Computer Architecture and Signal Processing

E-ISSN: 3124-0003

Research Article

Phonics Recognition for Indonesian Dialects Using PNCC and RNN-GRU

Nayya Kamila Putri Yulianto[1*], Ratih Nur Esti Anggraini[2], Dwi Sunaryono[3]

1 Universitas Diponegoro, Indonesia; e-mail : nayya.kamila@gmail.com
2 Institut Teknologi Sepuluh Nopember, Indonesia; e-mail : ratih_nea@its.ac.id
3 Institut Teknologi Sepuluh Nopember, Indonesia; e-mail : dwi@its.ac.id
* Corresponding Author: nayya.kamila@gmail.com

Abstract: Phonics plays an important role in helping learners develop reading and spelling skills by linking sounds to letters. However, applying phonics in multilingual environments such as Indonesia can be challenging due to the influence of regional dialects, especially for non-native English speakers. In this study, a phonics recognition system is developed using several deep learning approaches. The dataset consists of 986 audio recordings collected from 38 speakers, including both native and non-native English speakers from different regions in Indonesia. To improve data diversity, augmentation techniques such as pitch shifting and speed perturbation are applied. Feature extraction is performed using MFCC and PNCC, followed by classification using CNN, RNN-GRU, and Transformer models. The results show that the RNN-GRU model with PNCC features achieves the best performance, with an accuracy of 94.59% and an F1 Score of 0.946. Compared to previous work using SVM and MFCC, this approach provides better results. It is also observed that PNCC is more robust in handling pronunciation variations, and that dialect differences have a noticeable impact on model performance. Overall, this study highlights the importance of considering dialect variation in phonics recognition and shows how deep learning can be used to build more adaptive speech-based learning systems.

Keywords: Deep Learning; Indonesian Dialect; MFCC; Phonics Recognition; PNCC.

1. Introduction

Reading proficiency is one of the key foundations of a child's academic development. In the context of language learning, particularly English as a foreign language, literacy remains a significant challenge in many countries, including Indonesia. One approach that has proven effective in improving reading skills is the phonics method, which emphasizes the relationship between letters and sounds. Through this method, learners can recognize sound patterns within words, thereby improving their reading, pronunciation, and overall comprehension [1], [2].

Although the phonics method has been widely adopted, its implementation in multilingual environments like Indonesia presents unique challenges. Indonesia has a high diversity of dialects, which affects the pronunciation of language sounds, particularly among non-native English speakers [3]. These variations in pronunciation lead to differences in phonics representation that can hinder the learning process as well as the development of sound-based automatic systems. Therefore, an approach is needed that can accommodate this diversity of pronunciation so that phonics recognition systems can operate more accurately and adaptively.

Previous research by Esti Anggraini has explored phonics recognition using conventional machine learning approaches, including Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Random Forest, with Mel-Frequency Cepstral Coefficients (MFCC) as the primary feature extraction method. The study reported promising performance, achieving an accuracy of up to 95% and an F1 score above 0.92 [4]. Despite these results, several limitations remain. First, the approach relies on conventional machine learning models, which may have limited capability in capturing complex temporal and spectral patterns in speech signals compared to deep learning methods [5]. Second, the dataset

Received: 01, January 2025
Revised: 28, February 2025
Accepted: 20, April 2025
Published: 30, June 2026
Curr. Ver.: 30, June 2026

Copyright: © 2026 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY SA) license (https://creativecommons.org/licenses/by-sa/4.0/)

used in the study is relatively limited in both size and diversity, particularly in terms of dialectal variation. Third, the study does not provide an in-depth analysis of how pronunciation differences across dialects affect model performance. These limitations highlight the need for a more comprehensive approach that not only improves model performance but also considers the impact of dialect variation in phonics recognition.

With advancements in technology, deep learning methods such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Transformers have demonstrated superior performance in various speech recognition tasks. CNNs are effective at capturing local patterns in audio signals, while RNNs can model temporal dependencies in sequential data [5], [6]. Furthermore, the Transformer architecture with its self-attention mechanism has proven capable of capturing complex relationships within audio signals and improving performance in Automatic Speech Recognition [7].

On the other hand, the choice of feature extraction methods is also a critical factor in determining system performance. MFCC, as one of the commonly used methods, has limitations when dealing with pronunciation variations and unstable signal conditions. Alternatives such as Power-Normalized Cepstral Coefficients (PNCC) offer a feature representation that is more robust against noise and signal variations, thereby potentially improving the performance of phonics recognition systems [8].

Based on this background, this study aims to develop a deep learning-based phonics recognition system capable of handling pronunciation variations caused by dialect differences. Unlike previous studies, this research not only focuses on improving model performance but also explicitly evaluates the impact of dialect variations on phonics recognition results. Additionally, this study examines the effectiveness of two feature extraction methods, namely MFCC and PNCC, on several deep learning architectures, including CNN, RNN-GRU, and Transformer.

The main contributions of this study are as follows. This study explores phonics recognition while taking into account dialectal variations in Indonesia, a topic that has not been extensively addressed in previous phonics recognition research. In addition, this study compares the performance of MFCC and PNCC feature extraction methods across various deep learning models and analyzes model performance on data from both native and non-native speakers, including variations across Indonesian dialects. The experimental results demonstrate that the combination of PNCC and RNN-GRU provides the best performance and is more stable in handling pronunciation variations.

2. Literature Review

Phonics Recognition Studies

Research on phonics recognition remains relatively limited, particularly in multilingual contexts involving non-native speakers. One important study in this field was conducted by Esti Anggraini, who investigated phoneme recognition using conventional machine learning approaches, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Random Forest with Mel-Frequency Cepstral Coefficients (MFCC) as feature extraction. This study utilized a dataset comprising native and non-native English speakers, with a total of 16 participants, including ten native speakers and six non-native speakers from regions such as Java and Batam. Each participant contributed audio recordings of phonics alphabet pronunciations, and the dataset was further enriched using data augmentation techniques such as noise addition and time shifting.

