Green Intelligent Systems and Applications

Research Article
Volume 5(1), 2025, 74–92
https://doi.org/10.53623/gisa.v5i1.605

Comparison Of Feature Extraction Techniques For Long Short-Term Memory Models In Indonesian Automatic Speech Recognition

Dimas Dwi Armaisya[1], Panca Dewi Pamungkasari[1], Achmad Pratama Rifai[2], Ira Diana Sholihati[1]*, Gopal Sakarkar[3]

[1]Universitas Nasional Faculty of Communication and Information Technology, Jakarta, Indonesia
[2]Department of Mechanical and Industrial Engineering, Universitas Gadjah Mada, Indonesia
[3]Dr.Vishwanath Karad MIT World Peace University, Pune, India

*Correspondence: ira.diana@civitas.unas.ac.id

SUBMITTED: 17 February 2025; REVISED: 26 March 2025; ACCEPTED: 7 April 2025

ABSTRACT: Automatic Speech Recognition (ASR) faced challenges in accuracy and noise robustness, particularly in Bahasa Indonesia. This research addressed the limitations of single feature extraction methods, such as Mel-Frequency Cepstral Coefficients (MFCC), which were sensitive to noise, and Relative Spectral Transform - Perceptual Linear Predictive (RASTA-PLP), which was less effective in frequency representation, by proposing a hybrid approach that combined both techniques using Long Short-Term Memory (LSTM) models. MFCC enhanced spectral accuracy, while RASTA-PLP improved noise robustness, resulting in a more adaptive and informative acoustic representation. The evaluation demonstrated that the hybrid method outperformed single and non-extraction approaches, achieving a Character Error Rate (CER) of 0.5245 on clean data and 0.8811 on noisy data, as well as a Word Error Rate (WER) of 0.9229 on clean data and 1.0015 on noisy data. Although the hybrid approach required longer training times and higher memory usage, it remained stable and effective in reducing transcription errors. These findings suggested that the hybrid method was an optimal solution for Indonesian speech recognition in various acoustic conditions.

KEYWORDS: ASR; LSTM; MFCC; RASTA-PLP; Hybrid

1. Introduction

ASR was a branch of deep learning that had been widely adopted [1]. ASR was the process of using algorithms in computing machines to modify, analyze, and recognize certain patterns in audio signals [2]. This technology enabled devices to recognize and transcribe human speech into text. ASR dealt with digital signal processing that was related to recognizing people based on their voice or speech [3]. In modern applications, ASR played a crucial role in enhancing user interactions, enabling hands-free control, and improving accessibility for individuals with disabilities. Its integration into virtual assistants, customer service automation, and real-time transcription services highlighted its growing significance in various industries.

One of the important stages in developing an accurate ASR system was the feature extraction process, which involved retrieving important information from audio signals.

Reference [4] discussed the most commonly used feature extraction techniques, such as Linear Predictive Coding (LPC), Mel-Frequency Cepstral Coefficients (MFCC), zero-crossing with peak amplitude (ZCPA), Discrete Wavelet Transforms (DWT), and RASTA. In this research, the ASR model was built using LSTM as the algorithm to handle sequential data such as speech signals. LSTM, as a type of Recurrent Neural Network (RNN), was effective in processing long-duration speech signals because it retained important information from long time sequences and overcame the vanishing gradient problem [5].

In addition, this study compared one non-feature extraction approach and three feature extraction techniques used in building ASR models: non-feature extraction, MFCC, RASTA-PLP, and a combination of both (a hybrid approach). MFCC was known to be effective in capturing the frequency characteristics of sound and was commonly used in the field of speech recognition due to its high recognition accuracy, good discriminative ability, and low coefficient correlation, making it excellent for identifying sound frequencies [6]. However, MFCC had a major drawback—its poor robustness to noise. Even minor deviations in frequency bands could significantly alter MFCC coefficients, affecting recognition accuracy. RASTA-PLP, on the other hand, combined the RASTA technique with the PLP method to increase the robustness of PLP features. It was known to be effective in reducing the influence of noise and environmental variations that often occurred when audio recordings were made in non-ideal conditions, making RASTA-PLP superior in handling noise [6, 7].

In this research, a hybrid approach was applied to feature extraction by combining the advantages of MFCC and RASTA-PLP. MFCC offered lower complexity, high recognition accuracy, and strong performance in identifying voice frequencies, while RASTA-PLP was effective in reducing noise and helped mitigate temporal distortion caused by the recording environment. Thus, the hybrid feature combination (MFCC & RASTA-PLP) was expected to improve ASR performance under noisy environmental conditions, despite potentially increasing the complexity of the feature extraction process. Therefore, this study aimed to test whether the hybrid approach could provide significant improvements in the context of Indonesian language speech recognition.

2. Materials and Methods

This section described the key components used to develop an ASR system for Indonesian speech recognition. The system applied the LSTM technique to transcribe speech into text accurately. The components used in this research included feature extraction using MFCC and RASTA-PLP, and the performance of the ASR model was evaluated using CER and WER.

2.1. Speech recognition.

Speech Recognition, also known as speech-to-text, was a technology that enabled the identification of spoken words and converted acoustic signals into written text. A speech recognition system usually consisted of modules such as acoustic-related models, language-related models, decoders, and acoustic feature extraction processing modules [8]. The working principle of the speech recognition system was to collect the characteristic information of the speech model, use training or other methods to construct an acoustic model, adapt it to the speech model, and use scientific algorithms to construct an acoustic information model. ASR was defined as the process of translating and transcribing spoken language using acoustic input

and algorithms [9]. ASR technology had been successfully integrated into computer-aided interpretation tools with high precision and low latency, further expanding its scope of application [10]. The results of this identification process could be displayed in written form or used by technological devices as commands to perform various tasks [11]. ASR continued to evolve with a wide range of practical applications, from virtual assistants to communication aids for people with disabilities, and contributed significantly to human-machine interaction..

2.2. Mel-frequency cepstral coefficients.

Mel-Frequency Cepstral Coefficients (MFCC) was one of the most widely used feature extraction methods in the field of ASR [7]. It represented sound signals in the frequency domain based on human auditory perception. The main purpose of the MFCC feature extraction method was to mimic the human ear [6, 12], so MFCC was designed to capture how humans heard at various frequencies more accurately than other methods such as LPC, PLP, and others. Figure 1 illustrated the step-by-step process of MFCC feature extraction.

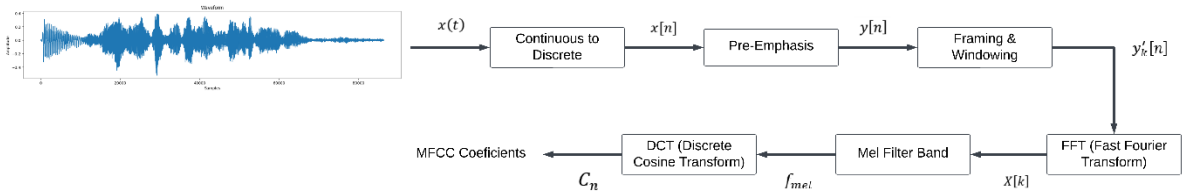

Figure 1. Flow of MFCC feature extraction.

