ICIC Express Letters
Volume **18**, Number **9**, September **2024**

ICIC International ©2024 ISSN 1881-803X
pp. **987–995**

IMPROVEMENT IN SPEECH RECOGNITION OF INDONESIAN LANGUAGE USING MEL FREQUENCY CEPSTRAL COEFFICIENTS AND LONG SHORT-TERM MEMORY METHOD

Adriani[1], Ridwang[1,*], Agustan Syamsuddin[2], Muliadi[3] and Usman Umar[4]

[1]Department of Electrical Engineering
[2]Magister of Elementary Education, Postgraduate Program
Universitas Muhammadiyah Makassar
Jl. Sultan Alauddin No. 259, Makassar 90221, Indonesia
{ adriani; agustan }@unismuh.ac.id
*Corresponding author: ridwang@unismuh.ac.id

[3]Department of Informatics and Computer Engineering Education
Universitas Negeri Makassar
Jl. AP. Pettarani Makassar, Sulawesi Selatan 90222, Indonesia
muliadi7404@unm.ac.id

[4]Department of Medical Electrotechnology
Politeknik Muhammadiyah Makassar
Jl. DR. Ratulangi No. 101, Makassar, Sulawesi Selatan 90132, Indonesia
usmanumar@poltekkesmu.ac.id

Received September 2023; accepted December 2023

Abstract. *Speech recognition technology has witnessed significant advancements in recent years, revolutionizing the way humans interact with machines and devices. The Indonesian language, with its rich phonetic diversity, presents unique challenges for automatic speech recognition systems. This research aims to enhance the accuracy and efficiency of Indonesian speech recognition by employing Mel Frequency Cepstral Coefficients and Long Short-Term Memory neural networks. The study begins by collecting a comprehensive dataset of spoken Indonesian phrases from various speakers, capturing a wide range of dialects and accents. Preprocessing techniques are applied to clean and prepare the audio data, including noise reduction and feature extraction using MFCCs. These MFCCs are used to represent the spectral characteristics of the audio, providing a compact and informative input for subsequent recognition. The core of the research lies in the implementation of LSTM neural networks, a type of recurrent neural network (RNN) known for its ability to capture long-term dependencies in sequential data. The LSTM model is trained on the preprocessed audio data to learn the underlying patterns and relationships in the spoken Indonesian language. The model is fine-tuned through iterations to optimize its performance. Experimental results demonstrate a significant improvement in the accuracy and robustness of the Indonesian speech recognition system when compared to conventional methods. The incorporation of MFCCs and LSTM networks not only enhances the system's ability to handle diverse dialects but also increases its tolerance to background noise and speaker variations. The achieved recognition rates exhibit promising outcomes for practical applications in voice assistants, transcription services, and other voice-controlled technologies.*

Keywords: Speech, Recognition, MFCC, Indonesian language, LSTM

1. Introduction.

Speech recognition technology has made remarkable strides in recent years, revolutionizing the way humans interact with machines and devices [1]. From voice-activated personal assistants to automated transcription services [2], the applications of speech recognition are both diverse and pervasive [3]. However, achieving accurate and

DOI: 10.24507/icicel.18.09.987

efficient speech recognition in languages with complex phonetic structures and dialectal variations remains a significant challenge. Indonesia, as one of the world's most linguistically diverse countries, exemplifies this challenge with its multitude of regional dialects and accents. Indonesia, with its vast archipelago comprising thousands of islands, is home to over 700 living languages. Indonesian language, the standardized form of the language, serves as the official language of the nation and is widely spoken. Yet, the linguistic diversity of the country extends far beyond this standardized variant. Various regional languages and dialects coexist [4], each characterized by distinct phonetic nuances and pronunciation patterns. As a result, creating a robust and adaptable speech recognition system for Indonesian poses unique obstacles [3].

This research undertakes the ambitious task of improving the accuracy and efficacy of Indonesian speech recognition through the utilization of two key components: Mel Frequency Cepstral Coefficients [5] and Long Short-Term Memory [6] neural networks. By combining these technologies, we aim to address the challenges posed by the diverse phonetic landscape of Indonesian and contribute to the ongoing advancement of speech recognition systems. MFCCs have proven to be a powerful tool for audio signal processing [7]. It is used to extract essential spectral features from audio data, capturing critical information about the frequency content and spectral characteristics of the spoken language. These coefficients serve as a robust representation of speech signals and play a pivotal role in the subsequent stages of speech recognition [8].

