Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (192-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49040

JURNAL TEKNIK INFORMATIKA
Homepage : http://journal.uinjkt.ac.id/index.php/ti

Impact of Wavelet Denoising on LSTM-Based Greeting Sentence Recognition Using the IndSpeech Teldialog SVCR Dataset

Shabira Zhillan Zhalila[1], Luh Kesuma Wardhani[2*], Nenny Anggraini[3], Nashrul Hakiem[4], Imam Marzuki Shofi[5]

[1,2,3,4,5] Informatics Engineering, Faculty of Science and Technology
[1,2,3,4,5] State Islamic University Syarif Hidayatullah Jakarta
[1.2,3,4,5] Jl. Ir H. Juanda No.95, Ciputat, Kec. Ciputat Tim., Tangerang Selatan

Article:
Accepted: January 18, 2026
Revised: November 20, 2025
Issued: April 30, 2026

***Correspondence Address:**
luhkesuma@uinjkt.ac.id

ABSTRACT

Speech signals play a crucial role in human communication, particularly in speech recognition systems. However, speech recognition performance is often compromised by noise in the audio signal. This study aims to examine the effect of wavelet denoising technique on greeting sentence data containing artificial white noise before performing speech recognition using Long Short-Term Memory (LSTM). Mel Frequency Cepstral Coefficient (MFCC) is used as speech feature extraction. The results show that speech recognition accuracy reaches 90% on clean data. Accuracy drops to 51% when tested on data with noise, indicating a significant decrease of 39 percentage points. After applying the wavelet denoising method, accuracy improved using the two best parameter combinations. The combination with the highest SNR value resulted in an improvement of 18 percentage points, while the combination with the highest PESQ value resulted in an improvement of 13 percentage points. These findings indicate that the wavelet denoising method is capable of improving the performance of LSTM-based speech recognition in noisy environments

Keywords : *Denoising; Wavelet Transform; Long Short-Term Memory; Speech Recognition; MFCC* .

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (192-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49040

I. INTRODUCTION

Speech signal recognition is one of the research fields that is widely used in the development of human–machine interaction systems. This technology enables a system to characterize and identify speech signals automatically [1]. This technology has been widely applied in various electronic devices that interact using voice, such as audio and video transcription systems as well as security systems [2].

In the process of speech signal recognition, there is one problem that is often experienced by researchers, namely the presence of disturbances in speech signals which are often referred to as noise. Noise in a signal can be caused by many factors, such as being in a crowded place, bad weather, or other factors [3]. This noise can adversely affect the quality and clarity of speech pronunciation [3]. One type of noise that is often used in research, both for data testing and augmentation processes, is white noise.

To overcome the problems caused by noise disturbances, a technique called denoising can be used. Denoising is a technique used to separate the mixture between speech data and noise disturbances within it so that clean speech data is produced. The denoising technique is used to improve the quality of speech signals [4].

One denoising method that can be used to reduce or even eliminate noise disturbances in a speech signal is the wavelet denoising technique. Wavelet is a technique that is capable of representing signals in the time and frequency domains simultaneously [5]. Wavelet can be used to remove disturbances in speech data during the signal processing stage [6]. In several applications, wavelet is considered quite effective in mapping a speech signal into the time–frequency domain simultaneously without the risk of signal loss [1]. Therefore, wavelet has become one of the methods that is widely used in speech signal processing, both as a denoising technique and for speech feature extraction.

One algorithm that can be used in speech signal recognition systems is Long Short-Term Memory (LSTM). LSTM is one of the developments of Recurrent Neural Network (RNN) [7]. LSTM is used because it has the ability to handle sequential data over a long period of time and is capable of storing data over a relatively long duration [8].

This study uses a public dataset, namely the INDSpeech TELDIALOG SVCR Dataset, which contains small vocabulary continuous speech recognition data. This study focuses on three formal greeting sentences, namely “Halo, Selamat Pagi”, “Halo, Selamat Siang”, and “Halo, Selamat Malam”. This dataset has previously been used in research to develop an Indonesian speech recognition system for the deaf and mute using the Hidden Markov Model (HMM) with an average recognition rate of 98% [9].

Based on the discussion presented, it can be identified that the process of speech signal recognition still faces problems due to noise disturbances that degrade the quality and clarity of speech signals. One technique that can be used as a solution is the wavelet denoising technique. However, studies that specifically evaluate the effect of wavelet denoising techniques on the performance of models using Long Short-Term Memory are still limited, especially in the case of greeting sentence recognition. Therefore, this study is conducted to analyze the effect of wavelet denoising techniques on the performance of LSTM models in the case of greeting sentence recognition using the INDSpeech TELDIALOG SVCR dataset.