Definition of each stage in the MFCC feature extraction process: Pre-emphasis: Enhancing high-frequency components to improve signal quality; Framing: Dividing the signal into smaller time frames for analysis; Windowing: Applying a Hamming window to minimize spectral leakage; Fast Fourier Transform (FFT): Converting the time-domain signal into a frequency-domain representation; Mel Filter Bank: Mapping the frequency spectrum onto a Mel scale, which better represents human auditory perception; Discrete Cosine Transform (DCT): Reducing data dimensionality to obtain the final MFCC coefficients.

The discretization process (Continuous to Discrete) converts a continuous-time audio signal x(t) into a discrete-time signal x[n] through sampling at specific time intervals. This discretization was performed according to the sampling rate, which determined how often the continuous signal was sampled to form a discrete signal. This discrete signal x[n] was then used as input for the pre-emphasis stage, where the high-frequency components were amplified to reduce the damping effect and improve the quality of features in the sound signal processing [13]. Equations (1) and (2) represented the processes involved in this discretization.

$$x(t) = A\,cos(w_0 t + \emptyset) \tag{1}$$

The sampling process is performed, resulting in x[n]

$$x[n] = A\,cos(\Omega_0 n + \emptyset) \tag{2}$$

The results were included in the pre-emphasis stage of both MFCC and RASTA-PLP. At this stage, the sound signal passed through a filter that strengthened the high-frequency components to compensate for the signal at lower sound frequencies. This process produced y[n] which emphasized the high-frequency components. The pre-emphasis calculation resulted in the signal y[n]. Once this value was obtained, y[n] was segmented into several frames or

smaller time windows during the framing stage. The calculation performed during the pre-emphasis stage is shown in Equation (3).

$$y[n] = x[n] - \alpha \cdot x[n-1] \tag{3}$$

Where x[n] is audio signal, α is pre-emphasis coefficient (in this research we used 0.97), y[n] is the result of pre-emphasis calculation, and n is index or position of the sample in the signal.

After the pre-emphasis calculation, the resulting signal was used in the framing stage, where it was divided into smaller segments (frames) to ensure local stationarity. The equation that was applied to the signal after it had been segmented into frames is shown as follows:

$$y_k[n] = x[k\ .\ L + n] \tag{4}$$

Where x[n] is audio signal, K is number of frames, N is frame length, L is steps per frame, n is Index or position of the sample in the signal.

The purpose of the framing stage was to maintain the stability of the signal by preventing information loss when the signal was extended. As a result, the output of the framing stage consisted of signal segments that were processed individually in the subsequent stage. Each frame or segment that had been extracted was then passed through the windowing stage, where it was multiplied by a windowing function w[n]w[n]w[n] to reduce spectral leakage. In this research, the Hamming function was used as the windowing function. Windowing was necessary to reduce discontinuities at the edges of each frame, to prevent leakage effects during the Fourier Transform, and to attenuate the signal at the frame's edges while emphasizing the center. The equation used in the windowing stage is presented below:

$$w[n] = 0.54 - 0.46 \cdot cos\left(\frac{2\pi n}{N-1}\right) \tag{5}$$

The windowing result of the signal:

$$y'_k[n] = y_k[n] \cdot w[n] \tag{6}$$

Where $w[n]$ is the windowing function, $y_k[n]$ is the result of the framing, and N is frame length.

After the windowing process, the results from this stage were used as input for the FFT (Fast Fourier Transform) stage. The signal in each frame was transformed using the FFT to obtain a spectral representation of the signal, which depicted the intensity at various frequencies. The FFT process was applied to each frame of the windowing result, converting the time domain into the frequency domain. This produced x[k]x[k]x[k], which represented the frequency components of the signal. Equation (7) shows the calculation performed at the FFT stage, as follows:

$$X[k] = \sum_{n=0}^{N-1} y'_k[n] \cdot e^{-J2k\pi mn/N} \tag{7}$$

Where $X[k]$ is the windowing function, m is filter index Mel, n is MFCC coefficient index, and N is frame length.

Next, the process proceeds to the Mel Frequency Transform stage, where the frequency spectrum generated by the FFT is converted into a Mel scale, which is based on how the human ear responds to different frequencies. Since the human ear is more sensitive to low frequencies than high frequencies, the Mel scale emphasizes low frequencies and provides finer resolution at lower frequencies. The equation that can be written is as follows:

$$f_{mel} = 2595 . \log_{10}(1 + \frac{f}{700}) \quad (8)$$

Finally, at the last stage, the Discrete Cosine Transform (DCT) is performed to reduce the data dimensions and eliminate redundancy in the frequency data. The output from the Mel filter is used in the DCT stage, and as a result, the MFCC coefficients are obtained, which represent the key features of the audio signal in a more compact form. Equation (9) shows the calculation performed at the DCT stage, as follows:

$$C_n = \sum_{m=0}^{M} log(S_m) \cdot cos\left[n \cdot (m - 0.5) \cdot \frac{2}{M}\right] \quad (9)$$

2.3. Relative spectra–perceptual linear prediction.

RASTA-PLP is a feature extraction method introduced to improve robustness against noise and distortion in changing environments. This method is based on the premise that the temporal characteristics of the audio signal environment differ from the temporal characteristics of the audio signal itself. RASTA-PLP combines the RASTA technique with the PLP method to enhance the robustness of PLP features. By applying band-pass filtering to each frequency sub-band of the speech signal, this approach reduces the effects of channel mismatch between the training and testing environments, while also smoothing short-term sound variations and eliminating constant offsets in the speech channel [6]. This process involves modifying the spectral amplitude using a nonlinear compression transformation, followed by filtering the temporal trajectory of each transformed spectral component. The next step simulates the auditory power law, and finally, the model spectrum is calculated for all poles, as in the conventional PLP method. The flowchart below provides an overview of the steps in the RASTA-PLP method.

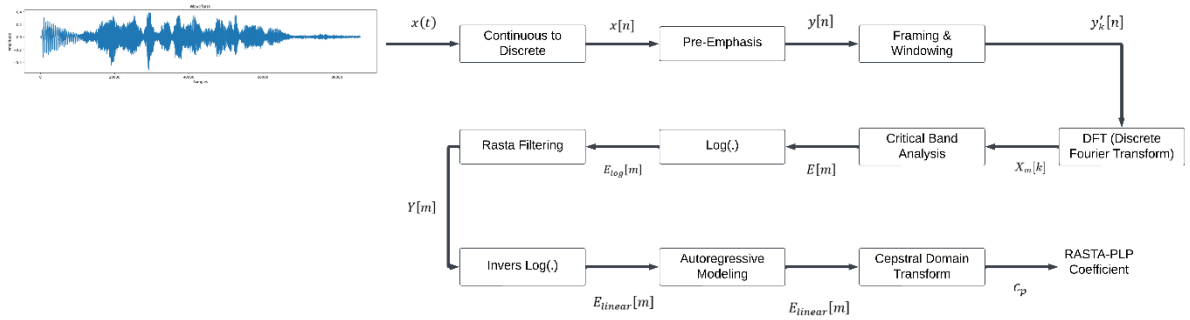

Figure 2. Flow of RASTA-PLP feature extraction.