Several experimental scenarios were conducted to evaluate the impact of different configurations, including variations in data augmentation, feature extraction, and classification selection. The results show that the SVM model consistently achieved the best performance, with the highest accuracy reaching approximately 95% and strong performance on evaluation metrics such as F1-score, precision, and recall. Furthermore, this study reveals that factors such as data quality, pronunciation consistency, and sample size significantly influence model performance. Differences in pronunciation between native and non-native speakers also affect classification results, with data from native speakers generally yielding higher accuracy than data from non-native speakers.

However, despite achieving promising results, this study has several limitations. The size and dialect diversity of the dataset are relatively limited, as it includes only a small number of non-native speakers from a few regions. Furthermore, this approach relies solely on conventional machine learning models and does not explore deep learning techniques, which have proven to be more effective at capturing complex patterns in speech signals. Additionally, this study does not provide a comprehensive analysis of how dialectal variations affect phonics recognition performance [4].

Feature Extraction Techniques in Speech Processing

In Mel-Frequency Cepstral Coefficients (MFCC) are one of the most commonly used feature extraction methods in speech signal processing since their introduction by Davis and Mermelstein in the 1980s. MFCCs represent audio signals in the form of coefficients that reflect the spectral characteristics of sound by adjusting to human auditory perception of frequency [9], [10].

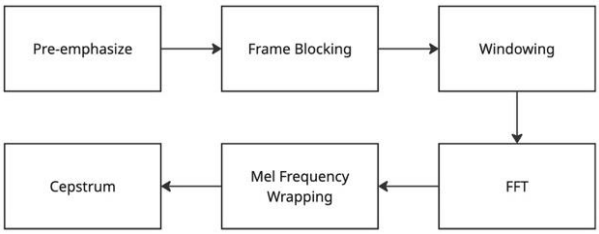


Figure 1. Steps of MFCC feature extraction.

The MFCC extraction process involves several main stages, namely pre-emphasis, frame blocking, windowing, Fast Fourier Transform (FFT), Mel filterbank, and Discrete Cosine Transform (DCT) which is pictured in Figure 1 [11]. These stages transform the audio signal from the time domain into a more informative frequency representation while mapping it into the Mel scale, which aligns with human auditory perception. In general, MFCC is widely used due to its ability to produce compact and informative features while preserving essential characteristics of the audio signal. However, this method has limitations when dealing with complex signal variations, such as noise and pronunciation differences, which can degrade the performance of speech recognition systems.

To address these limitations, alternative feature extraction methods have been developed, one of which is Power-Normalized Cepstral Coefficients (PNCC). PNCC was introduced by Kim and Stern as a more robust feature extraction technique for speech recognition, particularly in noisy environments and real-world conditions [8]. Unlike MFCC, which relies on a Mel filterbank and logarithmic nonlinearity, PNCC utilizes a gammatone filterbank and power-law nonlinearity, thereby providing a representation that is more robust against noise and signal variability.

The PNCC extraction process consists of several stages designed to improve robustness, as illustrated in Figure 2. Similar to MFCC, the process begins with pre-emphasis and transformation into the frequency domain using Short-Time Fourier Transform (STFT), followed by magnitude squared computation to represent signal energy. The signal is then processed using a gammatone filterbank, which better simulates human auditory filtering. PNCC also incorporates medium-time processing to capture longer-term signal characteristics and reduce the impact of stationary noise. Furthermore, mean power normalization is applied to stabilize energy variations, followed by power-law nonlinearity to enhance robustness against intensity fluctuations. Finally, DCT and mean normalization are used to produce compact and balanced cepstral coefficients.

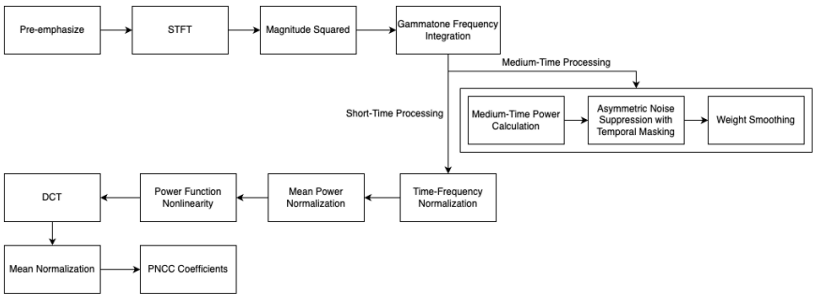


Figure 2. Steps of PNCC feature extraction.

Based on several previous studies, MFCC and PNCC each have their own advantages depending on the test conditions. In the study by Soumeya et al., PNCC appeared to perform better in noisy and reverberant environments because it was more effective at suppressing unstable noise. Its performance even improved when combined with MFCC, particularly at low SNR levels [12]. These results are also supported by Al-Kaltakchi et al., who found that PNCC is more accurate in the presence of white noise, although it requires higher computational power, while MFCC performs better for male voices and PNCC stands out for female voices [13]. Research by Lin et al. further reinforces these findings, showing that PNCC offers more comprehensive feature extraction capabilities and superior denoising

performance compared to MFCC [14]. On the other hand, Soumeya et al. demonstrated that PNCC tends to be more stable under moderate noise conditions, but MFCC remains superior under clean voice conditions and is even slightly better under very noisy conditions, while developments such as AFE-MFCC can deliver the most optimal results across various conditions [15]. In general, PNCC is more robust against noise, whereas MFCC remains relevant because it is simpler and still competitive under certain conditions.

Deep Learning for Speech Recognition

Advances in deep learning have brought significant changes to the field of audio signal processing, particularly in various applications such as speech recognition, voice pathology detection, and voice-based control systems. Compared to conventional methods, deep learning approaches can capture complex patterns in audio signals, both spatially and temporally, thereby delivering superior performance under various conditions [4].