LSTMs excel in capturing long-range dependencies within sequences, making them particularly well-suited for natural language and speech recognition tasks. Their ability to learn intricate patterns in sequential data, such as spoken language, positions them as a powerful tool for enhancing the accuracy and robustness of speech recognition systems [9]. This research centers on the integration of MFCCs and LSTM networks in the realm of Indonesian speech recognition. Our goal is to develop a system that harnesses the spectral information derived from MFCCs and utilizes the sequence modeling prowess of LSTMs to accommodate the intricacies of diverse Indonesian dialects, enhance accuracy in noisy environments, and ultimately close the communication divide between humans and machines for Indonesian speakers [10].

Recently, some studies have investigated MFCC methods and deep learning to detect audio signals. The authors propose an automatic speech recognition system based on convolutional neural networks (CNNs) and Mel Frequency Cepstral Coefficients (MFCCs). They investigate different deep models' architectures with various hyperparameters, such as dropout rate and learning rate. They collected the dataset from the Kaggle TensorFlow Speech Recognition Challenge. Each audio file in the dataset contains one word with one second of length, and the total number of words in the dataset is 30 categories, with one category for background noise. The dataset consists of 64,721 files, which are divided into 51,088 for the training set, 6,798 for the validation set, and 6,835 for the testing set. The authors evaluate three models with different hyperparameter configurations to choose the best model with higher accuracy. The highest accuracy achieved is 88.21% [11].

The researchers propose a novel approach to speech emotion recognition (SER) called MFF-SAug, which significantly improves the accuracy of emotion classification from human speech compared to traditional methods. The proposed method employs a combination of noise removal, white noise injection, pitch tuning, and feature extraction techniques to enhance speech representation learning and voice emotion classification. We utilize the contrastive loss function to maximize agreement between differently augmented samples in the latent space and reconstruct the loss of input representation, leading to improved accuracy prediction. The proposed method is evaluated using four benchmark datasets: RAVDESS, CREMA, SAVEE, and TESS, achieving impressive accuracy rates of 92.6%, 89.9%, 84.9%, and 99.6%, respectively [12].

In addition, [13] investigated 3D-CNN to recognize audio for Indian language modeling. In this research, a dataset was created by reproducing the sound of 20 words from conventional (ISL) to train this model. Consequently, propose a novel deep recurrent neural network (RNN)-based framework to detect depression and predict its severity level from speech. They extracted low-level and high-level audio features from audio recordings to predict the 24 scores of the Patient Health Questionnaire (PHQ) and the binary class of depression diagnosis. To address the small size of Speech Depression Recognition (SDR) datasets, they considered expanding training labels and transferring features. Their proposed approach outperforms state-of-the-art methods on the DAIC-WOZ database, achieving an overall accuracy of 76.27% and a root mean square error (RMSE) of 0.4 in assessing depression, while an RMSE of 0.168 is achieved in predicting depression severity levels [14]. The authors illustrated the Deep GRU method for audio recognition [15] in this study, which examined the seventh publicly available dataset. This dataset contains a significant number of samples and encompasses a broad spectrum of interactions. On cross-subject and cross-view tests of the NTU RGB+D dataset, it achieves recognition accuracy of 84.9 per cent and 92.3 per cent, respectively, and 100 per cent recognition accuracy on the Kaggle dataset [16].

However, none of the above-mentioned studies used MFCC with LSTM as a method of extraction of features and rapid classification on continuous speech. The main contributions of this paper are highlighted below.

1) The proposed method can be applied to large series datasets with higher accuracy and quick processes without additional computing layers to input data in LSTM.
2) The proposed method can be applied to identifying accents and dialects in Indonesian language translation systems with higher accuracy and faster extract data.

In this study, MFCC extracted spectrogram images into three-dimensional vector data, and LSTM can calcify audio signals quickly, so it can help predict words in Indonesian sentences. This work is split into four sections: Section 1 is the introduction; Section 2 outlines research method; Section 3 covers results and discussion, and Section 4 presents the conclusions.