The key stages in the RASTA-PLP process consist of: Pre-emphasis and framing: This step boosts high-frequency components and divides the signal into frames, similar to the MFCC process; Short-Time Fourier Transform (STFT): Converts time-domain signals into their corresponding frequency-domain representation; Critical Band Analysis: Uses a Bark-scale filter to simulate human auditory perception; RASTA Filtering: Reduces temporal fluctuations and eliminates noise artifacts; Cepstral Domain Conversion: Transforms the frequency-domain features into a more compact and efficient representation.

As shown in Figure 2, the initial stages of RASTA-PLP feature extraction are similar to those of the MFCC process, with the steps including pre-emphasis, framing, and windowing. These stages are crucial before entering the core process of RASTA-PLP feature extraction. After the windowing process is complete, the results are further processed using the Short-

Time Fourier Transform (STFT). At this stage, the STFT is applied to calculate the Discrete Fourier Transform (DFT) on each signal frame, which serves to transform the signal from the time domain to the frequency domain. This process produces a frequency spectrum that describes the distribution of energy across various frequencies for each specific time interval [13]. The following equation illustrates the DFT calculation applied to each signal frame using STFT.

$$X_m[k] = \sum_{n=0}^{N-1} y'_k[n] \cdot e^{-J\frac{2\pi kn}{N}} \quad (10)$$

Where $y'_k[n]$ is the result of signal windowing, N is frame length, m is filter index Mel, and k is index frequency.

Then, the calculation results from the DCT stage were used as input for the next stage, the Critical Band Analysis stage. This stage served to simulate human auditory perception by using a Bark filter band. This filter was used because it better reflected the way the human ear processed sound, focusing on the critical bands that underlie human hearing. The following is the energy calculation process performed at the Critical Band Analysis stage.

$$E[m] = \sum_{k=0}^{N/2} |X_m[k]|^2 \cdot H_m[k] \quad (11)$$

Where $H_m[k]$ is filter response for the M band, and $E[m]$ is the energy generated in the band.

In the next stage, the results of the Critical Band Analysis process were used to model the non-linear nature of human hearing in response to sound intensity. The spectral amplitude was then compressed logarithmically, with the logarithm helping to normalize the sound energy range. The equation that was calculated at this stage is as follows:

$$E_{log}[m] = \log(E[m]) \quad (12)$$

After the results were produced, the RASTA filtering stage was carried out to smooth out rapid changes, remove noise, and retain important information, ensuring that the result of this process produced filtered critical band energy. A band-pass filter was applied to the logarithmic signal to ensure that only relevant frequency information was retained. Equation (13) represents the calculation process carried out at the RASTA filtering stage as follows:

$$Y[m] = \frac{E_{log}[m]}{1+\frac{\Delta^2}{E_{log}[m]}} \quad (13)$$

After the RASTA filtering results were obtained, they were processed to adjust the energy at various frequencies in accordance with the sensitivity of human hearing to certain frequencies, typically around 1 kHz to 5 kHz. This process gave more weight to frequencies that are more sensitive to the human ear, resulting in a spectral representation that was closer to auditory perception. The results were then processed with the intensity loudness power law, where the signal intensity was mapped based on the non-linear relationship between physical intensity and perceived loudness. The process results were then processed in the inverse logarithm stage, which served to return the modified spectrum to a linear form. The following is the calculation equation for the inverse logarithm stage process.

$$E_{linear}[m] = \exp(Y[m]) \quad (14)$$

After the inverse logarithm stage, the process continued to autoregressive modeling. At this stage, an autoregressive model was used to represent the power spectrum at the output of the Bark filter, using the LPC technique. This stage served to reduce the spectral information to a small number of parameters that described the main frequency patterns. The signal was modeled as an autoregressive model for linear predictor coefficients, which were then further processed at the Cepstral Domain Transform stage. The following is the equation for the autoregressive modeling stage.

$$E_{linear}[m] = \sum_{p=1}^{P} a_p \, E_{linear}[m-p] \tag{15}$$

Where is a_p is autoregressive coefficient (Sum of LPC coefficients), P is order of the autoregressive model, p is autoregressive coefficient index.

This transformation was done to convert LPC coefficients to the cepstral domain, or from the frequency domain to the cepstral domain, resulting in cepstral coefficients or final features. The following is the calculation or formula for the cepstrum coefficient at the Cepstral Domain Transform stage to obtain the RASTA-PLP feature.

$$c_p = a_p + \sum_{k=1}^{p-1} \left(\frac{k}{p}\right) c_k a_{p-k} \tag{16}$$

Where c_p is the p-th cepstral coefficient, p is index of autoregressive or cepstral coefficients, and k : frequency index in the frequency domain.

The coefficient results obtained from the Cepstral Domain Transform stage were used as the final feature of RASTA-PLP, which included important information for speech recognition.

2.4. Evaluation with CER & WER.

WER and CER are performance metrics for automatic spontaneous speech recognition. WER measured the performance of predicting the correct order of recognized words, while CER was calculated based on the phoneme order. Both metrics were derived from the Levenshtein Distance formula and were useful for evaluating improvements to acoustic models. WER was used to measure the ratio of prediction errors at the word level. A good WER value was close to zero. The WER formula [14] was expressed in the following equation:

$$WER = \frac{S+I+D}{N} \tag{17}$$

The CER is a metric similar to WER but applied at the character level. CER measures the ratio of prediction errors at the character level and provides finer details compared to WER. The CER equation is as follows:

$$CER = \frac{S+I+D}{N} \tag{18}$$

With (S) as the number of substitutions, i.e., words/characters that were incorrectly recognized as other words; (I) as the number of insertions, i.e., words/characters that should not have been present but were recognized by the system; and (D) as the number of deletions, i.e., words/characters that should have been present but were not recognized by the system. The total number of words/characters in the correct reference transcription was denoted by (N).

Evaluation using WER and CER made it possible to comprehensively assess the performance of the ASR model and identify areas that required improvement in the model. These two metrics provided different but complementary perspectives when measuring the accuracy of an ASR system.

3. Results and Discussion

In this section, the results of the feature extraction preprocessing process, as well as the ASR model training analysis, will be discussed. The discussion begins by evaluating the results of feature extraction from the audio signal, which includes feature representation using MFCC, RASTA-PLP, and hybrid approaches. This process aimed to show the results of transforming the raw audio signal into a more informative form so that the model could better recognize sound patterns. Next, the training analysis of the ASR model was discussed to evaluate the performance of the model based on the extracted features. The results from this stage provided insights into the effectiveness of various feature extraction techniques in improving the accuracy and robustness of the model in handling speech signal variations.

3.1. Discussion of feature extraction preprocessing results.

The audio data was represented as a waveform in the time domain prior to preprocessing. Figure 3 shows the amplitude of an audio signal against time, known as a time-domain waveform representation. This preprocessed data contained information that only included the amplitude of the signal against time, without separating the energy contributions of the various frequencies. This representation was only useful for understanding the general pattern of the signal, but did not provide insight into the underlying frequency characteristics. For this reason, preprocessing was performed to help simplify the data by extracting key features from the raw audio signal, thus obtaining information that was more relevant to the speech recognition task and better equipped to be utilized by the ASR model.