Several studies indicate that different deep learning architectures possess distinct characteristics and advantages. A study by Ali compared the performance of several deep learning models, namely Convolutional Neural Networks (CNN), Recurrent Neural Networks based on Long Short-Term Memory (RNN-LSTM), and Transformers, in audio classification tasks. The results show that CNNs achieve the highest accuracy early in training, reaching approximately 75% after 50 epochs. Meanwhile, the Transformer model exhibits a more stable improvement pattern without hitting a plateau, although it requires a longer training time. On the other hand, RNN-LSTM tends to show stagnant performance and ranks below both CNNs and Transformers in this scenario [16].

The strengths of each architecture are also evident in different application contexts. CNN are known to be effective at extracting local features from audio signals, especially when combined with representations such as MFCC or spectrograms. This is demonstrated in a study by Jellali et al., who developed a voice command recognition system for mobile robots using a Raspberry Pi. In that study, a CNN combined with MFCC features achieved an accuracy of 89.27% and demonstrated the best performance compared to RNN-LSTM, particularly in real-time scenarios [17].

Meanwhile, RNNs and their variants, such as LSTM and GRU, excel at modelling temporal dependencies in sequential data. This capability is crucial in speech recognition, where the sequence and dynamics of the signal play a primary role. The combination of CNNs and RNNs is also widely used to leverage the strengths of both simultaneously. Research by Liu and Pei shows that integrating CNNs for local feature extraction and LSTMs for temporal modelling can improve the performance of speech recognition systems in language learning contexts. Furthermore, the use of Transformer-based models in this study yielded the best performance with a word error rate (WER) of 13.1% and an accuracy of 86.5%, which is significantly better than traditional Hidden Markov Model (HMM-GMM)-based methods [18].

Beyond general speech recognition tasks, deep learning has also demonstrated strong performance in more complex vocal pattern analysis, such as the detection of voice pathologies. Research by Amami et al. and Lee indicates that the use of a hybrid architecture combining CNNs and Bidirectional LSTM (BiLSTM) can identify various voice disorder conditions, including laryngitis, vocal nodules, and early indications of Parkinson's disease [19]. Additionally, another study comparing CNN and RNN-LSTM in voice pathology detection showed that CNN were slightly superior, achieving an accuracy of 87.11% compared to RNN-LSTM, which reached 86.52% [20].

Dialect Variation in Speech Recognition

In the context of language and dialect variations, deep learning approaches have also been used for dialect classification tasks. A study by Ye et al. using a dataset consisting of hundreds of phonics samples from several dialects showed that a hybrid model integrating CNN, GRU, and Deep Neural Network (DNN) was able to achieve the highest accuracy of 85.29%, outperforming individual models such as LSTM, GRU, and CNN. This indicates that combining architectures can enhance a model's ability to capture the complexity of pronunciation variations [21].

Research by Humayun et al. proposes a CNN-based ensemble model that combines analysis of vocal segments and the entire speech signal for native language identification. This approach utilizes a hierarchical CNN along with augmentation techniques such as low-pass filtering and achieves an average accuracy of up to 83.6% on a dataset covering five Indian native languages [22]. Furthermore, Wubet et al. evaluated the use of CNN and LSTM in classifying non-native English accents. Experimental results showed that LSTM achieved higher accuracy (84.33%) compared to CNN (79.69%), indicating the superiority of sequence-based models in capturing the temporal characteristics of accents [23].

In another study, Wubet and Lian developed a hybrid CNN-SVM model for speaker accent and gender recognition. The combination of these two methods yielded the best performance with an accuracy of 97%, surpassing the performance of each model individually, namely CNN (95%) and SVM (88%) [24]. Meanwhile, Chen et al. proposed a Mandarin accent recognition model based on pretrained Wav2Vec 2.0 combined with LSTM and feature aggregation techniques. This model achieved an accuracy of 88.6% on a dataset of non-native learners from eight countries and demonstrated better performance in distinguishing highly similar accents compared to conventional methods [25].

3. Proposed Method

This study will involve several stages, ranging from preprocessing and model development to model evaluation. Figure 3 illustrates each component and its relationship within a clearer workflow. Figure 4, meanwhile, shows the website testing workflow using the results of the best model.

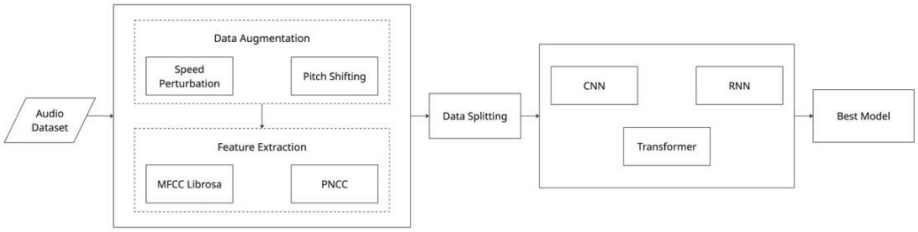


Figure 3. Model Development Flowchart.

The dataset used in this study is based on Esti Anggraini's previous research, which originally included 10 native speakers and 6 non-native speakers [4]. In this study, the sample size has been expanded to 19 native speakers and 19 non-native speakers. Of these 19 non-native speakers, 6 were from Java, 3 from Batam, 4 from Papua, 3 from Bali, and 3 from Madura. This dataset was collected by recording voice notes using a cell phone in MP3 format, which were then saved in a Google Drive folder.

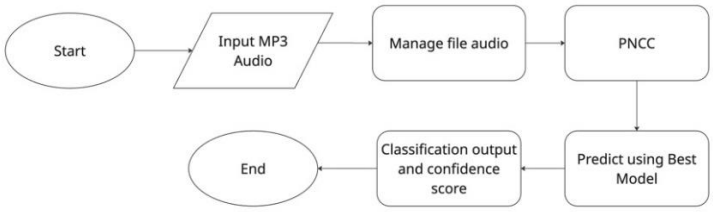


Figure 4. Website Development Flowchart.

The total number of data from native and non-native speakers is 986 audio files. However, some data is incomplete, specifically regarding the letter 'q'. A visualization of the overall dataset and the distribution of data for each speaker can be seen in Figure 5 and Figure 6.

