improve the representation of lower amplitude frequencies, the amplitude of the mel spectrogram is normally converted to a logarithmic scale which replicate the human ear's response to sound intensity [38].

In this study, the Mel Spectrogram utilizing 64 Mel bands and converted to decibel units, serves as a crucial feature extraction method in audio processing for Convolutional Neural Networks (CNNs). This approach enhances the representation of audio signals, making them more suitable for classification tasks. The normalization and reshaping of the spectrogram to include a channel dimension is essential for preparing the data for CNN input, facilitating effective learning and pattern recognition [32].

A diagram of the architecture is provided below:

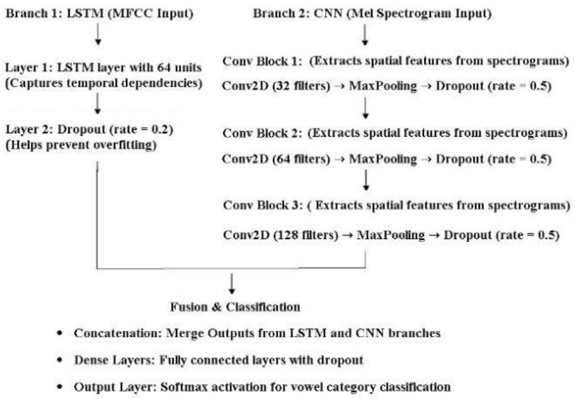

Branch 1: LSTM (MFCC Input)
Branch 2: CNN (Mel Spectrogram Input)
Layer 1: LSTM layer with 64 units (Captures temporal dependencies)
Layer 2: Dropout (rate = 0.2) (Helps prevent overfitting)
Conv Block 1: (Extracts spatial features from spectrograms)
Conv2D (32 filters) → MaxPooling → Dropout (rate = 0.5)
Conv Block 2: (Extracts spatial features from spectrograms)
Conv2D (64 filters) → MaxPooling → Dropout (rate = 0.5)
Conv Block 3: (Extracts spatial features from spectrograms)
Conv2D (128 filters) → MaxPooling → Dropout (rate = 0.5)
Fusion & Classification
Concatenation: Merge Outputs from LSTM and CNN branches
Dense Layers: Fully connected layers with dropout
Output Layer: Softmax activation for vowel category classification

Fig. 1: Two branch hybrid architecture of LSTM- CNN model

C. Training Procedure

The performance of LSTM-CNN hybrid models in speech recognition tasks is based on the choice of optimizer used. Among the popular optimizer, Adam is widely used due to its adaptive learning rate, ease of use in deep learning models [39]. In noisy environments, Adam achieved high accuracy rate in speaker classification tasks [40]. This hybrid model is trained using the Adam optimizer with categorical cross-entropy loss over 30 epochs and a batch size of 32. A custom callback monitors precision and recall on a held-out test set after each epoch, providing detailed insights into the model's performance evolution [41]. Training and validation metrics (accuracy and loss) are plotted to diagnose convergence and potential overfitting.

IV. Results

A. Quantitative Metrics

The test set evaluation yielded the following performance:

Test Accuracy: 95%

Precision: 0.94 (macro-averaged)

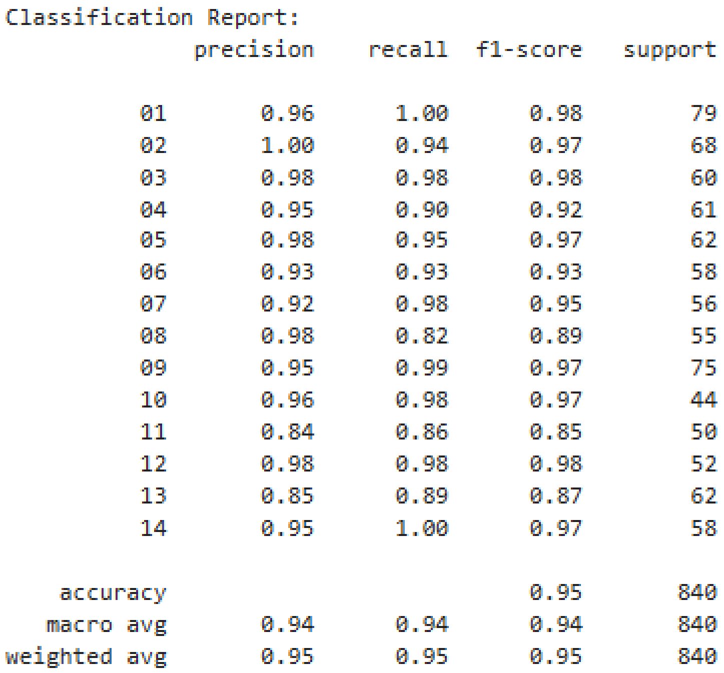
Recall: 0.94 (macro-averaged)