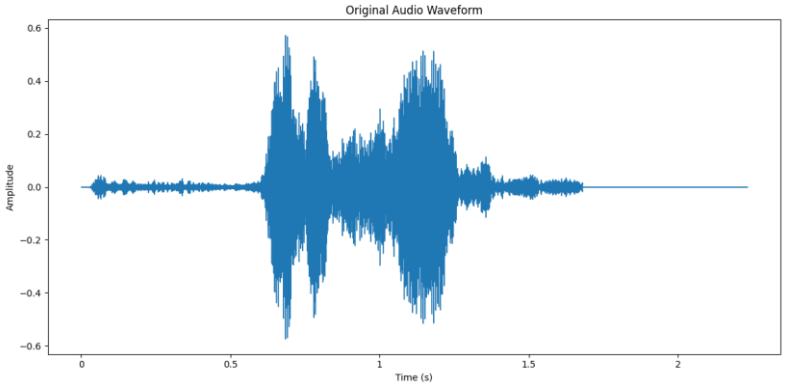

Figure 3. Data before preprocessing.

The results of feature extraction preprocessing using the MFCC technique were represented in Figure 4, which shows the visual representation of the MFCC. In the visualization, the horizontal axis represented time (in seconds), while the vertical axis displayed the MFCC coefficients. The colors in the figure represented the intensity or energy of each coefficient at a given time, with the color scale on the right indicating the range of values in decibels (dB). The 13 MFCC coefficients obtained per frame were extracted to be used as input to the LSTM-based ASR model. This MFCC visualization helped to understand

how audio features were processed and ensured that the model input contained an optimal representation of the sound.

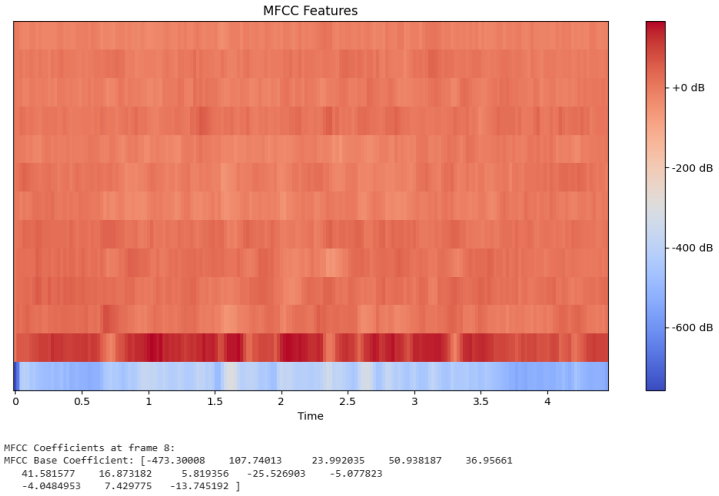

Figure 4. Visualization results of preprocessing on mfcc feature extraction.

The results of the feature extraction preprocessing using the RASTA-PLP technique were visualized in Figure 5, which represented the audio features resulting from processing the sound signal with the RASTA-PLP technique. In the graph above, the horizontal axis represented time (seconds), and the vertical axis represented the RASTA-PLP coefficients at different time points. The colors shown reflected the intensity values of the coefficients in decibels (dB), with the color scale on the right indicating their numerical values. Red indicated high intensity, while blue indicated low intensity. The coefficient results displayed the numerical values of the base coefficients for a particular frame (in this case, the 8th frame). The number of coefficients generated from each frame in the application of RASTA-PLP feature extraction was 13, which were later used as input data in the LSTM model to study the temporal patterns and long-term relationships between the coefficients.

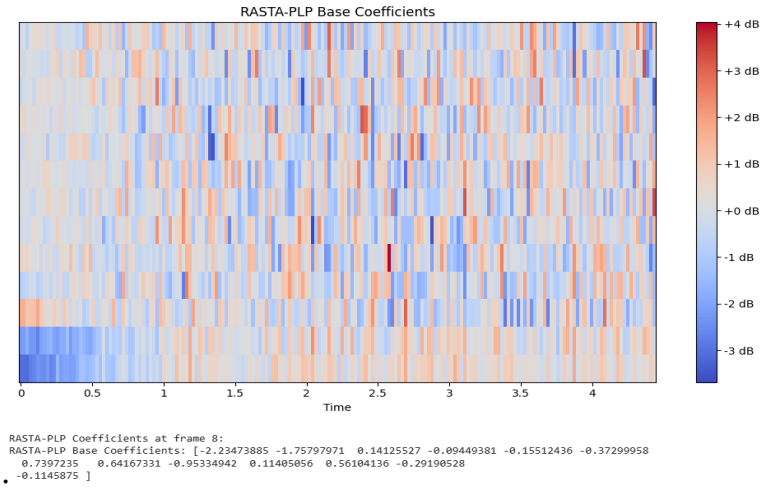

Figure 5. Visualization of preprocessing results on rasta-plp feature extraction.

Figure 6 illustrated the feature representation obtained from the hybrid approach, where the combined MFCC and RASTA-PLP features resulted in a more informative and robust spectral representation. The hybrid coefficients, which were visualized in the figure with feature dimensions on the y-axis and time frames on the x-axis, enhanced the feature space by capturing both fine spectral details from MFCC and noise-resistant temporal characteristics

from RASTA-PLP. Each row in the figure represented the combined value of both feature types, including detailed spectral information of the audio signal, with a total of 26 coefficients. These hybrid features were then used as numerical inputs for the LSTM model, where each time frame was processed to capture both short-term and long-term relationships in the sequential data, ultimately improving phoneme discrimination and robustness against environmental noise.

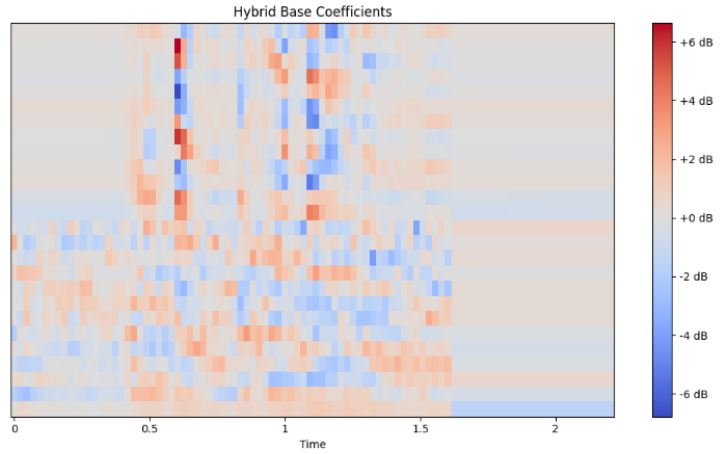

Figure 6. Visualization of preprocessing results on hybrid feature extraction.

3.2. ASR model training analysis results.

Table 1. Summary of noise sStatistics and data quality (in percent).