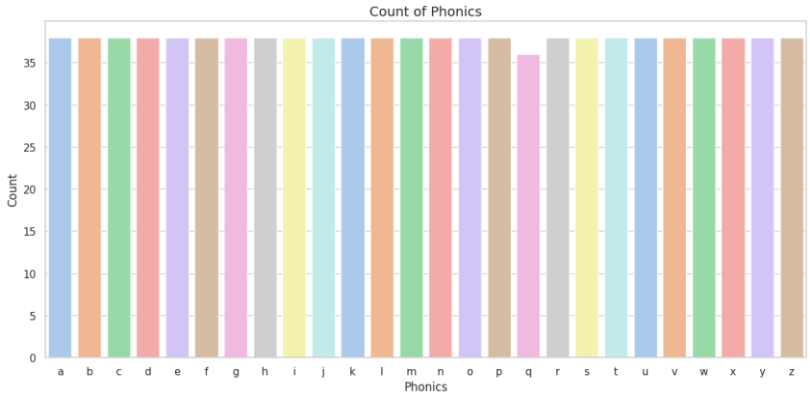


Figure 5. Overall dataset class distribution.

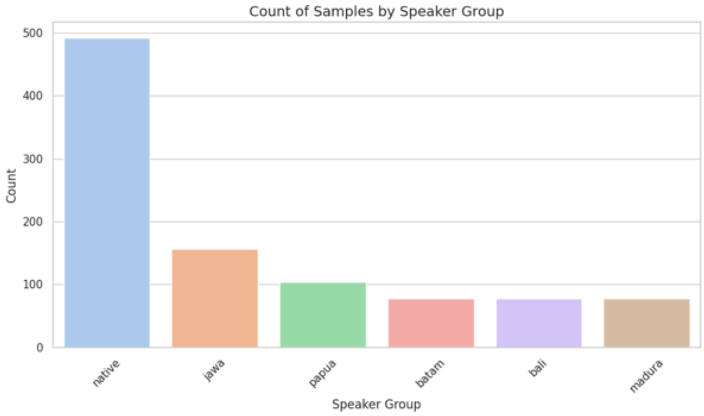


Figure 6. Speaker group class distribution.

Preprocessing

Data preprocessing in this study involves several steps, including data augmentation and feature extraction.

Data Augmentation

The initial pre-processing stage involves data augmentation to increase the variety of data from a limited dataset, enabling the model to learn more patterns and avoid overfitting. This study employs two methods: constant pitch shifting (slightly altering the pitch to keep the voice sounding natural) and constant speed perturbation (changing the overall audio speed). A visual representation of the spectrogram comparing the original signal with the results of speed perturbation and pitch shifting can be seen in Figure 7.

Unlike Esti Anggraini's study, which employed noise addition and time shifting, the approach in this study focuses more on altering the phonics characteristics of the voice rather than introducing external disturbances. Noise addition aims to make the model more robust against noise, while time shifting merely shifts the signal's temporal position. Meanwhile, speed perturbation fundamentally alters the audio's tempo, resulting in a more comprehensive range of variations.

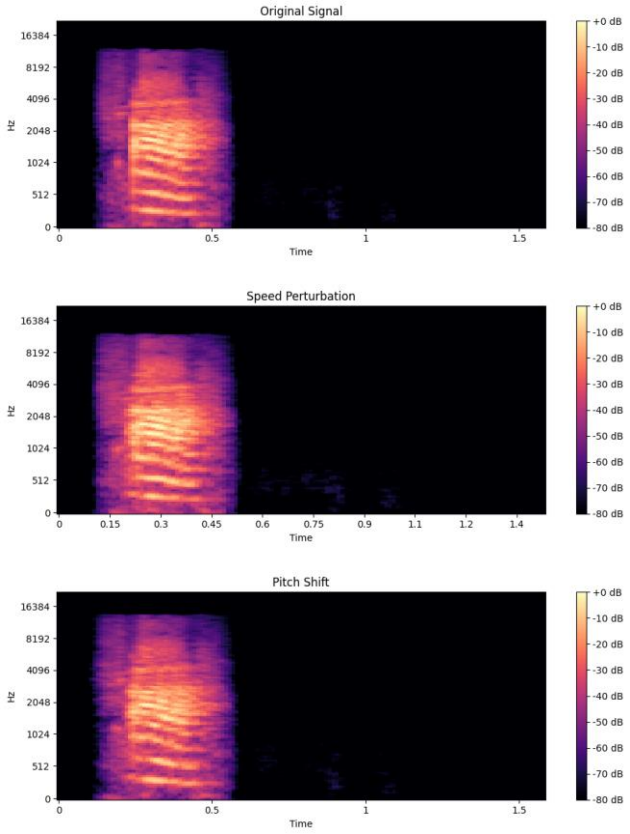


Figure 7. Data augmentation visualization.

MFCC Extraction

MFCC is a widely used feature extraction method in speech processing. The results of the MFCC feature extraction are shown in Figure 8.