A. Related Work

Research on sentence recognition, wavelet denoising techniques, and speech signal recognition has been conducted in several previous studies.

Several studies have applied Long Short-Term Memory (LSTM) in audio-based classification tasks. For example, study [7] implemented LSTM to identify coughing and sneezing sounds, achieving an accuracy of 68.52%. Another study [10] also implemented LSTM for the classification of animal name pronunciations in Sundanese, achieving an accuracy of 97.50%. These results indicate that LSTM has strong capability in handling sequential audio data; however, these studies have not addressed the issue of noise interference in speech signals.

In addition, studies that apply wavelet techniques in signal processing have also been conducted. In study [11], wavelet and LSTM were used for automatic segmentation of pulse

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (192-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49040

signals, achieving an accuracy of 92.8%. Meanwhile, study [1] applied the wavelet-MFCC method for speaker classification using a Hidden Markov Model (HMM), achieving an accuracy of 95%. These studies demonstrate that wavelet is effective in signal processing, particularly as a feature extraction method; however, they have not specifically evaluated the impact of wavelet as a denoising technique on improving speech recognition performance in noisy conditions.

Furthermore, several studies have developed hybrid approaches to improve system performance. Study [12] proposed a hybrid model by incorporating a convolutional layer before a bidirectional LSTM, which successfully reduced the Word Error Rate compared to the standard model. However, this study focuses on model architecture development and does not consider noise reduction techniques in speech signals.

Based on previous studies, it can be observed that various methods have been used in speech signal processing, such as LSTM, Bidirectional LSTM, and Hidden Markov Model. In addition, feature extraction techniques such as MFCC and wavelet have been widely applied due to their ability to improve performance in speech signal processing tasks.

However, from these studies, the use of wavelet as a denoising technique prior to speech recognition has not been widely explored, particularly in combination with LSTM models for greeting sentence recognition. Most studies focus on model architecture development or feature combinations without specifically evaluating the impact of wavelet denoising on system performance.

Therefore, this study proposes the use of wavelet denoising as a preprocessing step combined with MFCC feature extraction and an LSTM model, with the aim of improving speech recognition performance under noisy conditions.

II. METHODS

Figure 1 is a research workflow used in this paper. In this paper, the mode development flow used is the machine learning life cycle.

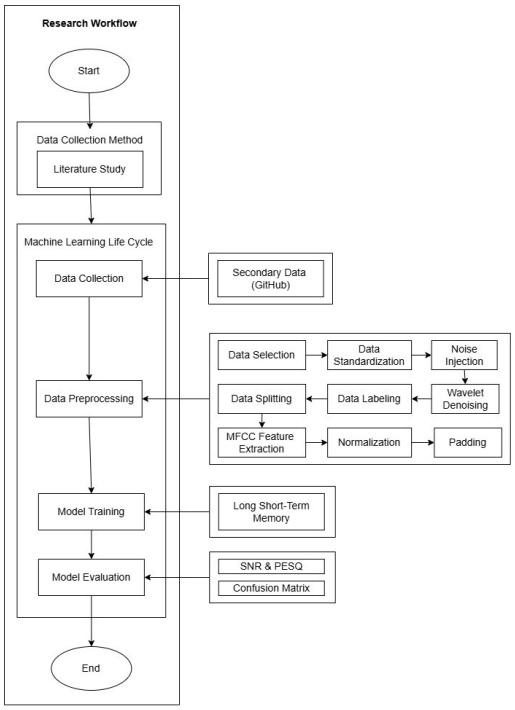

Figure 1. *Research Workflow*

B. Data Collection Method

In this study, a literature review was conducted to search for and collect references from various sources that could support the research. The references used were obtained from books, e-books, journals, articles, and websites with proven credibility.

C. Machine Learning Life Cycle

The model development method used in this study is the Machine Learning Life Cycle. The Machine Learning Life Cycle is a series of model development flows that ensure that each process is carried out properly and produces effective and accurate results.

Based on research [13], there are six stages in the Machine Learning Life Cycle, including Data Collection, Data Preprocessing, Model Training, Model Evaluation, Model Serving, and Prediction & Inference. Since this study only focuses on developing a system that can eliminate noise in a sound signal and then assess its effectiveness from the model evaluation results, the Model Serving and Prediction & Inference stages are not used in this study.

In this paper, each stage of the Machine Learning Life Cycle has its own process. The following section describes each stage carried out in this study.

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (192-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49040

1. Data Collection

In this first stage, researchers collected secondary data used in the study. This secondary data used a voice dataset obtained from Github named INDSpeech TELDIALOG SVCR. Please refer to Figure 2 below.