2. **Research Method.**

2.1. **Architecture design.**

Figure 1 demonstrates the architectural design of the proposed system. According to [17], learning both spatial and temporal information for dynamic speech identification is difficult using handmade feature extraction methods. We presented a model to address this problem, as shown in Figure 1.

2.2. **Data collection.**

In this step, we recorded the voices of 15 different people to generate a diverse dataset. Users are divided into two groups: an adult group and a child group. In the adult group of 10 people with a loud voice compared to the children's group of 5 people, users have to follow some processes to eliminate bias in order to improve the dataset. The sequence of speeches recorded by each user consists of 5 seconds for each word. Each word must be spoken 20 times by the user. Each user must pronounce a word in a different tone for each experiment, adjusting the speed and intonation of their speech. However, during training data collection, we found some invalid data. Before pre-processing, we remove invalid data to improve classification.

To speed up the data collection process, the data collector used must be of high quality, and the data collected must consist of a human-generated speech sequence. In addition, speech is captured using microphones with a variety of training data collection scenarios: using sound intonation from loud to small, voice intonations from small to large, and also adjusting speaking speeds, sometimes fast and sometimes even slow. To incorporate modality data into the proposed model, microphones must be high quality and be used as the primary tool for recording data. Audio was extracted using MFCC during the

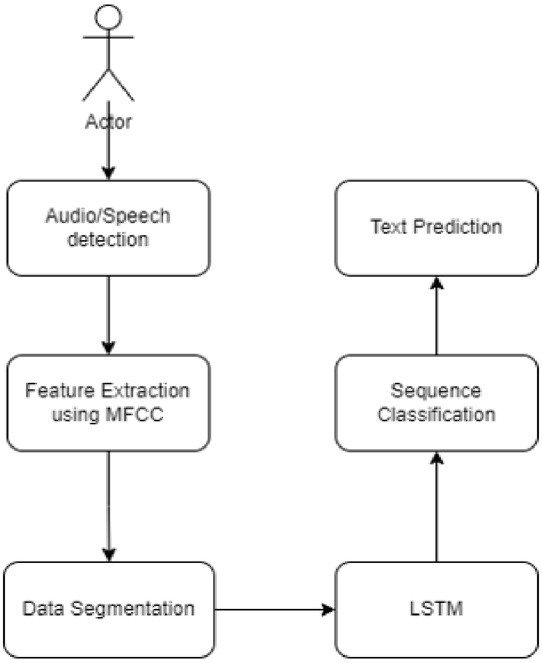


Figure 1. Architecture design of system

pre-training period. The result of the extraction is a vector that will be the input for the next process.

2.3. **Mel Frequency Cepstral Coefficients (MFCC).** The MFCC processing circuit consists of pre-emphasis, frame blocking, windowing, fast Fourier transform, mail frequency wrapping, and last cepstrum [9]. This method will generate data extracting sound characteristics in the form of mail vectors or voice characteristic vectors that will later be used as LSTM input data to create sound models [18]. Pre-emphasis is a phase of preprocessing that acts as a filter that can eliminate noise from recorded sounds and balance high and low frequencies in human voices. Frame blocking aims to divide a sound signal that has passed the pre-emphasis phase into several frames with a time interval of 25 milliseconds and a frame interval of 10 milliseconds. Windowing aims to reduce the signal discontinuity at the beginning and end of each frame. Suppose an audio file has 348 frames; then each frame will go through the windowing phase. FFT works to convert any frame from a time domain to a frequency domain. The FFT is a quick algorithm for the implementation of the discrete Fourier transform (DFT) operated on a discrete time signal consisting of N samples. The size of the FFT used in the research is 512; the result of the windowing will go through the process of FFT and will have a vector size of $512 * 348$ of the original $1102 * 348$. Mel frequency wrapping is aimed at applying a bank filter used to extract the strength of each band of frequencies, although the number of filters used is 26. Cepstrum is the final stage in MFCC and aims to normalize the results of a mail frequency wrap using the DCT algorithm to simplify the bank filter because not all the filters on the bank filter contribute to the uniqueness of the sound [5].