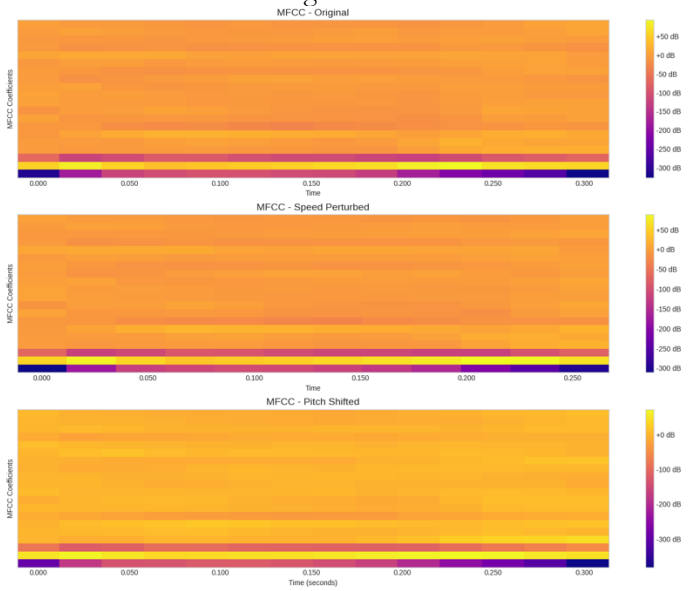


Figure 8. MFCC visualization

Pre-emphasis is an initial step aimed at amplifying the high-frequency components in an audio signal, typically by applying a high-pass filter. Afterward, the signal is divided into short, overlapping segments (frames). The frame size and the distance between frames (hop size) determine how finely the signal is divided. Each frame is then windowed, generally using a Hamming window, to reduce edge distortion and prevent spectral leakage during the Fast Fourier Transform (FFT). This process converts the signal from the time domain to the frequency domain. Next, the frequency transform results are processed using a Mel filterbank to map frequencies to the Mel scale, which approximates how humans perceive sound. After that, a DCT (Discrete Cosine Transform) is performed to simplify the data and generate MFCC coefficients. Several key coefficients are selected to represent the important characteristics of the sound. This process can be adjusted as needed by setting the number of coefficients, frame size, or hop size to ensure the feature extraction results are truly optimal.

PNCC Extraction

PNCC is a feature extraction method designed to be more noise-resistant than MFCC, involving a series of processes to stabilize and highlight the main signal. The process begins with pre-emphasis to emphasize high frequencies; the signal is then transformed into the frequency domain using STFT and processed into a power spectrogram, which is subsequently mapped to a mel scale to align with human auditory perception. Next, medium-time power is calculated using a moving average to smooth energy across frames, preventing drastic fluctuations. To separate noise, an asymmetric low-pass filter is applied to form the signal's envelope, which is then subtracted from the main signal, followed by rectification to eliminate negative values. Subsequent steps include temporal masking to suppress weak energy, energy-threshold-based selection, and weight smoothing to reduce spectral fluctuations. Next, mean power normalization is performed to stabilize energy differences between frames, followed by power-law-based non-linear compression, which is more robust against noise than logarithmic compression. In the final stage, features are normalized using mean-standard deviation normalization, resulting in a two-dimensional matrix representation of PNCC ready for classification. The results of the MFCC feature extraction are shown in Figure 9.

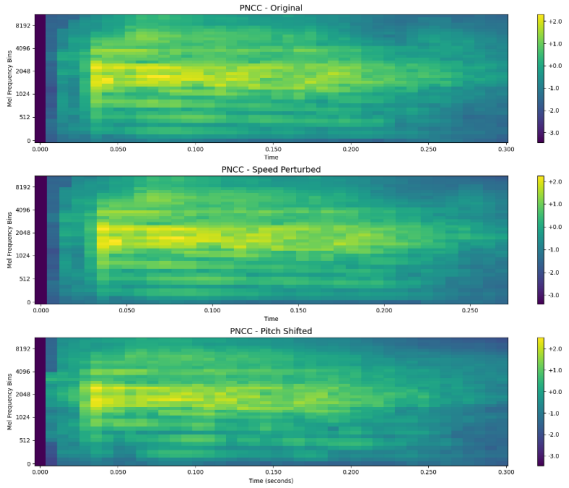


Figure 9. PNCC visualization.

Modelling

This study uses three deep learning methods, namely Convolutional Neural Network (CNN), GRU-based Recurrent Neural Network (RNN), and Transformer to classify phonics. All models are optimized using the Optuna hyperparameter tuning technique, which automatically searches for the best parameter combination through pruning and sampling based on validation performance.

The CNN model in this study was built with several convolution blocks equipped with SEBlock (Squeeze-and- Excitation). SEBlock aims to adjust the weight of each channel based on its actual contribution to the final result, so that the most prominent features in the signal can be strengthened. This technique helps the model focus on the important elements of the audio signal. Each CNN block consists of two convolution layers, batch normalization, ReLU activation, SEBlock, max pooling, and dropout. Following this, adaptive average pooling is applied, followed by a fully connected layer with dense units and dropout for final classification.

The RNN model in this study was constructed using multiple layers of Gated Recurrent Units (GRU) capable of capturing sequential patterns in the audio signal. This GRU architecture is designed to be bidirectional, so it can process data sequences in both directions, from start to end and vice versa, to understand the relationships between sounds in a single sound sequence comprehensively. The number of GRU layers, hidden state size, bidirectional usage, and other parameters are determined using Optuna to produce the most optimal configuration. The output from the GRU is then further processed by the Multihead Attention Layer, which identifies important parts of the sequence based on the inter-frame audio correlation patterns. The attention results are averaged using mean pooling and then passed into the classification block, which consists of a dense layer, LayerNorm, and dropout. The activation functions used, such as Mish, GELU, or ReLU, are also selected based on the best tuning results.

The Transformer model is built with an encoder stack of several TransformerEncoderLayers, equipped with self- attention and feed-forward networks. The initial input is projected to dimension d_model and then added with positional encoding to preserve sequence information. The output from the Transformer encoder is temporally averaged (mean pooling) before entering the classifier. Parameters such as the number of layers, number of heads, hidden size, and dense layer dimension are determined using Optuna to achieve the best performance.

Experimental Scenarios

This study conducted five experiments to obtain the best results in identifying phonics pronunciation. The experiments conducted are shown in Table 1.

Table 1. Experimental scenarios used.

Scenario	Description
A	MFCC Librosa Extraction
B	PNCC Extraction
C	Native PNCC Extraction
D	Nonative PNCC Extraction
E	Nonative per Region PNCC Extraction

Scenario A and B were tested using the entire dataset to evaluate the performance of the MFCC Librosa and PNCC feature extraction methods. Scenarios C and D used PNCC feature

extraction to evaluate the performance of the model and the feature extraction method on datasets of native and non-native speakers. Finally, Scenario E used PNCC feature extraction and was tested on a dataset of non-native speakers by region to evaluate the model's recognition performance on datasets of speakers from five various regions in Indonesia.