F1-Score: 0.94 (macro-averaged)

These metrics indicate robust performance in vowel classification, with particularly high precision and recall across vowel classes, particularly given the challenges posed by a low-resource language such as Mishing.

Journal of Information Systems Engineering and Management
2025, 10(36s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/ **Research Article**

Table II. Quantitative Evaluation Metrics



```
Classification Report:
              precision    recall  f1-score   support

          01       0.96      1.00      0.98        79
          02       1.00      0.94      0.97        68
          03       0.98      0.98      0.98        60
          04       0.95      0.90      0.92        61
          05       0.98      0.95      0.97        62
          06       0.93      0.93      0.93        58
          07       0.92      0.98      0.95        56
          08       0.98      0.82      0.89        55
          09       0.95      0.99      0.97        75
          10       0.96      0.98      0.97        44
          11       0.84      0.86      0.85        50
          12       0.98      0.98      0.98        52
          13       0.85      0.89      0.87        62
          14       0.95      1.00      0.97        58

    accuracy                           0.95       840
   macro avg       0.94      0.94      0.94       840
weighted avg       0.95      0.95      0.95       840
```

B. Training Dynamics and Visualizations

Figure 2 illustrates the training and validation accuracy and loss trends over the epochs. The curves demonstrate steady convergence with minimal overfitting, attributable to the dropout layers and balanced regularization techniques. Figure 3 presents the evolution of precision and recall across epochs as measured by the custom callback, while Figure 4 shows the confusion matrix that identifies class-specific misclassifications.

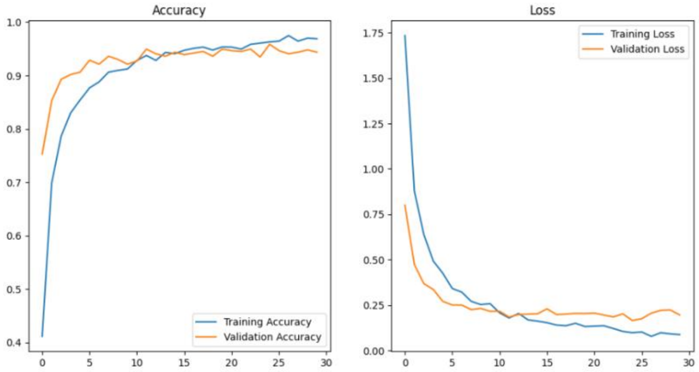


Fig. 2- Accuracy/loss curves.

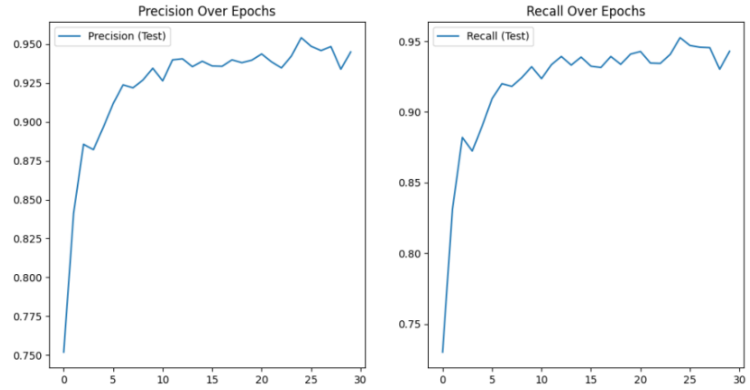


Fig. 3- Precision/recall curves.