	Noise_percentage	Signal_clarity	Quality_score
Count	5041.00	5041.00	5041.00
Mean	0.09	99.91	99.82
Std	0.19	0.19	0.37
Min	0.00	98.76	97.51
25%	0.00	99.93	99.86
50%	0.00	100.00	100.00
75%	0.07	100.00	100.00
Max	1.39	100.00	100.00

ASR models with MFCC, RASTA-PLP, Hybrid, and Non-feature extraction techniques were trained using two different types of data: data with low noise and data with added Gaussian Noise. This approach aimed to produce an analysis of the ASR model's performance on clean sound (without interference) and sound affected by noise. The Signal-to-Noise Ratio (SNR) is the ratio between the strength of a useful signal and the strength of noise in a signal. The higher the SNR value, the clearer the signal and the lower the noise, indicating good audio quality [15]. For example, in the dataset, the first data set had an SNR of 38.43 dB, which indicated that the received signal was much stronger than the noise. The original data used were analyzed as a whole, resulting in the noise summary in Table 1. The overall average noise percentage was 0.09%, with the minimum noise in the dataset being 0.00% and the maximum noise being only 1.39%. This ensured that the dataset had clean and high-quality sound. The following are the results of the noise analysis summary performed on all data. The results of the training performed on the ASR models with the non-feature extraction approach and the MFCC, RASTA-PLP, and Hybrid feature extraction approaches showed variations in the CER,

WER, and loss levels for each method. The results obtained highlighted the performance of each model that was trained. The following are the results of ASR model training based on each feature extraction and non-extraction approach.

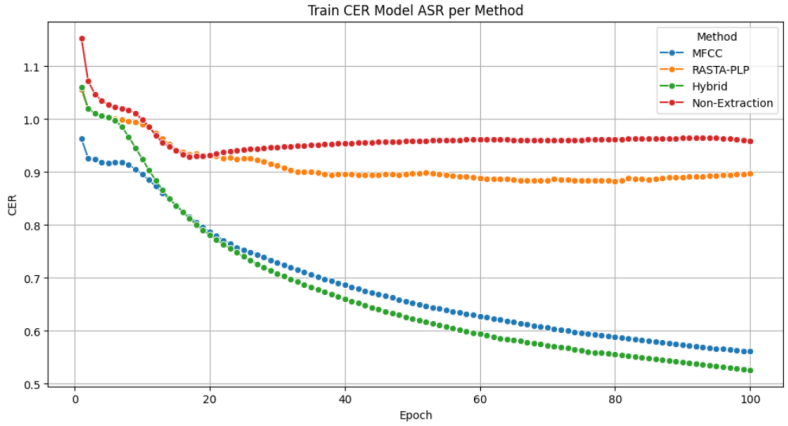

Figure 7. Graph of CER evaluation metrics in the ASR model.

Based on the CER graph in Figure 7, the performance of the four models—MFCC, RASTA-PLP, Hybrid, and non-feature extraction methods—during the training process, with 100 training epochs, is shown. From the graph, it can be seen that the non-feature extraction method had the highest CER value, maintaining a stable trend until the end of training. The model using the RASTA-PLP method also showed a relatively high error rate. This indicates that the models with the non-feature extraction and RASTA-PLP methods had difficulty correcting the prediction errors of the characters. ASR models using MFCC and the Hybrid method demonstrated a more significant decrease in CER as the epochs increased, showing an improvement in error correction. Notably, the Hybrid method achieved the lowest CER value, which indicates that the combination of MFCC and RASTA-PLP features helped the model recognize character patterns more effectively, leading to higher accuracy results.

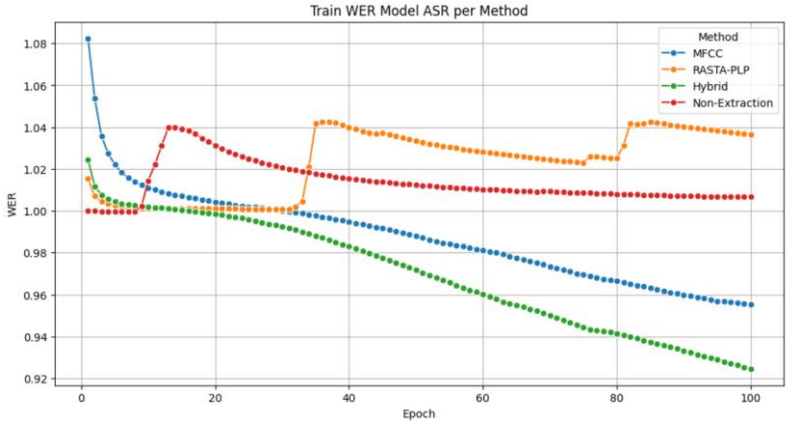

Figure 8. Graph of WER evaluation metrics in the ASR model.

In Figure 8, the WER shows the error rate in word recognition for the four models trained with the original data. The graph reveals that the ASR model with the non-feature extraction method had a relatively stable trend until the end of training. In contrast, the ASR model with the RASTA-PLP method exhibited instability and did not show a decrease in the word error rate, resulting in a higher WER compared to the other methods. Other methods, such as MFCC, showed a significant improvement in performance with a lower error rate compared to the non-feature extraction and RASTA-PLP methods. On the other hand, the Hybrid method displayed

a very significant decrease in WER compared to the other methods. This indicates that the use of combined MFCC and RASTA-PLP features made a positive contribution to improving word recognition accuracy, with fewer errors compared to other methods.

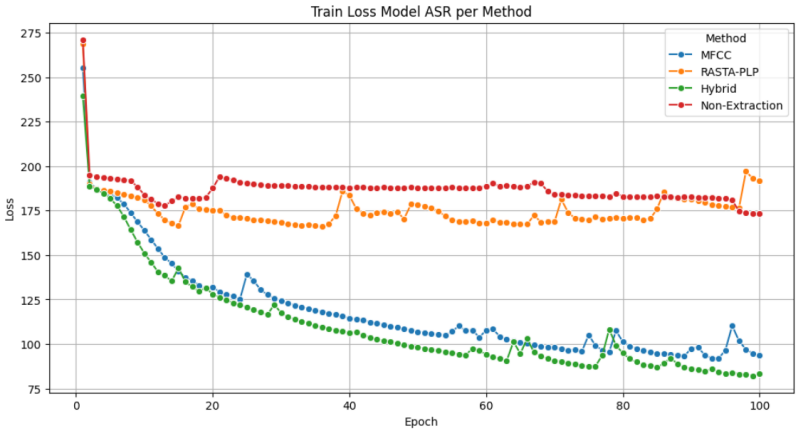

Figure 9. Graph of loss evaluation metrics in the ASR model.

The last graph, shown in Figure 9, is the loss graph, which illustrates how well the model minimizes the prediction error during the training process. In the trained models, the loss pattern decreases significantly and consistently in the ASR models with the MFCC and Hybrid methods throughout the training process. It is noted that the Hybrid method has the lowest loss value at the end of the epochs, which indicates the ability of the ASR model with the Hybrid method to reduce prediction errors more effectively. The pattern in the Hybrid method shows that the combination of the single feature extraction methods, MFCC and RASTA-PLP, had a positive impact on the learning process of the ASR model. On the other hand, it can be seen that the non-feature extraction method has a relatively high loss value with fairly stable results during the training process, indicating that the model was less able to learn effectively from the training data. The RASTA-PLP method also shows a very high loss with a fluctuating trend in the graph, resulting in a higher loss rate at the end of training compared to the other methods.