4. Results and Discussion

The experiment results are presented in three groups according to each experiment described before, namely evaluation of feature extraction methods, speaker types, and non-native speaker regions of origin. Each test was conducted using CNN, RNN-GRU, and Transformer models, then evaluated using four main metrics: accuracy, precision, recall, and F1 Score. Scenario A and B aim to compare the performance of the model when using two different feature extraction methods, namely MFCC from Librosa and PNCC. The test results are shown in Table 2.

Table 2. Results of Scenario A and Scenario B.

Feature Extraction	Model	Accuracy	F1-Score
MFCC	CNN	0.8919	0.891
	RNN	0.9291	0.9289
	Transformer	0.8125	0.8126
PNCC	CNN	0.9071	0.9067
	RNN	0.9459	0.946
	Transformer	0.897	0.8966

From these results, the PNCC feature produces improved performance across all models compared to MFCC. The RNN model recorded the best results with an accuracy of 94.59% and a F1 Score of 0.946 when using PNCC, an improvement from an accuracy of 92.91% and a F1 Score of 0.9289 with MFCC. CNN also saw an improvement from a F1 Score of 0.891 to 0.9067, and Transformer improved from 0.8126 to 0.8966. This shows that PNCC is more effective in representing relevant acoustic information for phonics classification, especially in handling pronunciation variations between speakers. The visualization of the confusion matrix comparison for the best model (RNN-GRU) using MFCC Librosa and PNCC feature extraction is shown in Figure 10.

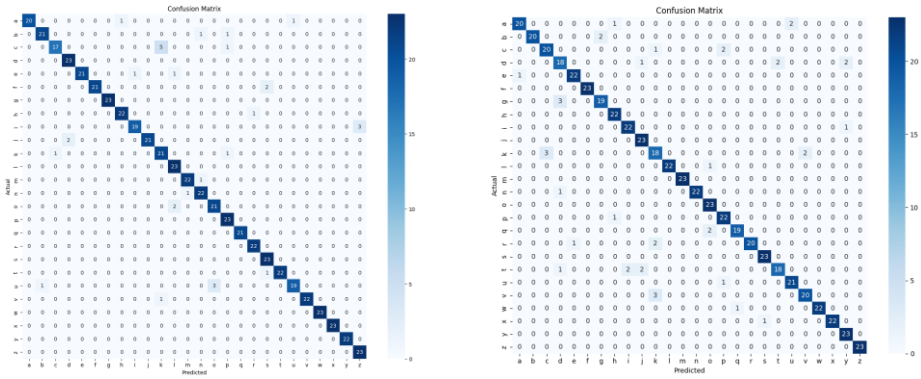


Figure 10. Confusion matrix visualization (a) MFCC Librosa with RNN-GRU model; (b) PNCC with RNN-GRU model.

Scenario C and D were conducted to measure the impact of speaker type on model performance. The data was separated into native and non-native speaker groups. The test results are shown in Table 3. The results show that all models achieved higher scores when tested with data from native speakers. RNN-GRU remained the most superior model with a F1 Score of 0.9385 and an accuracy of 93.92 on native data, compared to 0.9305 and 92.93 on non-native data. Improvements were also seen in CNN and Transformer. This indicates that the consistency of articulation and more standard pronunciation of native speakers makes it easier for models to recognize phonics patterns.

Table 3. Results of Scenario C and Scenario D.

Speaker Dataset	Model	Accuracy	F1-Score
Native Speaker	CNN	0.9324	0.9329
	RNN	0.9392	0.9385
	Transformer	0.9257	0.9246
Non-native Speaker	CNN	0.8923	0.8906
	RNN	0.9293	0.9305
	Transformer	0.8519	0.8555

Scenario E aims to analyse the influence of regional accents on the accuracy of phonics classification using PNCC feature extraction. The test results from five regions of origin, namely Bali, Batam, Java, Madura, and Papua, are shown in Table 4. The results show that speakers from Papua recorded the highest scores in all models. CNN and RNN-GRU both achieved an accuracy of 99.04% and an F1 score of 0.9902. Meanwhile, the lowest performance was recorded in data from Madura and Bali, particularly in the Transformer model, which achieved an F1 Score of 0.8912 (Madura), and the CNN model, which reached an F1 Score of 0.8102 (Bali). These differences indicate that the distinctive dialectal characteristics of each region influence the models' effectiveness in recognizing phonics.

Table 4. Evaluation Results Based on Non-native Speaker Origin.

Model	Region	Accuracy	F1-Score
CNN	Batam	0.9487	0.9341
	Bali	0.8205	0.8102
	Jawa	0.9103	0.9135
	Madura	0.9487	0.9385
	Papua	0.9904	0.9902
RNN	Batam	0.9615	0.9464
	Bali	0.9615	0.9451
	Jawa	0.9615	0.961
	Madura	0.9615	0.9487
	Papua	0.9904	0.9902
Transformer	Batam	0.9615	0.9464
	Bali	0.9615	0.9487
	Jawa	0.9359	0.935
	Madura	0.8974	0.8912
	Papua	0.9808	0.9806

Based on the experimental results, the best-performing model is the PNCC-based RNN-GRU. This model is then implemented in a web-based system for phonics classification. The testing website developed in this study is designed to classify phonics directly through a web-based interface. The process begins when the user uploads an MP3 audio file. Once the file is received, the system extracts features using the Power-Normalized Cepstral Coefficients (PNCC) method to convert the audio signal into a numerical representation. The extraction results are then processed by the best classification model, namely RNN-GRU. The model generates phonics predictions in the form of letter labels accompanied by confidence values. The website interface is designed to be simple and intuitive, featuring file upload, direct recording, and real-time classification result display. The system's interface can be seen in Figure 10, which shows the audio input form and visualization of prediction results.

