International Journal of Electronics and Communications System
Volume 4, Issue 1, 27-41.
ISSN: 2798-2610
http://ejournal.radenintan.ac.id/index.php/IJECS/index
DOI: 10.24042/ijecs.v4i1.22918

Digital Signal Processing for The Development of Deep Learning-Based Speech Recognition Technology

Dita Novita Sari*
Institut Teknologi dan Bisnis Bakti Nusantara (IBN), Lampung, INDONESIA

Danang Kusnadi
Institut Teknologi dan Bisnis Bakti Nusantara (IBN), Lampung, INDONESIA

Ricco Herdiyan Saputra
Institut Teknologi dan Bisnis Bakti Nusantara (IBN), Lampung, INDONESIA

Mujeeb Ullah Khan
Govt Degree College Takhtaband, Q8JG+FV5, Mingora, Swat, Khyber Pakhtunkhwa, PAKISTAN

Article Info

Article history:

Received: March 28, 2024
Revised: June 4, 2024
Accepted: June 26, 2024

Keywords:

Digital Signal Processing;
Deep Learning;
Speech Recognition;
Technology.

Abstract

This research discusses digital signal processing in the context of developing deep learning-based speech recognition technology. Given the increasing demand for accurate and efficient speech recognition systems, digital signal processing techniques are essential. The research method used is an experimental method with a quantitative approach. This research method consists of several stages: introduction, research design, data collection, data preprocessing, Deep Learning Model Development, performance training and evaluation, experiments and testing, and data analysis. These findings are expected to contribute to developing more sophisticated and applicable speech recognition systems in various fields. For example, in virtual assistants such as Siri and Google Assistant, improved speech recognition accuracy will allow for more natural interactions and faster responses, improving the user experience. This technology can be used in security systems for safer and more reliable voice authentication, replacing or supplementing passwords and fingerprints. Additionally, in accessibility technology, more accurate voice recognition will be particularly beneficial for individuals with visual impairments or mobility, allowing them to control devices and access information with just voice commands. Other benefits include improvements in automated phone apps, automatic transcription for meetings or conferences, and the development of smart home devices that can be fully voice-operated.

To cite this article: D. N. Sari, D. Kusnadi, R. H. Saputra, and M. U. Khan, "Digital signal processing for the development of deep learning-based speech recognition technology," *Int. J. Electron. Commun. Syst.* Vol. 4, No. 1, pp. 27-41, 2024.

INTRODUCTION

The development of information and communication technology has had a significant impact on various aspects of human life [1]. Information and communication technology development has increased demand for more sophisticated and accurate speech recognition systems [2]. Speech recognition technology allows machines to understand and respond to human voice commands, opening up various application opportunities ranging from virtual assistants, and security systems, to accessibility technologies. Information technology, used to process data, plays a central role in this advancement by providing the tools and methods necessary to handle and analyze complex voice data [3]. This includes processing, obtaining, compiling, storing, and manipulating data in a variety of ways to produce quality information, i.e., relevant, accurate, and timely information used for individual, corporate, and governmental purposes, as well as strategic information for decision-making [4]. The advancement of

- **Corresponding author:**
Ricco Herdiyan Saputra, Information System, Bakti Nusantara Institute Lampung, Pringsewu Sub-district, Pringsewu, Lampung, INDONESIA. ✉ saputraherdiyanricco@gmail.com

information technology is very important for the progress of the times. Education, economy, health, government, and socio-culture are some of the important areas where technological advancements affect the progress of the country [5]. Technology was created to facilitate human work, but now it has become a necessity for humans and has been used in all aspects of human life.

The development of information and communication technology has driven an increase in demand for more sophisticated and accurate speech recognition systems [6]. Speech recognition technology allows machines to understand and respond to human voice commands, opening up various application opportunities ranging from virtual assistants, and security systems, to accessibility technologies. One of the main challenges in the development of speech recognition systems is how to handle and process complex and varied voice signals. This is where the important role of digital signal processing (DSP) comes in.