2.4. **LSTM-sequence classification method.** The LSTM method uses a recurrent neural network to effectively represent sequential data and identify out-of-the-ordinary values [19]. Recurrent neural network (RNN), which has been demonstrated in numerous research to have a good ability in time-series forecasting and to manage issues in long-term dependency models [20], is the predecessor of LSTM, a deep learning technique. There are input gates, output gates, forget gates, and memory blocks in a basic LSTM's recurrent layer. Memory blocks comprise gates that manage the information flow as well as memory cells that use self-connections to maintain the network's temporal state [21]. The input gate controls how many activations go into the memory cells. The output gate controls how cell activation output gets sent to the rest of the network. The forget gate scales the internal state of the cell before adding it as input to the cell via the self-recurrent link, enabling adaptive forgetting of the cell's memory data [22].

Sequence classification is used to check and identify data output by the LSTM in order to accurately verify all data. The proposed categorization using the LSTM sequence is shown in Figure 2 (below). The suggested method consolidates input from the third LSTM level into a final predictive value using a five-layer network made up of three LSTM layers and two dense layers. For layers 1, 2, 3, 4, and 5, correspondingly, there are 64, 128, 64, 64, and 32 units. The softmax function is typically used as the network's final activation function to define the classification output.

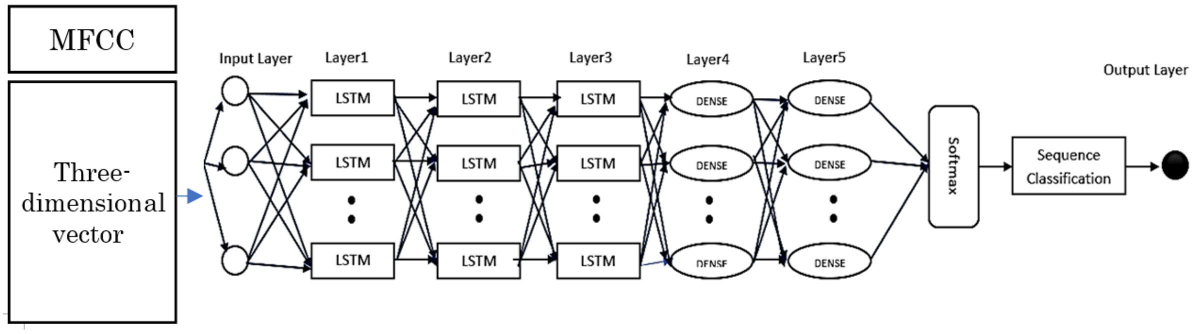


FIGURE 2. The proposed LSTM-sequence classification

2.5. **Model validation.** The dataset would be used in each trial for training and validation to the extents of 80% and 20%, respectively. Five trials were run after grouping the dataset into five different categories (folds). Each trial's testing set was chosen from one of the folds, while the training sets were chosen from the others. After that, the training and testing sets were used to practice the model and confirm its validity. A confusion matrix for nine classes was derived for each experiment's validation along with the overall accuracy. The 9-class confusion matrix is changed into a different matrix that contains true positive (TP), true negative (TN), false positive (FP), and false negative (FN), as shown in Table 1.

Accuracy (ACC), sensitivity (Se), and specificity (Sp) can be determined for each class using the estimated values for TP, TN, FP, and FN. The model's accuracy depends on how well it can identify instances. Sensitivity is the percentage of "actual" positives that are correctly classified as positives, whereas specificity is the percentage of "genuine" negatives that are accurately classified as negatives. The following is a representation of the accuracy, sensitivity, and specificity equations in terms of TP, TN, FP, and FN:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Se = \frac{TP}{TP + FN} \tag{2}$$

$$Sp = \frac{TN}{TN + FP} \tag{3}$$

The Matthews correlation coefficient (MCC), Fowkes-Mallows index (FM), and Bookmaker informedness (BM) can also be computed using the TP, TN, FP, and FN to show the statistical relevance of each class. MCC is a metric for comparing binary categorization that is seen and predictable. The similarity of observed and estimated binary classifications is compared using FM. For estimating the likelihood of making an educated decision, BM is used.