Based on research [16], Gaussian noise is added using two parameters: mean and variance. In this study, the mean parameter is set to 0, while the variance is set to 1, resulting in an SNR of 13.13 dB when added to the same sample data used in the original data. Gaussian noise applied to all data resulted in a noise summary of 6.83% for the average noise percentage, with a minimum noise percentage of 6.34% and a maximum noise percentage of 7.11%. The complete analysis of the noise summary for the dataset after adding Gaussian noise can be seen in Table 2. This analysis shows that the dataset with Gaussian noise has significant noise variability. The following is a summary of the noise analysis for all datasets after adding Gaussian noise.

Table 2. Summary of noise statistics and data quality with Gaussian noise (in percent).

	noise_percentage	signal_clarity	quality_score
Count	5041.00	5041.00	5041.00
Mean	6.83	93.17	86.33
Std	0.08	0.08	0.16
Min	6.34	92.89	85.77
25%	6.79	93.11	86.23
50%	6.84	93.16	86.32
75%	6.89	93.21	86.42
Max	7.11	93.66	87.32

The data that has been added with Gaussian noise is used when training the ASR model with the Non-extraction approach, as well as the MFCC, RASTA-PLP, and Hybrid extraction features. These methods show differences in CER, WER, and Loss levels for each approach. The following is a graph of the ASR model training process based on each feature extraction and non-feature extraction method.

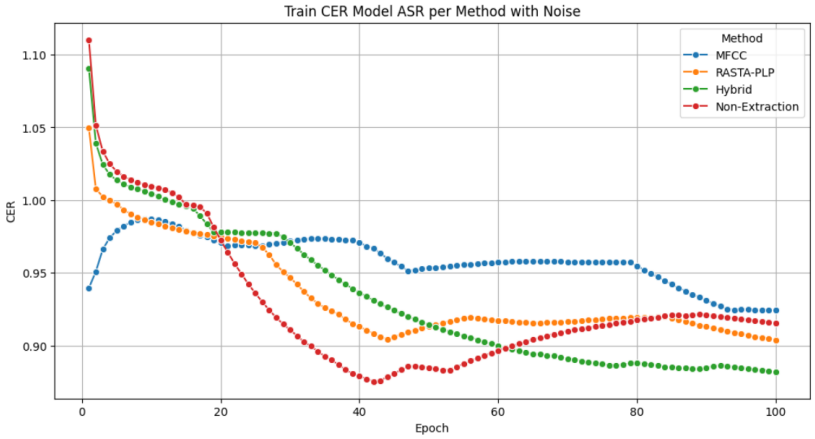

Figure 10. Graph of CER evaluation metrics with noise data in the ASR model.

Based on Figure 10, the graph shows the training performance of the four models: MFCC, RASTA-PLP, Hybrid, and the non-feature extraction methods in reducing CER during the training process with data added with Gaussian noise, using 100 epochs of training. From the graph, in general, the CER trend decreases as the epoch increases, but the non-feature extraction method experiences an increase in the CER rate after the 40th epoch. This suggests that the performance of the model with non-feature extraction has difficulty recognizing characters [17]. The ASR model with the Hybrid method shows a steady downward trend compared to the other methods, with a consistent decrease in CER rate, indicating that the Hybrid method is able to effectively recognize characters throughout the model training. The ASR model with the RASTA-PLP method has the second-lowest CER rate after the Hybrid method, indicating that the RASTA-PLP method performs relatively well at recognizing characters in noisy data, but not as well as the Hybrid method. On the other hand, the ASR model with the MFCC method has the highest CER error rate compared to the other methods, which indicates that this method is less effective in reducing character errors, especially in data with Gaussian noise.

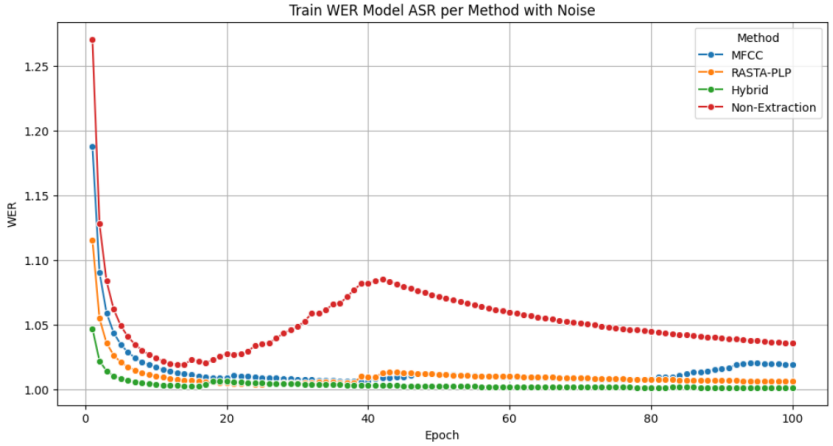

Figure 11. Graph of WER evaluation metrics with noise data in the ASR model.

The WER graph in Figure 11 shows the error rate in word recognition for the four models trained with data that has been added with Gaussian noise, illustrating the performance of the ASR model for each method [18]. Similar to CER, the WER trend also shows a decrease, although the Non-feature extraction method experiences an increase after the 20th epoch, making it the highest word recognition error rate compared to the other methods. The trend for the RASTA-PLP method is slightly similar to the Hybrid method, where after the 80th epoch, the RASTA-PLP method experiences a slight decrease, indicating that this method is fairly effective in reducing word recognition errors, although not better than the Hybrid method. The Hybrid method outperforms the other methods in word recognition, with the WER rate for the Hybrid method remaining quite stable throughout the training period, thus showing the best performance among all the methods trained.

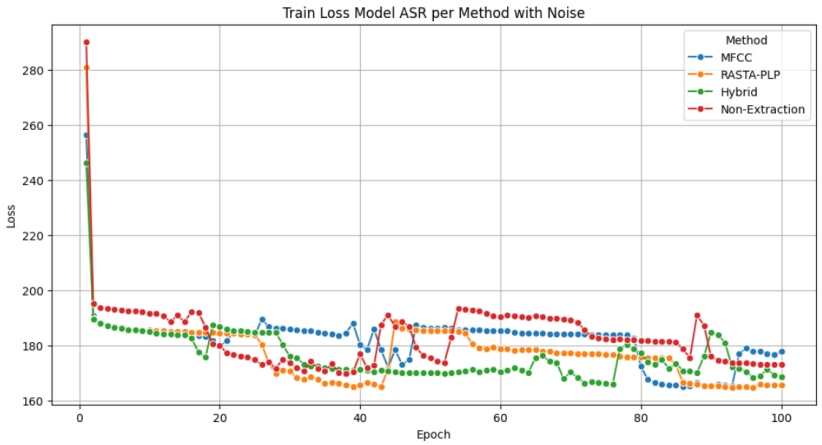

Figure 12. Graph of loss evaluation metrics with noise data in the ASR model.

In Figure 12, the loss graph illustrates how well the model minimizes the prediction error during the training process. At the beginning of training, all methods show high loss values, but these values decrease dramatically in the first few epochs. However, the trend of the loss metric is quite volatile across all methods, indicating instability in the training process. The Hybrid method performed quite well, with the second-lowest loss value after the RASTA-PLP method. The MFCC method has the highest loss value, which makes it the method with the highest loss compared to the other methods, indicating a lack of stability in the learning process [19]. The RASTA-PLP method shows a similar fluctuating trend as the other methods but has the lowest loss value compared to the others. This trend shows that although all methods experience fluctuations in loss value during training, the RASTA-PLP method still maintains the lowest loss value compared to the other methods.