One of the technologies that continues to grow rapidly is voice recognition, which allows interaction between humans and machines through voice commands [7]. The ability to convert incoming sounds on the computer into text. Speech Recognition combines computer science and linguistics to identify spoken words and convert them into text [8]. This technology has been applied to a wide range of devices, from virtual assistants like Siri and Google Assistant to biometric-based security systems [9]. There has been significant progress in the development of speech recognition technology, while there are still many challenges that need to be overcome [10]. Digital Signal Processing (DSP) plays a crucial role in improving the performance of speech recognition systems [11]. DSP enables the processing and analysis of sound signals to extract important features necessary for accurate recognition [12], [13].

The features referenced from the study refer to the important elements or characteristics of the voice signal that are identified and used for speech recognition purposes. The main features commonly extracted in sound signal processing for accurate recognition such as Mel-Frequency Cepstral Coefficients (MFCCs) which are representations of the amplitude spectrum of sound signals commonly used because they are effective in capturing frequency information relevant to speech recognition [14], spectrograms of visual representations of the frequency spectrum of sound signals that change over time, chroma features that summarize pitch information, zero-crossing rate (ZCR) is the number of legs of a sound signal crossing the zero axis in a given period of time, short-time energy that measures the energy of a sound signal, formant frequencies that are essential for vocal recognition and speech characteristics, linear predictive coding (LPC) coefficients to estimate the spectral characteristics of a sound signal, delta and delta-delta coefficients a feature that represents the dynamic change of spectral features such as MFCCs [15], Pitch, and harmonics and is useful for the recognition of vocal characteristics and intonation, and then there is the band energy ratio which is useful for measuring energy in various frequency bands [16].

Transformations such as Fast Fourier Transform (FFT), Mel-Frequency Cepstral Coefficients (MFCC), and filtering techniques are used to prepare the sound signal before it is further processed by the recognition algorithm [17], [18]. Computer learning is classified directly from images or sounds in deep learning, which is a branch of machine learning science based on the Artificial Neural Networks (ANNs) [19]. One of the deep learning algorithms, the Convolutional Neural Network (CNN/ConvNet), was developed from the Multilayer Perceptron (MPL). CNNs function to process two-dimensional data, such as sound or images [20]. With the ability to learn directly from images, CNN reduces the burden of programming [21].

On the other hand, advances in deep learning have shown great potential in various pattern recognition applications, including speech recognition [22]. Deep learning, is a branch of artificial intelligence, that uses artificial neural networks capable of learning and recognizing patterns from huge amounts of data. DSP integration with deep learning enables more comprehensive and accurate processing of voice signals. Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) are some of the deep learning architectures that are effective in handling sequential and complex voice data.

Deep learning-based approaches have shown great potential in improving speech

recognition performance [23]. Deep learning models, particularly neural networks, can learn complex data representations and generalize patterns in voice data [24]. This approach allows speech recognition systems to be more adaptive and resistant to variations in voice input [25]. The implementation of deep learning in speech recognition also requires the support of effective digital signal-processing techniques [26]. Precise signal processing can improve the quality of the data provided to deep learning models, thereby improving the accuracy and speed of recognition. Therefore, research on the integration of digital signal processing techniques with deep learning algorithms is of great importance [27].

Previous research focusing on the development of deep learning-based speech recognition technology has covered various implementations, including web-based environmental security design [28], human speech recognition case study of lecturer's voice [29], real-time object identification on Android-based platforms [30], speech recognition for home automation systems with command control [31], improving English speech sounds based on deep learning from signal processing to semantic recognition [32], and speech signal processing using deep learning in the form of an app [33]. However, as far as researchers have found, there has been no comprehensive study discussing digital signal processing in the context of developing deep learning-based speech recognition technology.

Therefore, the great potential of the combination of DSP and deep learning, This research aims to identify the most effective digital signal processing techniques for voice feature extraction, evaluate the performance of various deep learning architectures in speech recognition, Develop and test speech recognition systems that combine DSP and deep learning to achieve higher accuracy and efficiency compared to conventional methods. By combining the right DSP techniques and innovative deep learning architectures, this research is expected to make a significant contribution to the development of more advanced and applicable speech recognition technologies in various fields.