Figure 2. *INDSpeech TELDIALOG SVCR Dataset*

This dataset consists of 20,000 utterances from 70 dialogue vocabulary words comprising 100 sentences. This voice data was spoken by 200 speakers (100 women and 100 men) aged between 20 and 40 years old who speak a variety of spoken dialects from various ethnic groups.

However, this study specifically uses 3 greeting sentences, namely:

a. Halo, Selamat Pagi (Hello, Good Morning)
b. Halo, Selamat Siang (Hello, Good Afternoon)
c. Halo, Selamat Malam (Hello, Good Evening).

2. Data Preprocessing

a. Data Selection

The first stage in the Data Preprocessing flow is Data Selection. At this stage, data selection is carried out by identifying voice data containing three types of greeting sentences in the form of formal greetings from a total of 200 voice data folders. The total data used in this study is 600 data, with each formal greeting divided into 200 data.

The next stage is to change the naming format of all audio files to facilitate research. The file naming format is based on morning, afternoon, and evening greetings. Table 1 shows examples of the file names used.

Table 1. *Data Naming Format*

Sentence	File Name	Format Number
Halo, Selamat Pagi	Pagi_001	001 – 200
Halo, Selamat Siang	Siang_206	201 – 400
Halo, Selamat Malam	Malam_521	401 - 600

b. Data Standardization

The next stage is Data Standardization. This process is carried out in order to ensure that all audio data have consistent characteristics settings before proceeding to the next stage. The Data Standardization process includes the following steps:

- Converting stereo track to mono track
- Removing DC offset
- Applying a high-pass filter
- Resampling data to 16 kHz
- Adjusting the amplitude to -18 dB to maintain a consistent audio level across all samples while avoiding clipping and preserving signal quality during further processing.

c. Noise Injection

The next stage after the data has been standardized is to add artificial noise to the voice data. The artificial noise used in this study is white noise. The addition of noise intensity refers to previous research that applied artificial noise to evaluate the robustness of speech data. This study uses five levels of intensity, namely SNR 0, 5, 10, 15, and 20 dB. Based on the referenced study, two intensity levels were selected for further experimentation, namely SNR 10 dB and 20 dB [14].

d. Wavelet Denoising

After artificial noise is added to the data, the next stage is to perform denoising using wavelets. This denoising process is carried out to reduce the artificial noise in the audio data so that more representative features can be obtained during the feature extraction stage for use in the evaluation process

In this study, the two best parameters combinations identified by the researcher were used for testing with the dataset. Table 2 shows the two best parameters combinations used.

Table 2. *Best Parameters from the Research*

Type of Wavelet	Vanishing Moment	Threshold	Decomposition Level
Daubechies	7	Soft-Rigrsure	6
Coiflet	5	Soft-Rigrsure	-

These parameters are based on previous research. In study [15], it was found that coiflet 5 with a decomposition level of 6 and Soft-Rigrsure thresholding effectively eliminated white noise in PCG signals.

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (192-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49040

Meanwhile, in study [16], it was found that Daubechies 7 with Soft-Rigrsure thresholding could better reduce white noise in magnetic signals.

As a variation, the researchers also added another thresholding method, namely Soft-VisuShrink. Table 3 shows the parameters used in this study.

Table 3. *Research parameters applied*

Type of Wavelet	Vanishing Moment	Threshold	Decomposition Level
Daubechies	7	Soft-Rigrsure	4, 5, 6
Daubechies	7	Soft-VisuShrink	4, 5, 6
Coiflet	5	Soft-Rigrsure	4, 5, 6
Coiflet	5	Soft-VisuShrink	4, 5, 6

The parameters were then tested on the audio dataset to determine the best parameter combinations results for reducing white noise. The evaluation process used SNR (Signal to Noise Ratio) and PESQ (Perceptual Evaluation of Speech Quality) metrics.

After the experiment, two best parameter combinations were obtained based on the SNR and PESQ values.

The first combination, based on the highest SNR value, was coiflet 5 with decomposition level 6 with Soft-VisuShrink with an alpha value of 0.8.

Meanwhile, the second combination, based on the highest PESQ value, was coiflet 5 with decomposition level 6 with Soft-VisuShrink with an alpha value of 0.6.

e. Data Labeling

The next stage is to label the data into three classes based on the words morning, afternoon, and night. Refer to Table 4 below.

Table 4. *Data Labeling*

Label	Condition
0	File Halo, Selamat Pagi
1	File Halo, Selamat Siang
2	File Halo, Selamat Malam

Data labeling results in a balanced distributions data across all labels, with each label consisting of 200 samples. Figure 3 visualizes the distribution of the labeled data.