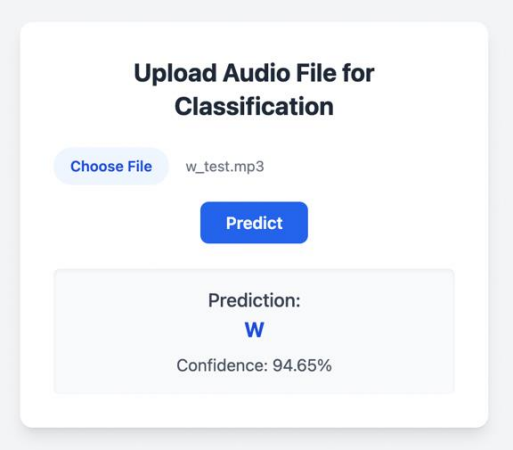


Figure 11. User Interface Design.

5. Comparison

Based on the experimental results, the method proposed in this study demonstrates competitive performance when compared to the study by Esti Anggraini. In that study, performance improvements were achieved through the use of augmentation techniques such as noise addition and time shifting, which focused on enhancing resistance to signal interference. Meanwhile, this study employs a different augmentation approach, specifically

pitch shifting and speed perturbation, which places greater emphasis on variations in phonics characteristics and natural changes in pronunciation, rather than merely noise robustness. This difference in approach indirectly influences the characteristics of the data learned by the model.

In terms of performance, the best model in this study, the PNCC-based RNN-GRU, achieved an accuracy of 94.59% and an F1 score of 0.946. Although these values are slightly below the results reported by Esti Anggraini, who produced the best SVM model using MFCC Librosa feature extraction with an accuracy of 0.95.57% and an F1 score of 0.9553, the difference is relatively small and still falls within a comparable performance range. This difference is likely influenced by variations in the dataset, augmentation methods, and the complexity of the data used. This study uses data with a wider variety of pronunciations, so naturally, the level of difficulty is also higher compared to more controlled scenarios.

Although both studies evaluated the differences between native and non-native speakers, this study provides a more detailed analysis by specifically examining the influence of regional dialect variations. The analysis was conducted across five different regions: Java, Batam, Madura, Papua, and Bali. The results indicate that differences in dialect characteristics have a significant impact on model performance. For example, data from Papua yielded the highest performance on some models, while data from Bali and Madura tended to be more challenging and resulted in lower performance. These findings suggest that dialect variations not only affect overall accuracy but can also be a key factor in understanding the limits of a model's ability to recognize phonics patterns. This approach provides a more realistic perspective on the real-world application of the system, where accent diversity is unavoidable.

Furthermore, the experimental results also show that the consistent use of PNCC features yields improved performance compared to MFCC across all tested models, including CNN, RNN-GRU, and Transformer. This reinforces the assumption that PNCC is better able to represent relevant acoustic information, particularly in conditions with pronunciation variations among speakers. Thus, the contribution of this research lies not only in model selection but also in the selection of feature representations that are more robust to speech variations.

In addition, this study does not stop at the model evaluation stage but also implements the best model into a web-based system for direct phonics classification. This implementation demonstrates that the proposed model is not only theoretically effective but can also be applied in real-world scenarios with a simple and user-friendly interface. This represents an added value compared to previous studies, which generally still focus on model evaluation without direct system implementation.

Overall, although the accuracy results obtained not surpassing the previous study, the proposed approach still makes a significant contribution. These contributions include the use of different augmentation strategies, a more in-depth analysis of performance across regional dialects, the utilization of more robust PNCC features, and the implementation of a web-based system. The combination of these aspects strengthens the practical value and relevance of this research in the development of phonics classification systems that are more adaptive to variations in pronunciation.

6. Conclusions

This study developed a deep learning model for phonics recognition through three main stages: data augmentation, feature extraction, and classification. In the augmentation stage, pitch shifting and speed perturbation methods were used to increase data variability. Feature extraction was performed using MFCC and PNCC, with PNCC consistently demonstrating better performance. In the classification stage, three model architectures were used: CNN, RNN-GRU, and Transformer. Based on all experiments, the RNN-GRU model with PNCC features yielded the best results, achieving an accuracy of 94.59% and an F1 score of 0.946.

Test results indicate that using PNCC improves performance compared to MFCC across all tested models, suggesting that PNCC is more effective at representing acoustic information for phonics recognition. Additionally, models trained on data from native speakers outperform those trained on non-native speakers, indicating that pronunciation consistency impacts classification accuracy. Testing on regional dialect variations in Indonesia also showed that dialect differences affect model performance, with data from Papua yielding the best results, while Bali and Madura tended to produce lower performance. Overall, this study confirms that the selection of feature extraction methods, speaker characteristics, and dialect variations are key factors in developing a phonics recognition system that is more adaptable to phonics pronunciation diversity.

For future development, there are several areas for improvement. First, the dataset should be expanded to include a wider variety of local dialects, as this study is currently limited to five regions: Bali, Batam, Java, Madura, and Papua. Adding dialects from other regions is expected to provide a broader picture of pronunciation variations. Second, recording quality needs to be improved to ensure the data used is cleaner from noise and more consistent across speakers. Third, future research could explore other classification methods not used in this study to assess potential performance improvements. Finally, the use of additional augmentation techniques such as frequency masking, time warping, or a combination of several augmentation methods could be considered to enrich data variation and enhance model robustness.