METHOD

The purpose of this research is to develop speech recognition technology that uses deep learning through digital signal processing. The research method used is an experimental method with a quantitative approach. This research method consists of several stages: introduction, research design, data collection, data preprocessing, deep learning model development, performance training and evaluation, experiments and testing, and data analysis [34].

1. Introduction

This section describes the steps and approaches that will be used in research on digital signal processing for the development of deep learning-based speech recognition technology [35]. This study aims to identify effective digital voice signal processing techniques to improve the accuracy of deep learning-based speech recognition systems. The techniques and approaches used are sound signal pre-processing, feature extraction, deep learning models, evaluation, and validation. This section also built a flowchart of the algorithm to make sure the process running well.

2. Research Design

This study uses an experimental research design involving several main stages, as follows:

a) **Data collection** to collect voice data from various sources to build a representative dataset.
b) **Data** pre-processing performs data pre-processing including noise removal, signal normalization, and framing and windowing.
c) **Development of deep learning models** to design deep learning model architectures suitable for sound signal processing
d) **Train and evaluate the model** to train the model using the processed dataset, then evaluate its performance using the cross-validation method
e) **Analyze results** to identify the model's strengths and weaknesses and iterate on improvements if needed

3. Data Collection

a) **Data Source:** Voice data will be collected from publicly available datasets such as the Librispeech ASR corpus or Google Speech Commands Dataset. In addition, additional data can be collected through voice recordings from various sources to add to the variety of datasets.

b) **Tools and Devices:** High-quality microphones will be used for voice recording to ensure good data quality.

c) **Collection Procedure:** Data will be recorded in WAV format with a sampling rate of 16 kHz and a resolution of 16-bit. The collected voice data will be divided into classes based on the words or phrases spoken.

4. Data Preprocessing

a) **Normalization:** The process of adjusting the signal amplitude to ensure that all data is at a consistent scale, facilitating further processing and improving the performance of deep learning models. There are several steps such as measuring the maximum amplitude to identify the maximum amplitude value of the entire sound dataset, dividing the maximum amplitude of each sample in the sound signal will be divided by the maximum amplitude value, then there is a scaling range to rescale the signal into a specific range, for example, -1 to 1 or 0 to 1.

b) **Noise Reduction:** Noise reduction techniques such as Spectral Gating or Wiener Filtering will be used to clear the signal from background noise.

c) **Framing and Windowing:** The sound signal will be broken down into small frames of a certain duration using a windowing function such as a Hamming or Hanning window.

d) **Feature Extraction:** Important features such as Mel-Frequency Cepstral Coefficients (MFCC), Spectrogram, and Mel-Spectrogram will be extracted from the sound signal [8], [36].

5. Deep Learning Model Development

a) **Model Architecture:** The deep learning model to be used is the Convolutional Neural Network (CNN), the Recurrent Neural Network (RNN), or a combination of the two (CRNN). The model structure will be customized for speech recognition, such as using Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU) to handle data sequences. LSTMs and GRUs are designed to address vanishing gradient problems common in sequence data processing, allowing them to retain information over longer periods and capture long-term dependencies in sound signals. This is particularly relevant in speech recognition because the temporal context of voice data is essential for accurately understanding and identifying sound patterns.

b) **Model Parameters:** Parameters such as the number of layers, kernel size, number of units on each layer, and activation function will be determined through initial experiments.

c) **Frameworks and Tools:** Models will be built using deep learning frameworks such as TensorFlow or PyTorch [37], [38].

6. Model Training and Evaluation

a) **Dataset Distribution:** The dataset will be divided into three parts, namely the training set (70%), the validation set (15%), and the test set (15%). This distribution was chosen to ensure the model could be effectively trained, validated for hyperparameter adjustment, and accurately tested.

1. Training Set (70%): The largest portion of the dataset is used to train the model, allowing the model to learn and adjust weights based on the available data. The 70% proportion ensures that the model has enough data to study patterns and features in the sound signal.