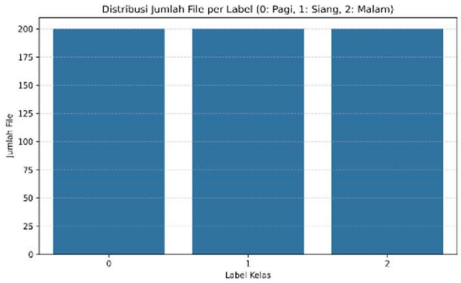

Figure 3. *Data Labeling Visualization*

f. Data Splitting

The next stage is to divide the data into three subsets: training data, validation data, and test data. This data splitting process uses the train_test_split function from the scikit-learn library.

The data is divided stratified based on the label column to maintain a balanced class distribution across the training, validation and test data. The data is split using a 70:15:15 ratio. The total number of samples data obtained is as follows:

- Training data: 420 data
- Validation data: 90 data
- Test data: 90 data

Figure 4 visualizes the distribution of data across the training, validation and test data.

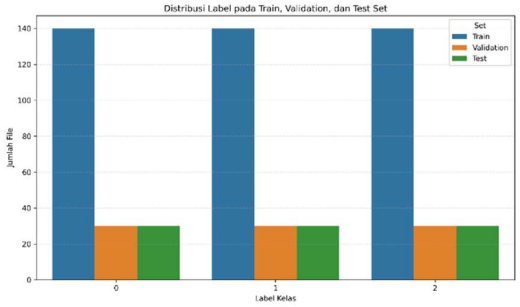

Figure 4. *Data Splitting Visualization*

g. Feature Extraction

The next stage is feature extraction using MFCC. However, before feature extraction is performed, the optimal padding length for MFCC features is determined, which is then used as input for the LSTM model. This process begins by calculating the number of frames resulting from MFCC feature extraction for each audio file.

Based on the frame count results, the shortest frame length was 93 and the longest was 137. After analysis, to avoid audio truncation and maintain input consistency for the model, zero post-padding was applied to all

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (192-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49040

audio files to reach a frame length of 138. This approach was chosen to balance data utilization and input length efficiency for the LSTM model.

After determining the padding length, the next process is MFCC Feature Extraction using the training, validation, and test datasets. The extracted MFCC features are then stored in .npy format to improve processing efficiency during model training.

h. Normalization and Padding

The final stage in the Data Preprocessing process is normalization and padding. The normalization process is used to standardize data with z-score normalization, which sets the mean to 0 and the standard deviation to 1. This is important to ensure that each feature has a uniform scale. The mean and standard deviation values are calculated from the training data, and stored for later use. The process continues with zero post-padding.

The padding results show that the training data has been successfully processed into a shape of (420, 138, 13). The validation and test data have also been processed into shapes of (90, 138, 13). All data has been successfully processed and stored in NumPy (.npy) format for use in the next stage.

3. Model Training

a. Data Preparation

The first stage in model training is to prepare the normalized and padded MFCC data. This data is then converted into a one-hot representation using the to_categorical function.

After the data has been successfully processed using one-hot encoding and the label outputs have been stored, the next step is to load the training, validation, and test datasets for model training. This process involves loading the previously generated one-hot encoded data. All arrays are converted to the float32 data type to improve memory efficiency during model training.

b. Modeling

The next stage is to train the model using the Long Short-Term Memory algorithm. The following step is to define the initial hyperparameter for experimentation. The parameters used include LSTM units, dropout rate, learning rate, optimizer, batch size, and epoch. For the first experiment, the default configuration is used to train the model with the parameter settings shown in Figure 5.

```
lstm_units    = 64
dropout_rate  = 0.2
learning_rate = 0.001
batch_size    = 64
epochs        = 50
optimizer = Adam(learning_rate=learning_rate)
```

Figure 5. *Default LSTM Modeling Code*

In the LSTM model training process, researchers applied callbacks to improve training efficiency and prevent overfitting. Three callbacks methods were used: Early Stopping, Model Checkpoint and ReduceLROnPlateau. Early Stopping was used to stop training when the validation accuracy did not improve for 5 epochs. Model Checkpoint was used to save the model with the best validation accuracy. ReduceLROnPlateau was used to reduce the learning rate when validation loss remained stagnant for 4 consecutive epochs. This combination ensures that training stops at the appropriate time and that the model is saved under optimal conditions.

After the model is compiled and trained, the training results are stored in Google Drive in .npy and .keras formats. In addition, the results of each experiment are visualized in the form of accuracy and loss graphs, which are used for further analysis.