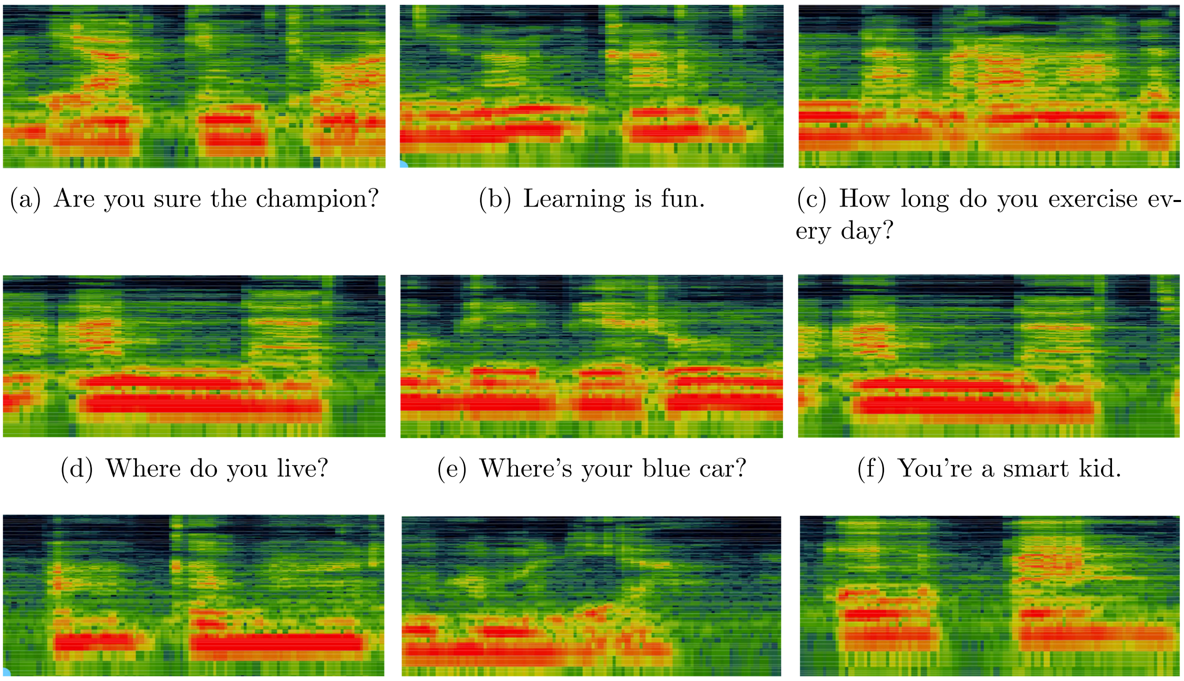
$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{4}$$

$$FM = \sqrt{\frac{TP}{TP+FP} \times \frac{TP}{TP+FN}} \tag{5}$$

$$BM = Se + Sp - 1 \tag{6}$$

3. Results and Discussion.

To justify this model, nine examples of visualization of Sentence Spectrograms are selected, as illustrated in the following Figure 3.



(a) Are you sure the champion?

(b) Learning is fun.

(c) How long do you exercise every day?

(d) Where do you live?

(e) Where's your blue car?

(f) You're a smart kid.

(g) What did you hurt yesterday?

(h) I was in college.

(i) I go to market.

Figure 3. The visualization of Sentence Spectrograms

In this study, 9-class confusion matrices were created for each of the five trials and then translated into TP, TN, FP, and FN matrices. Consequently, accuracy, sensitivity and specificity were computed. To appropriately analyze the results, an average of over five trials were collected for each letter for ACC, Se and Sp, as shown in Table 1. The model's per-class accuracy and sensitivity were calculated to be 100 percent, with the exception of the expressions *Are you sure the champion?*, *How long do you exercise every day?*,

Table 1. Average ACC, Se and Sp for 9-class

No	Class	Accuracy (%)	Sensitivity (%)	Specificity (%)
1	Are you sure the champion?	94	67	100
2	Learning is fun.	100	100	100
3	How long do you exercise every day?	93	100	93
4	Where do you live?	93	100	93
5	Where's your blue car?	94	50	100
6	You're a smart kid.	100	100	100
7	What did you hurt yesterday?	94	50	100
8	I was in college.	100	100	100
9	I go to market.	100	100	100

Where do you live?, *Where's your blue car?* and *What did you hurt yesterday?*, which amounted to just 50%, showing that the model has a high chance of correctly predicting and identifying a favorable result in each of the nine classes. As a result, the proportion of accurately identified instances would be higher. Except for the class for, *How long do you exercise every day?* sensitivity only reaches 100%.