References

[1] Ehri, L. C. (2020). The science of learning to read words: A case for systematic phonics instruction. Reading Research Quarterly, 55(S1). https://doi.org/10.1002/rrq.334

[2] Munggaraning Westhisi, S. (2019). Metode fonik dalam pembelajaran membaca permulaan bahasa Inggris anak usia dini. Tunas Siliwangi, 5(1), 29–43. https://doi.org/10.22460/ts.v5i1p29-43.1271

[3] Utami, S., & Musthafa, B. (2023). The utilization of phonics songs in phonics reading classes in Indonesia: Teachers' perspectives. Jurnal Pendidikan Bahasa dan Sastra, 22(2), 201–208. https://doi.org/10.17509/bs_jpbsp.v22i2.55911

[4] Anggraini, R. N. E., Razak, F. A., & Sunaryono, D. (2024). Phonics recognition using machine learning with MFCC extraction. In International Conference on Information and Communications Technology (ICOIACT) (pp. 170–175). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICOIACT64819.2024.10913214

[5] Mehrish, A., Majumder, N., Bharadwaj, R., Mihalcea, R., & Poria, S. (2023). A review of deep learning techniques for speech processing. Information Fusion, 99, 101869. https://doi.org/10.1016/j.inffus.2023.101869

[6] Zhang, Z., He, C., & Yang, K. (2020). A novel surface electromyographic signal-based hand gesture prediction using a recurrent neural network. Sensors, 20(14), 3994. https://doi.org/10.3390/s20143994

[7] Pham, N.-Q., Nguyen, T.-S., Niehues, J., Müller, M., & Waibel, A. (2019). Very deep self-attention networks for end-to-end speech recognition. In Interspeech 2019 (pp. 66–70). ISCA. https://doi.org/10.21437/Interspeech.2019-2702

[8] Kim, C., & Stern, R. M. (2016). Power-normalized cepstral coefficients (PNCC) for robust speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 24(7), 1315–1329. https://doi.org/10.1109/TASLP.2016.2545928

[9] Chamidy, T. (2016). Metode Mel Frequency Cepstral Coeffisients (MFCC) pada klasifikasi Hidden Markov Model (HMM) untuk kata Arabic pada penutur Indonesia. MATICS, 8(1), 36. https://doi.org/10.18860/mat.v8i1.3482

[10] Sasilo, A. A., Saputra, R. A., & Ningrum, I. P. (2022). Sistem pengenalan suara dengan metode Mel Frequency Cepstral Coefficients dan Gaussian Mixture Model. Komputika: Jurnal Sistem Komputer, 11(2), 203–210. https://doi.org/10.34010/komputika.v11i2.6655

[11] Maurya, A., Kumar, D., & Agarwal, R. K. (2018). Speaker recognition for Hindi speech signal using MFCC-GMM approach. Procedia Computer Science, 125, 880–887. https://doi.org/10.1016/j.procs.2017.12.112

[12] Soumeya, B., Djamel, A., & Ahmed, S. S. (2026). A multifaceted approach based on deep learning architectures for dysarthric speech recognition. Institute of Electrical and Electronics Engineers (IEEE), 1–5. https://doi.org/10.1109/IC2SDA68097.2025.11331376

[13] Al-Kaltakchi, M. T. S., Al-Raheem Taha, H. A., Shehab, M. A., & Abdullah, M. A. M. (2020). Comparison of feature extraction and normalization methods for speaker recognition using grid-audiovisual database. Indonesian Journal of Electrical Engineering and Computer Science, 18(2), 782–789. https://doi.org/10.11591/ijeecs.v18.i2.pp782-789

[14] Lin, H., et al. (2025). Recognition of wood-boring insect creeping signals based on residual denoising vision network. Sensors, 25(19). https://doi.org/10.3390/s25196176

[15] Soumeya, B., Djamel, A., & Ahmed, S. S. (2023). Improved ASR system based on CNN-LSTM acoustic model for mobile voice. In Proceedings of the 2023 2nd International Conference on Electronics, Energy and Measurement (IC2EM 2023). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/IC2EM59347.2023.10419854

[16] Ali, Z. (2025). A comprehensive overview and comparative analysis of CNN, RNN-LSTM and transformer. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5175090

[17] Jellali, A., Ben Fredj, I., Zouhir, Y., & Ouni, K. (2022). Applying a deep learning technique for speech recognition in robotics. In 2022 IEEE Information Technologies and Smart Industrial Systems (ITSIS 2022). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ITSIS56166.2022.10118430

[18] Liu, Y., & Pei, S. (2024). Research on Spanish classroom speech recognition system based on deep learning algorithm. In 2024 5th International Conference on Information Science and Education (ICISE-IE 2024) (pp. 119–122). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ICISE-IE64355.2024.11025398

[19] Amami, R., Amami, R., Trabelsi, C., Mabrouk, S. H., & Khalil, H. A. (2023). A robust voice pathology detection system based on the combined BiLSTM-CNN architecture. MENDEL, 29(2), 202–210. https://doi.org/10.13164/mendel.2023.2.202

[20] Lee, J.-Y. (2021). Experimental evaluation of deep learning methods for an intelligent pathological voice detection system using the Saarbruecken Voice Database. Applied Sciences, 11(15), 7149. https://doi.org/10.3390/app11157149

[21] Ye, S., Zhao, R., & Fang, X. (2019). An ensemble learning method for dialect classification. IOP Conference Series: Materials Science and Engineering, 569(5), 052064. https://doi.org/10.1088/1757-899X/569/5/052064

[22] Humayun, M. A., Yassin, H., & Abas, P. E. (2022). Native language identification for Indian-speakers by an ensemble of phoneme-specific, and text-independent convolutions. Speech Communication, 139, 92–101. https://doi.org/10.1016/j.specom.2022.03.007

[23] Wubet, Y. A., Balram, D., & Lian, K.-Y. (2023). Intra-native accent shared features for improving neural network-based accent classification and accent similarity evaluation. IEEE Access, 11, 32176–32186. https://doi.org/10.1109/ACCESS.2023.3259901

[24] Wubet, Y. A., & Lian, K.-Y. (2021). A hybrid model of CNN-SVM for speakers' gender and accent recognition using English keywords. In 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW) (pp. 1–2). IEEE. https://doi.org/10.1109/ICCE-TW52618.2021.9603210

[25] Chen, J., Zhou, H., Ma, H., Dai, C., Zhao, Y., & Wang, H. (2023). Non-native Chinese accent recognition model based on pretrained Wav2vec2.0 and feature aggregation. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4626817