In this hyperparameter tuning stage, the above process is repeated using different parameter combinations to obtain the most optimal model. In this study, 15 experiments were conducted using different parameter configurations. Table 5 shows the parameters used in the tuning process.

Table 5. *Model Parameter Tuning*

Parameter	Details
Layer LSTM	1 Layer, 2 Layer
Unit LSTM	32, 64 and 128
Dropout	0.2, 0.3 and 0.4
Learning Rate	0.001, 0.0001 and 0.0005
Optimizer	Adam and RMSProp
Batch Size	64
Epoch	50

4. Model Evaluation

At this stage, the researcher evaluate the model by selecting the best-performing model and testing it using the prepared test data. In this process, the test data used to analyze the effect of wavelet denoising on LSTM performance

consist of data with 20 dB white noise, to observe how noise degrades the quality of speech recognition, and denoised test data using the best parameter combinations to evaluate how wavelet denoising restores the quality of speech recognition affected by noise.

III. RESULTS AND DISCUSSION

A. Denoising using Wavelet

We performed a denoising process using wavelet methods to remove artificial noise added to the audio data, specifically white noise. The denoising process used parameters that had been tested in previous studies, including wavelet types (Daubechies 7 and Coiflet 5), the decomposition level (6), and the thresholding method (Soft-RigrSure).

For variation, two additional decomposition levels (4 and 5) and an additional thresholding method (VisuShrink) were included. The denoising results were evaluated using the Signal to Noise Ratio (SNR) and Perceptual Evaluation of Speech Quality (PESQ) metrics. Table 6 presents the sample experimental data used.

Table 6. *Experimental Data Sample*

Intensity	Mother Wavelet	Threshold	Level
10	Daubechies 7	Soft-RigrSure	4
20	Daubechies 7	Soft-RigrSure	5
10	Daubechies 7	Soft-RigrSure	6
20	Coiflet 5	Soft-VisuShrink	4
10	Coiflet 5	Soft-VisuShrink	5
20	Coiflet 5	Soft-VisuShrink	6
10	Coiflet 5	Soft-RigrSure	4
20	Coiflet 5	Soft-RigrSure	5
10	Coiflet 5	Soft-RigrSure	6

Based on the experiments conducted, the results show that the wavelet denoising process using coiflet 5 at level 6 with Soft-VisuShrink thresholding on white noise with an intensity of 20 dB produced relatively better compared to other experimental configurations (see Table 7).

Table 7. *Denoising Test Result*

Mother Wavelet	Level	Alpha Value	Final SNR	PESQ
Coiflet 5	6	0.6	16.42	2.971
Coiflet 5	6	0.8	17.27	2.571
Coiflet 5	6	1.0	17.22	2.290
Coiflet 5	5	0.6	13.97	2.906
Coiflet 5	5	0.8	14.36	2.534
Coiflet 5	5	1.0	14.38	2.277
Coiflet 5	4	0.6	11.53	2.825
Coiflet 5	4	0.8	11.54	2.523
Coiflet 5	4	1.0	11.45	2.325

Based on Table 7, it can be observed that the denoising process uses three alpha values, namely 0.6, 0.8, and 1.0. The evaluation process is then carried out using the SNR and PESQ metrics.

SNR (Signal-to-Noise Ratio) is a metric that measures the intensity of noise within a signal and is used to assess sound quality based on noise content. A higher SNR value indicates better signal quality. Based on Table 7, the SNR values vary depending on the combination of decomposition level and the alpha value. However, the best parameter combination achieved an SNR value of 17.27 dB, obtained using coiflet 5 with Soft-VisuShrink thresholding at level 6 with an alpha value of 0.8.

In addition to SNR, evaluation was also conducted using the PESQ value. PESQ is a metric used to evaluate sound signal quality based on objective assessments such as the human auditory system. The PESQ score ranges from 1 (very poor) to 4.5 (very good). Based on Table 7, the best combination achieved a PESQ score of 2.971, which indicates that the perceived speech quality is relatively good. This result was obtained using coiflet 5 with Soft-VisuShrink thresholding at level 6 with an alpha value of 0.6.

This indicates that different parameter settings may optimize objective noise reduction (SNR) and perceptual speech quality (PESQ) differently.

These two parameters combinations are then used in the subsequent data preparation stage for model evaluation using the Long Short-Term Memory algorithm.

B. Model Evaluation

The researchers trained the model using the Long Short-Term Memory algorithm. The data used consisted of the results of wavelet denoising with the two best parameter combinations. Clean speech data were used during the training process. During testing, three types of data were used: clean speech data, noisy speech data, and denoised speech data.