In the numerical experiments, all of the algorithms acquire higher accuracy (100%) in the classes of *Learning is fun*, *You're a smart kid*, *I was in college* and *I go to market*. Except for the class words for *Are you sure the champion?*, *How long do you exercise every day?*, *Where do you live?*, *Where's your blue car?* and *What did you hurt yesterday?*, we only received 94, 93, 94 and 94 percent, respectively. As a result, the proposed method's average accuracy for 9-class is 96.4 percent. Furthermore, the proposed technique predicted the 9-class categorization with statistically good accuracy, a higher level of similarity between observed and expected binary classifications, and a higher likelihood of calculating an informed decision. The sensitivity results were obtained from nine more dominant classes achieving a score of 100, but there are still three classes with a score below 100, "*Are you sure the champion?*" which just got a score of 67, "*Where's your blue car?*" and "*What did you hurt yesterday?*" because the FN value of the confusion matrix in that class is not equal to zero. For classes, "*How long do you exercise every day?*" and "*Where do you live?*", it reaches a sensitivity value of 100, but its accuracy and specificity values do not reach 100 because FN values are equal to zero so sensitivity values can reach 100. This is due to the absence of predictive errors by a model that predicts negatively while the actual class is positive. Table 2 introduces the statistical significance of the proposed method. The class for *Learning is fun*, *You're a smart kid*, *I was in college* and *I go to market* achieved 100%; only *Are you sure the champion?*, *How long do you exercise every day?*, *Where do you live?*, *Where's your blue car?* and *What did you hurt yesterday?* achieved less than 100%. Consequently, we can conclude that the proposed method has better accuracy and good prediction in the numerical analysis.

Table 2. Statistical significance for 9-class

No	Class	MCC (%)	FM (%)	BM (%)
1	Are you sure the champion?	79	82	67
2	Learning is fun.	100	100	100
3	How long do you exercise every day?	68	71	93
4	Where do you live?	68	71	93
5	Where's your blue car?	68	71	50
6	You're a smart kid.	100	100	100
7	What did you hurt yesterday?	68	71	50
8	I was in college.	100	100	100
9	I go to market.	100	100	100

In contrast, after 100 training epochs, it was calculated that the model's accuracy was significantly lower than projected; as a result, 80 training sessions were selected. As shown in Figure 4, a graph of model accuracy over epochs was created to assess the effectiveness of selection. As can be observed, the accuracy of the model increases with the number of epochs, and between 100 and 300 epochs, the graph for the testing set gradually flattens. On the other hand, as seen in Figure 5, model loss reduced but persisted after falling drastically in the first 100 epochs. Just prior to the 300th epoch, the loss graph for the testing set flattens out.

Utilizing a jump motion controller and LSTM, the proposed approach aims to recognize nine sentences. The suggested method demonstrates increased precision and faster processing, suggesting somewhat superior overall performance. LSTM enhances the capacity of neural networks to manage extensive time-series datasets.