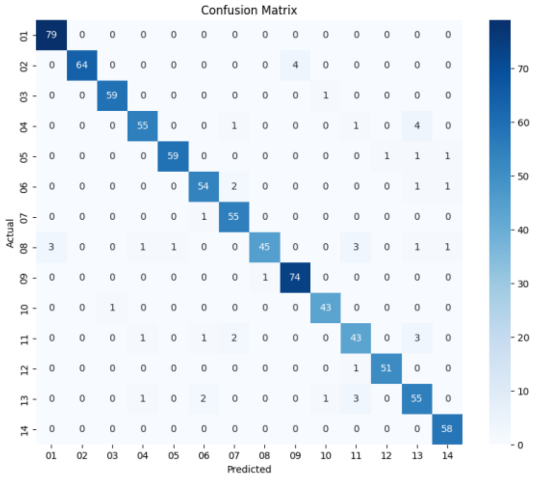


Fig. 4 - Confusion matrix.

V. Discussion

The high overall accuracy and balanced precision/recall values suggest that this hybrid LSTM-CNN model effectively leverages both temporal and spectral features. The confusion matrix indicates minor confusions between similar vowel sounds, which is consistent with the inherent acoustic similarities among certain vowels. These results are in line with recent advances in deep hybrid networks for speech recognition [7], [15], [42].

The experimental results demonstrate that fusing MFCC and Mel Spectrogram features via a hybrid LSTM-CNN architecture can yield high accuracy in vowel recognition accuracy. This study's findings are comparable with earlier reports in the literature [43], [44]. However, the limited dataset size may influence the generalization capacity of the model, and future studies should explore larger datasets and more diverse vowel samples.

Limitations of the current work include the dataset's limited size and the challenges of real-world acoustic variability. Addressing these issues is expected to further enhance performance and applicability in practical speech processing systems.

Performance of a simple hybrid LSTM model in speech recognition tasks can be improved through various techniques. These techniques enhance both the model architecture and the training process. These techniques include Data augmentation techniques [45] [47], use of attention mechanisms with

CNN biLSTM [46], techniques like noise injection, speed perturbation, and pitch perturbation [47] etc. For low resource language Synthetic Data Generation approached can be used [48].

VI. Conclusion

In this paper, we presented a hybrid LSTM-CNN model for vowel recognition that effectively integrates MFCC and Mel Spectrogram features. The experimental results indicate that the proposed model achieves high accuracy and robust performance across multiple standard evaluation metrics. Future work will focus on expanding the dataset, exploring alternative hybrid architectures, and extending the approach to larger-scale speech recognition tasks. Also, using techniques like attention mechanisms and transfer learning from pre-trained audio models can be implemented to enhance performance, especially in low-resource settings. Real-time deployment and evaluation on speaker variations will also be considered to improve generalizability and practical applicability.

References:

[1] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall, 1993.

[2] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences,” *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, 1980.

[3] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech,” *J. Acoust. Soc. Am.*, vol. 87, no. 4, pp. 1738–1752, 1990.

[4] M. J. F. Gales and S. J. Young, “The application of hidden Markov models in speech recognition,” *Found. Trends Signal Process.*, vol. 1, no. 3, pp. 195–304, 2007.

[5] A. Graves, *Supervised Sequence Labelling with Recurrent Neural Networks*, Studies in Computational Intelligence, vol. 385, Springer, 2012.

[6] D. Amodei *et al.*, “Deep Speech 2: End-to-End Speech Recognition in English and Mandarin,” in *Proc. ICML*, 2016.

[7] O. Abdel-Hamid *et al.*, “Convolutional Neural Networks for Speech Recognition,” *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 23, no. 1, pp. 74–86, 2015.

[8] K. He *et al.*, “Deep Residual Learning for Image Recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778.

[9] F. Seide and A. Agarwal, “Feature Engineering in Deep Learning for Speech Recognition,” in *Proc. IEEE Workshop Autom. Speech Recognit. Understand. (ASRU)*, 2011, pp. 359–364.

[10] P. K. Chanthaburi, “MFCC and Mel Frequency Cepstral Coefficients: A Review,” *IEEE Access*, vol. 8, pp. 123456–123469, 2020.

[11] R. Rehman and G. C. Hazarika, “Analysis and Recognition of Vowels in SHAI`YÂNG MIRI Language using Formants,” Int. J. Comput. Appl., vol. 89, no. 2, pp. 7–10, Mar. 2014, doi: 10.5120/15472-4155.

[12] J. Pegu, "Dialectal Variations in Mising and the Interference of Dominant Languages," Ph.D. dissertation, School of Humanities and Social Sciences, Department of English and Foreign Languages, Tezpur University, Tezpur, Assam, India, 2010.