During the model training process, several hyperparameter tuning experiments were conducted to identify the best performing. Based on these experiments, the best model was identified as the lstm_v6 model with a training accuracy of 97%.

The best model consists of two LSTM layers with 64 units, a dropout rate of 0.3, a

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (192-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49040

learning rate of 0.0001, a batch size of 64, the Adam optimizer and was trained for 50 epochs.

Based on Table 8, it can be observed that testing the model on different types of data resulted in varying accuracy values. The following section presents an analysis of the results based on the test data.

Table 8. *LSTM Model Test Result*

Data Condition	Accuracy
Clean Data	90%
Noise Data	51%
Denoised Data Based on the Highest SNR	69%
Denoised Data Based on the Highest PESQ	64%

1. Clean Data Test Result

The results of testing on clean data showed an accuracy of 90%. This result indicates that the dataset used is relatively clean and suitable for training the model. Figure 6 presents the confusion matrix for testing on clean data.

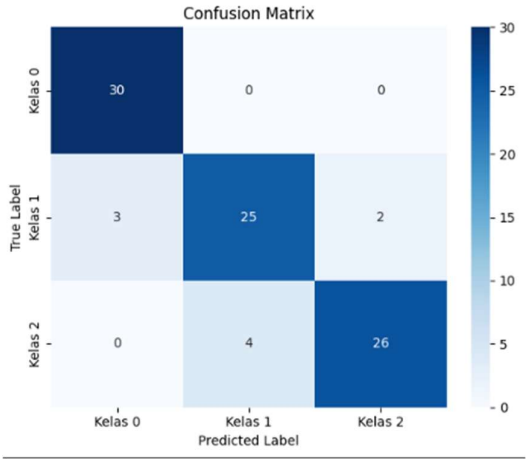

Figure 6. *Confusion Matrix for Clean Data*

Based on the confusion matrix, most samples are correctly classified into their respective classes, indicating that the model is able to learn distinguishing features of each greeting pattern effectively under noise-free conditions.

2. Noise Data Test Result

Based on Figure 7, it can be observed that speech data containing 20 dB white noise significantly affects the performance of the speech recognition system. This is reflected in the accuracy, which decreased to 51%, representing a substantial decline of 39 percentage points compared to clean data,

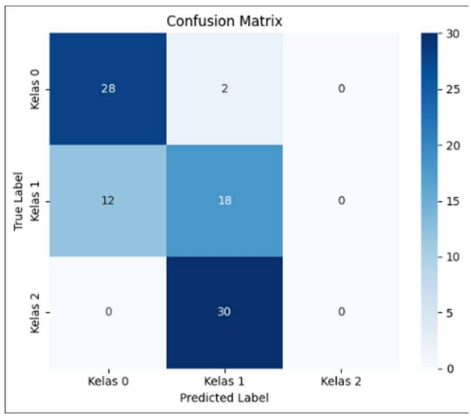

Figure 7. *Confusion Matrix for Noise Data*

In addition to accuracy, Figure 7 also shows that the model failed to identify all samples in class 2. Out of a total of 30 samples in class 2, none were predicted correctly. This indicates that the presence of noise severely disrupts the discriminative features of the speech signal, making it difficult for the model to distinguish between classes.

This performance degradation occurs due to the characteristic of white noise, which is uniformly distributed across the entire frequency spectrum of the signal. The addition of this noise directly interferes with the acoustic features of the speech signal utilized by the LSTM model during the classification process. Consequently, the overlapping feature representations between classes lead to misclassification, as evidence by the confusion matrix in Figure 7, where a considerable number of data samples were incorrectly assigned to other classes.

3. Denoise Data Test Result

Based on the results obtained, it can be seen that wavelet denoising successfully improved the quality of speech recognition compared to noisy data. Based on Figure 8, the highest SNR combination increases accuracy by 18 percentage points to 69%.

Jurnal Teknik Informatika Vol. 19 No. 1, April 2026 (192-202)
ISSN: p-ISSN 1979-9160 (Print)| e-ISSN 2549-7901 (Online)
DOI: https://doi.org/10.15408/jti.v19i1.49040

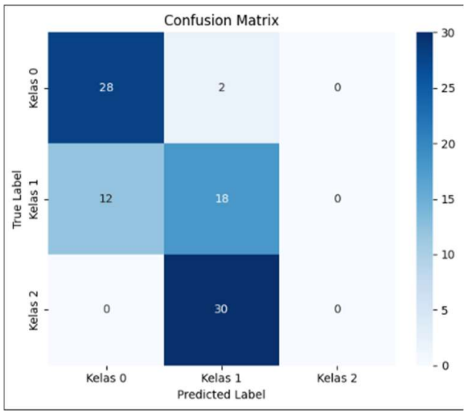

Figure 8. *Confusion Matrix for Denoise Data by the Highest SNR*

Meanwhile, in Figure 9, the highest PESQ combination increases accuracy by 13 percentage points to 64%.