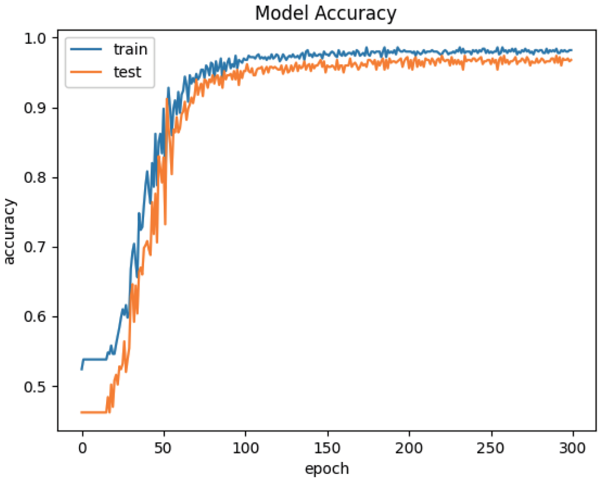


FIGURE 4. Model accuracy over epochs

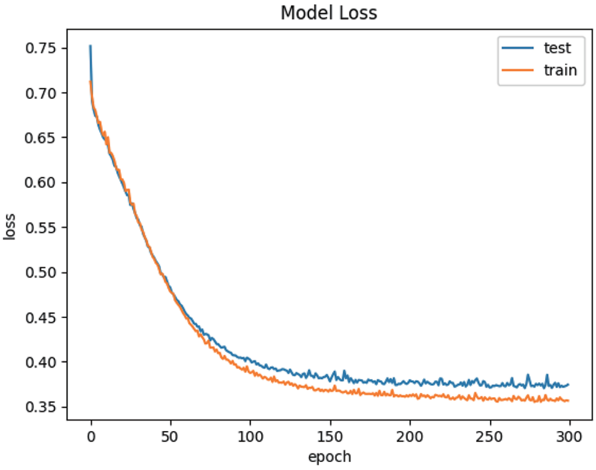


FIGURE 5. Model loss over epochs

4. **Conclusions.** The proposed research has developed an innovative model to detect dynamic speech in translation systems from voice to text. The spacetime properties of the speech sequence can be extracted using a combination of MFCC and LSTM, which is very useful for dynamic audio recognition. In terms of using them in real-time applications, integrating the classification and sequence models can reduce the search for classification of movement models to smaller ones, thus making model work easier and improving accuracy. Based on the results of the research, the accuracy of the proposed method reached 96.4 percent for ten types of speech tests. Therefore, the proposed method is suitable for identifying dynamic hand movements in the Indonesian language translation system from voice to text. For future research, it is suggested to develop new strategies that can be used to eliminate the transition of speech from each speech in a sentence to improve the accuracy of dynamic signal detection in the Indonesian language.

Acknowledgment. The authors would like to take this opportunity to thank the Ministry of Education, Culture Research and Technology of the Republic Indonesia, and Universitas Muhammadiyah Makassar for research facilities.

REFERENCES

[1] Z. Fang, H. Tanyas, T. Gorum, A. Dahal, Y. Wang and L. Lombardo, Speech-recognition in landslide predictive modelling: A case for a next generation early warning system, *Environmental Modelling & Software*, vol.170, 105833, DOI: 10.1016/J.ENVSOFT.2023.105833, 2023.

[2] J. Jeon, S. Lee and H. Choe, Beyond ChatGPT: A conceptual framework and systematic review of speech-recognition chatbots for language learning, *Computers & Education*, vol.206, 104898, DOI: 10.1016/J.COMPEDU.2023.104898, 2023.

[3] P. G. Shivakumar and S. Narayanan, End-to-end neural systems for automatic children speech recognition: An empirical study, *Computer Speech & Language*, vol.72, 101289, DOI: 10.1016/J.CSL.2021.101289, 2022.

[4] J. Zhao et al., Self-powered speech recognition system for deaf users, *Cell Reports Physical Science*, vol.3, no.12, 101168, DOI: 10.1016/J.XCRP.2022.101168, 2022.

[5] E. R. Swedia, A. B. Mutiara, M. Subali and Ernastuti, Deep learning Long-Short Term Memory (LSTM) for Indonesian speech digit recognition using LPC and MFCC feature, *Proc. of the 3rd International Conference on Informatics and Computing (ICIC 2018)*, DOI: 10.1109/IAC.2018.8780566, 2018.

[6] T. Wahyono, Y. Heryadi, H. Soeparno and B. S. Abbas, Enhanced LSTM multivariate time series forecasting for crop pest attack prediction, *ICIC Express Letters*, vol.14, no.10, pp.943-949, DOI: 10.24507/icicel.14.10.943, 2020.

[7] S. Ghaffarzadegan, H. Bořil and J. H. L. Hansen, Generative modeling of pseudo-whisper for robust whispered speech recognition, *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol.24, no.10, pp.1705-1720, DOI: 10.1109/TASLP.2016.2580944, 2016.