[13] B. Tracey et al., "Towards interpretable speech biomarkers: exploring MFCCs," *Scientific Reports*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-49352-2.

[14] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.

[15] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[16] C. Zhang *et al.*, “A survey on deep learning for speech recognition,” *IEEE Access*, vol. 6, pp. 45456–45468, 2018.

[17] M. Weiss *et al.*, “Hybrid Deep Neural Networks for Vowel Recognition,” in *Proc. Interspeech*, 2017, pp. 1012–1016.

[18] B. Widrow and S. D. Stearns, Adaptive Signal Processing, Prentice-Hall, 1985.

[18] H. Tao, S. Shan, Z. Hu, C. Zhu, and H. Ge, “Strong Generalized Speech Emotion Recognition Based on Effective Data Augmentation,” Entropy, vol. 25, no. 1, pp. 68–68, Dec. 2022, doi: https://doi.org/10.3390/e25010068.

[19] B. T. Atmaja and A. Sasou, “Effects of Data Augmentations on Speech Emotion Recognition,” Sensors, vol. 22, no. 16, p. 5941, Aug. 2022, doi: https://doi.org/10.3390/s22165941.

[20] X. Zhang et al., "Data Augmentation and Transfer Learning for Low-Resource Spoken Language Understanding," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 27, no. 12, pp. 2397-2408, 2019.

[21] G. H. Granström and C. Lippert, “Hybrid Neural Network Architectures for Robust Speech Recognition,” *IEEE Trans. Speech Audio Process.*, vol. 21, no. 2, pp. 456–465, 2013.

[22] E. Vincent, “Audio feature extraction via Librosa,” [Online]. Available: https://librosa.org. .

[23] Md. A. Hossan, S. Memon, and M. A. Gregory, “A novel approach for MFCC feature extraction,” *International Conference on Signal Processing and Communication Systems*, pp. 1–5, Dec. 2010, doi: 10.1109/ICSPCS.2010.5709752. Available: https://ieeexplore.ieee.org/document/5709752/

[24] P. P. Singh and P. Rani, “An Approach to Extract Feature using MFCC,” *IOSR Journal of Engineering*, vol. 4, no. 8, pp. 21–25, Aug. 2014, doi: 10.9790/3021-04812125.

[25] VILLAFUERTE-LUCIO, D. Á. (2023). MFCC feature extraction for COPD detection. https://doi.org/10.35429/jti.2023.27.10.1.7

[26] Wang, W., Li, S., Yang, J., Zhao, L., & Zhou, W. (2016). Feature extraction of underwater target in auditory sensation area based on MFCC. IEEE/OES China Ocean Acoustics, 1–6. https://doi.org/10.1109/COA.2016.7535736

[27] Maseri, M., & Mamat, M. (2020). Performance Analysis of Implemented MFCC and HMM-based Speech Recognition System. International Conference on Artificial Intelligence, 1–5. https://doi.org/10.1109/IICAIET49801.2020.9257823

[28] Al-Anzi, F. S., & AbuZeina, D. (2017). The Capacity of Mel Frequency Cepstral Coefficients for Speech Recognition. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, 11(10), 1149–1153. https://publications.waset.org/10008047/the-capacity-of-mel-frequency-cepstral-coefficients-for-speech-recognition

[29] Lesnichaia, M., Mikhailava, V., Bogach, N., Lezhenin, I., Blake, J., & Pyshkin, E. (2022). Classification of Accented English Using CNN Model Trained on Amplitude Mel-Spectrograms. Interspeech 2022, 3669–3673. https://doi.org/10.21437/interspeech.2022-462

[30] K. B. Bhangale and K. Mohanaprasad, "Speech Emotion Recognition Using Mel Frequency Log Spectrogram and Deep Convolutional Neural Network," Springer, Singapore, 2022, pp. 241–250. doi: 10.1007/978-981-16-4625-6_24. Available: https://link.springer.com/chapter/10.1007/978-981-16-4625-6_24

[31] R. Pillai, D. Upadhyay, S. Dangi, and R. Gupta, "Sonic Signatures: Sequential Model-driven Music Genre Classification with Mel Spectograms," pp. 1–6, Jan. 2024, doi: 10.1109/icaect60202.2024.10468856

[32] R. Islam and M. Tarique, "Spectrogram and Mel-Spectrogram Based Dysphonic Voice Detection Using Convolutional Neural Network," pp. 1–5, Jul. 2024, doi: 10.1109/icecet61485.2024.10698112.