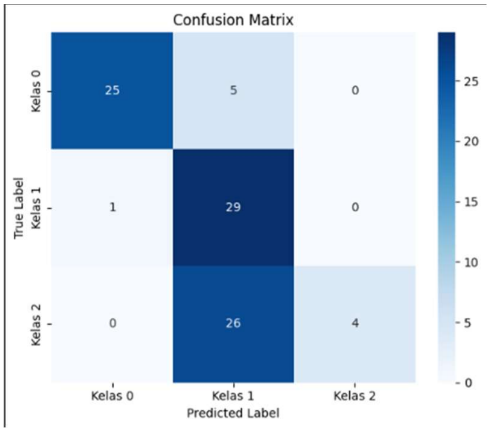

Figure 9. *Confusion Matrix for Denoise Data by the Highest PESQ*

Although both combinations improved the accuracy of speech recognition, there is a difference in the resulting confusion matrix. In the highest SNR combination, the model completely failed to correctly detect class 2. In contrast, the highest PESQ combination successfully predicted 4 samples of class 2 correctly, although this number remains considerably small. This indicates that a higher SNR metrics does not always correlate proportionally with the model ability to recognize a minority class.

This may also be caused by the characteristic of the thresholding method applied. In the highest SNR configuration, the thresholding process can be more aggressive in removing frequency components of the signal that are considered noise, causing the distinctive features of class 2 to be eliminated during the denoising process.

The higher accuracy improvement in the SNR-based combination (18 percentage points) compared to the PESQ-based one (13 percentage points) is primarily driven by the model's performance on Class 0 and Class 1. In the SNR-based configuration, which uses an alpha value of 0.8, the more aggressive noise suppression effectively clarifies the dominant acoustic features of "Pagi" and "Siang". This allows the LSTM model to achieve near-perfect recognition for Class 0 (28 correct) and Class 1 (18 correct), thereby boosting the overall accuracy score.

However, the confusion matrix analysis reveals a trade-off. While the SNR-based denoising excels at filtering background noise to highlight common features, it simultaneously treats the subtle, unique frequency components of Class 2 ("Malam") as noise. Consequently, all 30 samples of Class 2 were misclassified as Class 1. In contrast, the PESQ-based combination, with a lower alpha of 0.6, prioritizes perceptual speech quality and preserves more of the signal's original characteristics. Although this results in a lower overall accuracy because Class 0 and Class 1 are not as "cleanly" separated as in the SNR version, it retains enough information for the model to correctly identify some samples of Class 2. This suggests that the 5% accuracy gap between the two methods is a result of the SNR-based method's "over-cleansing" of the signal, which benefits majority-class recognition at the expense of class diversity.

CONCLUSION

This study aims to analyze the effect of wavelet denoising on the quality of speech recognition using the Long Short-Term Memory (LSTM) model. The results indicate that the presence of noise significantly degraded the model's performance, as reflected in an accuracy decrease of 39 percentage points from clean data (90%) to noisy data (51%).

The application of wavelet denoising using the Soft-VisuShrink method with the Coiflet 5 mother wavelet has been shown to improve speech recognition quality. The combination based on the highest SNR increases accuracy by 18 percentage points, while the combination based on the highest PESQ increases accuracy by 13 percentage points compared to the noisy data.

These findings suggest that the selection of denoising evaluation metric parameters affects the model's ability to recognize minority classes. The configuration with the highest SNR does not always produce better prediction distribution across all classes, as observed in the confusion matrix, where the highest PESQ combination demonstrated a better ability to detect a class with the limited number of samples.

In practical terms, these results can serve as a reference for developing speech recognition system that are more robust to noisy real-world environments, particularly in selecting appropriate denoising parameters prior to the classification process.

For future research, it is recommended to explore variations in noise types, more diverse datasets, and other deep learning algorithms to further strengthen the generalization of these findings.

REFERENCES

[1] S. Hidayat, A. S. Anas, S. Agrippina, A. Yusuf, and M. Tajuddin, "Sistem pengenalan pembicara dengan metode wavelet-MFCC dan pengklasifikasi hidden Markov models (HMM)," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 8, no. 1, pp. 119–126, 2021, doi: 10.25126/jtiik.202183284.