Based on the ASR model training conducted on the Non-feature extraction, MFCC, RASTA-PLP, and Hybrid approach methods in two different data conditions—namely the original data without added noise and data with added Gaussian noise—the ASR model performance shows relatively similar results. On the original data without the addition of noise, the ASR model with the Hybrid method shows a significant and consistent pattern of decreasing error rates on each metric, such as CER, WER, and Loss, indicating an improvement in performance over the training process. Meanwhile, on data with the addition of Gaussian noise, the Hybrid method still shows the best performance, with more stable CER and WER reductions than other methods, although the loss value is not always the lowest. Overall, the Hybrid method remains the best choice for producing a more stable and accurate ASR model, both on the original data and data with added noise.

The ASR model, trained in both conditions, was then evaluated using validation data to show that the ASR model with the Hybrid method remains superior to the other methods. This evaluation involves several important metrics, namely Character Error Rate, Word Error Rate, and loss, which reflect the error rate in character and word prediction on validation data using original data without added noise and data that has been added with Gaussian noise [20]. Additionally, the training duration, trainable parameters, and memory usage were analyzed for each method as part of the model efficiency assessment. The results of the model performance evaluation with the validation data are shown in Table 3 and Table 4.

Table 3. Results of performance evaluation of the ASR model with data validation.

100 Epoch	Non Ekstraksi	MFCC	RASTA-PLP	Hybrid
CER	0.9586	0.5590	0.9002	0.5245
WER	1.0065	0.9538	1.0359	0.9229
Loss	182.4409	80.0591	280.1849	78.1123
Training Duration	7008 Seconds (1 Hour 56 Minutes)	11087 Seconds (3 Hour 4 Minutes)	10396 Seconds (2 Hour 53 Minutes)	10485 Seconds (2 Hour 54 Minutes)
Trainable Parameters	2.955.297	2.839.073	2.839.073	2.845.729
Memory Usage	13113.41 MB	7802.06 MB	7846.09 MB	8275.69 MB

Based on Table 3, regarding the performance evaluation results of the ASR model with validation data without added noise, it shows that the Hybrid method is superior to the other methods. The CER for the Hybrid method recorded the lowest value of 0.5245, confirming that this method is able to recognize characters with greater precision compared to both feature extraction and non-feature extraction methods. Likewise, in terms of WER, the Hybrid method recorded the lowest value of 0.9229, indicating that this method is more effective in recognizing whole words and producing more accurate transcriptions compared to other methods. In terms of loss, the Hybrid method also shows the best performance with a value of 78.1123, which indicates that the model can adapt to the original data (or data without noise) and is able to learn acoustic patterns better than the other methods.

On the other hand, in terms of training duration, the Hybrid method took 10,485 seconds (2 hours and 54 minutes), slightly longer than the single feature extraction of RASTA-PLP but more efficient than the single feature extraction of MFCC. In terms of memory usage, the Hybrid method consumed 8275.69 MB, which is higher than the single feature extraction methods. The relatively high memory consumption in the Hybrid method is due to the process of combining two feature extraction techniques, namely MFCC and RASTA-PLP, which results in a richer and more informative feature representation [20]. Despite requiring more time and memory, the superior performance in CER, WER, and loss makes the Hybrid method the best choice for modeling ASR in a noiseless scenario.

Table 4. Results of evaluating the performance of the ASR model with noise added data validation.

100 Epoch	Non Ekstraksi	MFCC	RASTA-PLP	Hybrid
CER	0.9149	0.9249	0.9033	0.8811
WER	1.0355	1.0191	1.0063	1.0015
Loss	178.3424	172.5249	165.2172	163.5806
Training Duration	6727 Seconds (1 Hour 52 Minutes)	10292 Seconds (2 Hour 51 Minutes)	6422 Seconds (1 Hour 47 Minutes)	16692 Seconds (4 Hour 48 Minutes)
Trainable Parameters	2.955.297	2.839.073	2.839.073	2.845.729
Memory Usage	12565.00 MB	8189.54 MB	8439.76 MB	9159.19 MB

Based on Table 4, regarding the performance evaluation results of the ASR model on Non-feature extraction, MFCC, RASTA-PLP, and Hybrid methods using validation data that has been added with Gaussian noise, it can be seen that the performance of the ASR model

using the Hybrid method results in the lowest CER of 0.8811, indicating that this method is more robust in handling noise at the character level. In WER, this method recorded the lowest value of 1.0015, showing that the Hybrid method is more effective at capturing word context compared to other methods. Additionally, the loss assessment for the Hybrid method recorded the lowest value of 163.5806, which indicates that the model is more stable and capable of learning patterns from data with Gaussian noise, resulting in better generalization on the validation data.

In terms of training duration and memory usage, the Hybrid method required the longest training time, approximately 16,692 seconds (4 hours and 48 minutes), which is significantly higher than the other methods. This is understandable because the Hybrid feature extraction process involves combining the MFCC and RASTA-PLP techniques, enriching the acoustic representation but increasing computational complexity. In terms of memory usage, the Hybrid method also consumed the most memory, with 9159.19 MB, indicating that although this method provides the best performance, computational cost and resource usage are challenges that need to be considered.

Based on the performance evaluation results of the ASR model on the original dataset and the dataset with Gaussian noise added, the Hybrid method has proven to produce the best performance under both conditions. This success comes from combining two feature extraction techniques, MFCC and RASTA-PLP. As indicated in research [7], each single extraction feature has its own advantages. The advantage of using MFCC extraction features is the spectral representation that aligns with human auditory perception, which provides good accuracy on noise-free data. On the other hand, RASTA-PLP is more resistant to noise and better at capturing temporal patterns, making it adaptive to changes in acoustic signals. By combining these two techniques, the Hybrid method creates a feature representation that leverages the strengths of each approach.

The most influential stages in the Hybrid method that contribute to a richer and deeper acoustic representation are the Mel Filter Bank stage of MFCC and the RASTA Filtering stage of RASTA-PLP. The Mel Filter Bank of MFCC helps scale the frequencies in the mel domain, which resembles human auditory perception of sound, making relevant acoustic information more prominent. Meanwhile, the RASTA Filtering stage of RASTA-PLP helps reduce low-frequency components that are considered noise in the acoustic environment and adds robustness to temporal fluctuations and noise.

The advantage of the Hybrid method lies in its ability to utilize the accurate frequency representation of MFCC and the noise suppression and adaptive temporal patterning of RASTA-PLP. This results in a more informative and rich acoustic representation [21]. The synergy of these two techniques allows the model to learn from more informative data, significantly improving the performance across CER, WER, and Loss metrics, both on the original data (without noise) and on data with Gaussian noise. This approach leads to a significant reduction in the error rate compared to both single extraction and non-feature extraction methods.