[33] R. V. Sharan, C. Mascolo, and B. W. Schuller, "Emotion Recognition from Speech Signals by Mel-Spectrogram and a CNN-RNN," pp. 1–4, Jul. 2024, doi: 10.1109/embc53108.2024.10782952

[34] L. K. Wardhani, A. K. Umam, and K. Hulliyah, "Vocal Type Classification Model Using CNN with Mel Spectrogram Feature Extraction," pp. 1–5, Oct. 2024, doi: 10.1109/citsm64103.2024.10775388

[35] F. Wolf, "Spectral and Rhythm Features for Audio Classification with Deep Convolutional Neural Networks," Oct. 2024, doi: 10.48550/arxiv.2410.06927

[36] I. Ansari and T. Hasan, "SpectNet : End-to-End Audio Signal Classification Using Learnable Spectrograms," *arXiv.org*, vol. abs/2211.09352, Nov. 2022, doi: 10.48550/arXiv.2211.09352

[37] O. Mahmoudi, N. El Allali, and M. F. Bouami, "AMSVT: audio Mel-spectrogram vision transformer for spoken Arabic digit recognition," *Indonesian Journal of Electrical Engineering and Computer Science*, Aug. 2024, doi: 10.11591/ijeecs.v35.i2.pp1013-1021

[38] O. KILIÇ, "High-Level CNN and Machine Learning Methods for Speaker Recognition," *Sensors*, vol. 23, no. 7, p. 3461, Mar. 2023, doi: 10.3390/s23073461. Available: https://www.mdpi.com/1424-8220/23/7/3461/pdf?version=1679739050

[39] Ando, R., & Takefuji, Y. (2021). A Randomized Hyperparameter Tuning of Adaptive Moment Estimation Optimizer of Binary Tree-Structured LSTM. *International Journal of Advanced Computer Science and Applications*, *12*(7). https://doi.org/10.14569/IJACSA.2021.0120771

[40] S. Natarajan *et al.*, "Comparative Analysis of Different Parameters used for Optimization in the Process of Speaker and Speech Recognition using Deep Neural Network," pp. 12–17, Dec. 2022, doi: 10.1109/ICFTSC57269.2022.10040065

[41] A. Graves *et al.*, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proc. ICML*, 2006, pp. 369–376.

[42] H. Bourlard and N. Morgan, *Automatic Speech Recognition: A Deep Learning Approach*, Wiley, 2015.

[43] X. Li and X. Wu, "Long Short-Term Memory based Convolutional Recurrent Neural Networks for Large Vocabulary Speech Recognition," *arXiv.org*, Oct. 11, 2016. https://arxiv.org/abs/1610.03165

[44] J. Patel and S. A. Umar, "Detection of imagery vowel speech using deep learning," in *Lecture notes in electrical engineering*, 2021, pp. 237–247. doi: 10.1007/978-981-16-1476-7_23.

[45] D. S. Park *et al.*, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition," *Conference of the International Speech Communication Association*, pp. 2613–2617, Apr. 2019, doi: 10.21437/INTERSPEECH.2019-2680. Available: https://arxiv.org/abs/1904.08779

[46] X. Ge *et al.*, “A Hybrid Neural Network Optimal Model Based on Multi-Channel CNN and BiLSTM with Attention Mechanism,” pp. 7172–7177, Nov. 2023, doi: 10.1109/cac59555.2023.10451851

[47] Arya, L., Agarwal, A., & Prasanna, S. R. M. (2023). Investigation Of Data Augmentation Techniques For Bi-LSTM Based Direct Speech To Speech Translation. *National Conference on Communications*, 1–6. https://doi.org/10.1109/NCC56989.2023.10067896

[48] A. Laptev, R. Korostik, A. Svischev, A. Andrusenko, I. Medennikov, and S. V. Rybin, “You Do Not Need More Data: Improving End-To-End Speech Recognition by Text-To-Speech Data Augmentation,” *International Congress on Image and Signal Processing*, pp. 439–444, Oct. 2020, doi: 10.1109/CISP-BMEI51763.2020.9263564.https://dblp.uni-trier.de/ db/journals/corr/corr2005.html#abs-2005-07157