[2] O. Barkovska and A. Havrashenko, "Research of the impact of noise reduction methods on the quality of audio signal recovery," *Інформаційно-керуючі системи на залізничному транспорті*, vol. 29, no. 3, pp. 57–65, 2024, doi: 10.18664/ikszt.v29i3.313606.

[3] R. Dwi, P. Rahmasari, and N. Af, "Simulasi penghilangan noise pada sinyal suara menggunakan metode fast fourier transfrom (simulation of noise removal in sound signals by using fast fourier transform method)," vol. 1, no. 1, pp. 1–7, 2024, doi: 10.29303/semeton.v1i1.203.

[4] S. J. Lee and H. Y. Kwon, "A preprocessing strategy for denoising of speech data based on speech segment detection," *Appl. Sci.*, vol. 10, no. 20, pp. 1–24, 2020, doi: 10.3390/app10207385.

[5] K. Hulliyah, A. H. Setianingrum, and W. Santoso, "Sinyal elektroensefalografi untuk deteksi emosi saat mendengar stimulus pembacaan Al-Quran menggunakan wavelet transform," *Technomedia J.*, vol. 8, no. 2SP, pp. 175–188, 2023, doi: 10.33050/tmj.v8i2sp.2060.

[6] Yohannes and R. Wijaya, "Klasifikasi makna tangisan bayi menggunakan CNN," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, pp. 599–610, 2021, doi: 10.35957/jatisi.v8i2.470.

[7] S. B. Mulia, N. Wisma Nugraha, M. H. Robbani, T. Otomasi, M. & Mekatronika, and P. Manufaktur Bandung, "Implementasi machine learning untuk identifikasi orang batuk/bersin," *J. Energy Electr. Eng.*, vol. 81, no. 2, pp. 81–86, 2023, doi: 10.37058/jeee.v4i2.6836.

[8] P. Aliya Nabila, S. Soim, and A. Silvia Handayani, "Klasifikasi kondisi kendaraan berpotensi kecelakaan berbasis android menggunakan long short term memory," *J. Media Inform. Budidarma*, vol. 8, no. 1, pp. 30–40, 2024, doi: 10.30865/mib.v8i1.7005.

[9] S. Sakti, P. Hutagaol, A. A. Arman, and S. Nakamura, "Indonesian speech recognition for hearing and speaking impaired people," *8th Int. Conf. Spok. Lang. Process. ICSLP 2004*, pp. 1037–1040, 2004, doi: 10.21437/interspeech.2004-366.

[10] N. Aini Lailla Asri, R. Ibnu Adam, and B. Arif Dermawan, "Speech recognition untuk klasifikasi pengucapan nama hewan dalam bahasa sunda menggunakan metode long-short term memory," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 2, pp. 1242–1247, 2023, doi: 10.36040/jati.v7i2.6744.

[11] L. Huang, J. Yan, S. Cai, R. Guo, H. Yan, and Y. Wang, "Automated segmentation of the systolic and diastolic phases in wrist pulse signal using long short-term memory network," *Biomed Res. Int.*, vol. 2022, p. 9, 2022, doi: 10.1155/2022/2766321.

[12] S. N. Endah, R. Rismiyati, P. S. Sasongko, and A. P. F. Noiborhu, "Indonesian continuous speech recognition optimization with convolution bidirectional long short-

term memory architecture," *TELKOMNIKA (Telecommunication Comput. Electron. Control.*, vol. 23, no. 3, p. 807, 2025, doi: 10.12928/telkomnika.v23i3.24994.

[13] O. Spjuth, J. Frid, and A. Hellander, "The machine learning life cycle and the cloud: implications for drug discovery," *Expert Opin. Drug Discov.*, vol. 16, no. 9, pp. 1071–1079, 2021, doi: 10.1080/17460441.2021.1932812.

[14] S. Ranjan, R. Chakraborty, and S. K. Kopparapu, "Reinforcement learning based data augmentation for noise robust speech emotion recognition," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, pp. 1040–1044, 2024, doi: 10.21437/Interspeech.2024-921.

[15] S. K. Ghosh and R. N. Ponnalagu, "Investigation of discrete wavelet transform domain optimal parametric approach for denoising of phonocardiogram signal," *J. Mech. Med. Biol.*, vol. 22, no. 3, p. 19, 2022, doi: 10.1142/S0219519422500464.

[16] X. Li, K. Liao, G. He, and J. Zhao, "Research on improved wavelet threshold denoising method for non-contact force and magnetic signals," *Electron.*, vol. 12, no. 5, p. 1244, 2023, doi: 10.3390/electronics12051244.