[8] X. Xie, H. Cai, C. Li, Y. Wu and F. Ding, A voice disease detection method based on MFCCs and shallow CNN, *Journal of Voice*, DOI: 10.1016/J.JVOICE.2023.09.024, 2023.

[9] Z. Zhang, C. Xu, J. Xie, Y. Zhang, P. Liu and Z. Liu, MFCC-LSTM framework for leak detection and leak size identification in gas-liquid two-phase flow pipelines based on acoustic emission, *Measurement*, vol.219, 113238, DOI: 10.1016/J.MEASUREMENT.2023.113238, 2023.

[10] M. Wielgosz, A. Skoczeń and M. Mertik, Using LSTM recurrent neural networks for monitoring the LHC superconducting magnets, *Nucl. Instrum. Methods Phys. Res. A*, vol.867, pp.40-50, DOI: 10.1016/j.nima.2017.06.020, 2017.

[11] M. Sharafi, M. Yazdchi, R. Rasti and F. Nasimi, A novel spatio-temporal convolutional neural framework for multimodal emotion recognition, *Biomedical Signal Processing and Control*, vol.78, 103970, DOI: 10.1016/J.BSPC.2022.103970, 2022.

[12] S. Jothimani and K. Premalatha, MFF-SAug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network, *Chaos Solitons Fractals*, vol.162, 112512, DOI: 10.1016/J.CHAOS.2022.112512, 2022.

[13] Md. Z. Islam, Md. M. Islam and A. Asraf, A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images, *Informatics in Medicine Unlocked*, vol.20, 100412, DOI: https://doi.org/10.1016/j.imu.2020.100412, 2020.

[14] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi and A. Othmani, MFCC-based recurrent neural network for automatic clinical depression recognition and assessment from speech, *Biomedical Signal Processing and Control*, vol.71, 103107, DOI: 10.1016/J.BSPC.2021.103107, 2022.

[15] B. Subramanian, B. Olimov, S. M. Naik, S. Kim, K. H. Park and J. Kim, An integrated mediapipe-optimized GRU model for Indian sign language recognition, *Scientific Reports*, vol.12, no.1, DOI: 10.1038/s41598-022-15998-7, 2022.

[16] G. H. Samaan et al., MediaPipe's landmarks with RNN for dynamic sign language recognition, *Electronics (Switzerland)*, vol.11, no.19, DOI: 10.3390/electronics11193228, 2022.

[17] D. Efanov, P. Aleksandrov and N. Karapetyants, The BiLSTM-based synthesized speech recognition, *Procedia Computer Science*, vol.213, pp.415-421, DOI: 10.1016/j.procs.2022.11.086, 2022.

[18] Y. Shi, X. Liu and C. Wei, An event recognition method based on MFCC, superposition algorithm and deep learning for buried distributed optical fiber sensors, *Optics Communications*, vol.522, 128647, DOI: 10.1016/J.OPTCOM.2022.128647, 2022.

[19] R. Ridwang, I. Nurtanio, A. A. Ilham and S. Syafaruddin, Deaf sign language translation system with pose and hand gesture detection under LSTM-sequence classification model, *ICIC Express Letters*, vol.17, no.7, pp.809-816, DOI: 10.24507/icicel.17.07.809, 2023.

[20] C. Millar, N. Siddique and E. Kerr, LSTM network classification of dexterous individual finger movements, *Journal of Advanced Computational Intelligence and Intelligent Informatics*, vol.26, no.2, pp.113-124, DOI: 10.20965/jaciii.2022.p0113, 2022.

[21] Q. Zhang, J. Zhang, J. Zou and S. Fan, A novel fault diagnosis method based on stacked LSTM, *IFAC-PapersOnLine*, vol.53, no.2, pp.790-795, DOI: 10.1016/j.ifacol.2020.12.832, 2020.

[22] A. Mittal, P. Kumar, P. P. Roy, R. Balasubramanian and B. B. Chaudhuri, A modified LSTM model for continuous sign language recognition using leap motion, *IEEE Sensors Journal*, vol.19, no.16, pp.7056-7063, DOI: 10.1109/JSEN.2019.2909837, 2019.