4. Conclusions

The hybrid approach effectively enhances feature representation, leading to improved ASR performance in both clean and noisy conditions. This supports the assertion that combining MFCC and RASTA-PLP provides a more robust and adaptive acoustic representation for

Indonesian speech recognition. The Hybrid method successfully combines the strengths of MFCC in capturing voice frequency characteristics with the noise resistance advantages of RASTA-PLP. The evaluation results show that the Hybrid method achieves lower CER and WER, along with a smaller loss, compared to other methods, both in clean data conditions and data with added Gaussian noise. Although Hybrid requires more time and memory during the training process, the results demonstrate improved stability and accuracy of the ASR model. The combination of both techniques allows the ASR model to utilize a richer and more informative spectral representation, improving reliability in speech recognition. Evaluation using CER, WER, and loss metrics confirms that the Hybrid method remains the best choice, despite challenges in computational efficiency. By increasing the number of epochs, it is expected that the Hybrid-based ASR model will further optimize the learning process, reduce the error rate, and improve generalization across a wider variety of sounds and environments. Additionally, increasing the number of epochs may enhance the model's stability, ensuring more consistent performance in recognizing speech characteristics under both clean and noisy conditions.

Acknowledgments

We would like to thank our colleagues and research assistants for their invaluable contributions to the support and resources that made this research possible.

Competing Interest

The authors declare that they have no competing interests.

Authors Contributions

Dimas Dwi Armaisya: Conceptualization, Methodology, Data Collection, Writing – Original Draft Preparation, Visualization, Data Analysis; Panca Dewi Pamungkasari: Supervision, Writing – Review & Editing; Achmad Pratama Rifai: Supervision, Writing – Review & Editing.

References

[1] Deng, L.; Yu, D. (2013). The essence of knowledge deep learning methods and applications. *Foundations and Trends® in Signal Processing*, *7*(3-4), 197–387. https://doi.org/10.1561/2000000039.

[2] Maruf, M.R.; Faruque, M.O.; Mahmood, S.; Nelima, N.N.; Muhtasim, M.G.; Pervez, M.J.A. (2020). Effects of noise on RASTA-PLP and MFCC based Bangla ASR using CNN. 2020 IEEE Region 10 Symposium, TENSYMP 2020. https://doi.org/10.1109/TENSYMP50017.2020.9231034.

[3] Devi, K.J.; Devi, A.A.; Thongam, K. (2019).Automatic speaker recognition using MFCC and artificial neural network. *International Journal of Innovative Technology and Exploring Engineering*, *8*(10), 1160–1166. https://doi.org/10.35940/ijitee.a1010.1191s19.

[4] Kurzekar, P.K.; Deshmukh, R.R.; Waghmare, V.B.; Shrishrimal, P.P. (2014). A comparative study of feature extraction techniques for speech recognition system. *International Journal of Innovative Research in Science, Engineering and Technology*, *3*(12), 13703–13709. https://doi.org/10.15680/ijirset.2014.0312034.

[5] Oruh, J.; Viriri, S.; Adegun, A. (2022). Long short-term memory recurrent neural network for automatic speech recognition. *IEEE Access*, *10*, 11325–11334. https://doi.org/10.1109/ACCESS.2022.3159339.

[6] Anusuya, M.A.; Katti, S.K. (2011). Front end analysis of speech recognition: A review. *International Journal of Speech Technology*, *14*(4), 275–286. https://doi.org/10.1007/s10772-010-9088-7.

[7] Labied, M.; Belangour, A. (2021). Automatic speech recognition features extraction techniques: A multi-criteria comparison. *International Journal of Advanced Computer Science and Applications*, *12*(8), 40–48. https://doi.org/10.14569/IJACSA.2021.0120821.

[8] Leini, Z.; Xiaolei, S. (2021). Study on speech recognition method of artificial intelligence deep learning. *Journal of Physics: Conference Series*, *1754*(1), 1–5. https://doi.org/10.1088/1742-6596/1754/1/012183.

[9] Özkural, E. (2018). The foundations of deep learning with a path towards general intelligence. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10983*, 303–316. https://doi.org/10.1007/978-3-319-97676-1_16.

[10] Lei, N.; An, D.; Guo, Y.; Su, K.; Liu, S.; Luo, Z.; Yau, S.T.; Gu, X. (2020). A geometric understanding of deep learning. *Engineering*, *6*(1), 53–59. https://doi.org/10.1016/j.eng.2019.09.010.

[11] Broad, D.J. (1972). Basic directions in automatic speech recognition. *International Journal of Man-Machine Studies*, *4*(3), 251–268. https://doi.org/10.1016/S0020-7373(72)80026-9.

[12] Singh, N.A.; Khan, R.A.; Shree, R. (2012). MFCC and prosodic feature extraction techniques: A comparative study. *International Journal of Computer Applications*, *49*(19), 1–6. https://doi.org/10.5120/8529-2061.

[13] Raj, V.A.; Dhas, M.D.K. (2022). Analysis of audio signal using various transforms for enhanced audio processing. *International Journal of Health Sciences*, *6*(2), 8890–8897. https://doi.org/10.53730/ijhs.v6ns2.8890.

[14] Morris, A.C.; Maier, V.; Green, P. (2004). From WER and RIL to MER and WIL: Improved evaluation measures for connected speech recognition. *8th International Conference on Spoken Language Processing, ICSLP 2004*. https://doi.org/10.21437/interspeech.2004-668.

[15] Dhanalakshmi, M.; Priya, K.B.; Jyothi, G.; Madhuri, K.; Amulya, M.; Durga, J.B.; Jyothika, M. (2023). Signal-to-noise ratio (SNR): A cornerstone metric for quality and reliability in diverse applications. *International Journal of Research Publication and Reviews*, *4*(11), 356–359.

[16] Laxmi Narayana, M.; Kopparapu, S.K. (2009). Effect of noise-in-speech on MFCC parameters. Proceedings of the 9th WSEAS International Conference on Signal, Speech and Image Processing, SSIP ’09, 1–6. https://doi.org/10.1109/SSIP.2009.14.

[17] Lee, S.; Kim, J. (2019). A comparative analysis of feature extraction techniques in automatic speech recognition. *Journal of Computer Science and Technology*, *34*(2), 410–418. https://doi.org/10.1007/s11390-019-1910-1.

[18] Smith, J.; Brown, T.; Wang, H. (2020). Effect of Gaussian noise on speech recognition performance: A deep learning perspective. *Journal of Acoustic Modeling*, *45*(4), 222–231. https://doi.org/10.1016/j.jacmod.2020.04.010.

[19] Zhao, L.; Tan, X.; Liu, P. (2022). Performance analysis of MFCC in noisy speech recognition using deep learning models. *IEEE Access*, *10*, 112345–112355. https://doi.org/10.1109/ACCESS.2022.3176453.

[20] Ahmed, F.; Rahman, M.; Hasan, M. (2021). Hybrid feature extraction methods for robust speech recognition in noisy environments. *International Journal of Speech Technology*, *24*(3), 555–567. https://doi.org/10.1007/s10772-021-09813-2.

[21] Sharma, A.; Singh, R.; Jain, V. (2022). Performance enhancement of ASR using hybrid feature extraction combining MFCC and RASTA-PLP. *Journal of Signal Processing Systems, 94*(3), 281–295. https://doi.org/10.1007/s11265-021-01667-7.

© 2025 by the authors. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).