2. Validation Set (15%): Validation sets are used to perform hyperparameter tuning and prevent overfitting. Using 15% of the dataset, the model can be evaluated periodically during the training process, allowing researchers to make adjustments without using the same data as in the training set.
3. Test Set (15%): The test set is used to evaluate the final performance of the model after training is complete. With 15% of the dataset set aside for testing, the model can be tested on data that has never been seen before, providing an accurate picture of how the model will perform on real data.

This division ensures that the model is trained with enough data, validated with different data to prevent overfitting, and tested with independent data to provide a robust and objective evaluation of its performance.

b) **Training Method:** The model will be trained using an optimizer such as Adam or SGD with an adjusted learning rate. The loss function used is Categorical Cross-Entropy.
c) **Evaluation:** The model's performance will be evaluated using accuracy, precision, recall, and F1-score metrics. In addition, the Confusion Matrix will be used to see the model's performance in recognizing each class [39].

7. Experiment and Testing

a) **Experiments:** multiple experiments will be conducted with variations in parameters and model architecture to find the optimal configuration.
b) **Testing:** Test data is an important component in the development and evaluation of machine learning models, including federated learning. The test data should be prepared carefully to ensure that the test results truly reflect the model's ability to generalize to data that has never been seen before. Steps to prepare test data such as data separation using random splits and stratified splits, normalization, and preprocessing, and sampling techniques such as oversampling or undersampling are used to ensure that the data being tested is not biased towards a particular class

8. Data Analytics

Error analysis is the process of identifying and understanding the cause of errors in the system or model used. In the context of deep learning-based speech recognition, error analysis includes the following steps:

a) **Error Identification:** a classification error that specifies when and where the model misclassifies a sound, a detection error that identifies a model's failure to detect a relevant sound signal or capture noise as a signal, a labeling error that identifies training or test data that may be mislabeled.
b) **Error grouping:** systematic errors that occur consistently under certain conditions, such as background noise or different accents, and random errors that occur without a clear pattern caused by variations in sound signals or algorithm failures
c) **Model Performance Analysis:** there are 2 namely; the Confusion Matrix is a tool to evaluate the performance of model classification by showing the number of correct and incorrect predictions for each class and then evaluation methods using metrics such as accuracy, precision, recall, and F1-score to measure model performance
d) **Error Analysis by Features:** there are 2 namely; signal analysis to see how signal features such as frequency, amplitude, and duration affect errors. Then analyze the environment to evaluate how the recording environment affects the model's performance.

After identifying and understanding the errors, the next step is to develop methods to improve the performance of the speech recognition system. Here are some potential methods for improvement:

a) **Improved Data Quality:**
   1. Data Augmentation**:** Increase the amount and variety of training data by adding variations of sound signals such as pitch shifting, time stretching, and noise addition.
   2. Data Cleanup: Removes or repairs data that is mislabeled or contains excessive noise.

b) **Model Repair**:
   1. Better Model Architecture: Develop more complex and efficient deep learning architectures such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs) with Long Short-Term Memory (LSTM).
   2. Hyperparameter Tuning: Adjust hyperparameters such as learning rate, batch size, and number of layers to improve model performance.

c) **Use of Better Preprocessing Techniques:**
   1. Filtering and Normalization: Uses filtering techniques to eliminate noise and normalize to ensure the sound signal is within the desired range.
   2. Feature Extraction: Uses better feature extraction techniques such as Mel-Frequency Cepstral Coefficients (MFCC), Spectrograms, or Mel-Spectrograms to produce more informative features.

d) **Regularization and Overfitting Avoidance Techniques:**
   1. Dropout: A technique used to prevent overfitting by randomly removing neurons during training.
   2. More Effective Data Splitting: Uses cross-validation to ensure that the model does not overfit against the training data.

e) **Training with More Varied Data:**
   1. Transfer Learning: Using a model that has been trained on a large dataset and applying it to a smaller dataset through fine-tuning.
   2. Ensemble Methods: Combines multiple models to improve robustness and prediction performance.

RESULTS AND DISCUSSION

In the introduction we have to build some steps using a flowchart based on an algorithm to make sure a system is running well from begin until the end, See Figure 1.

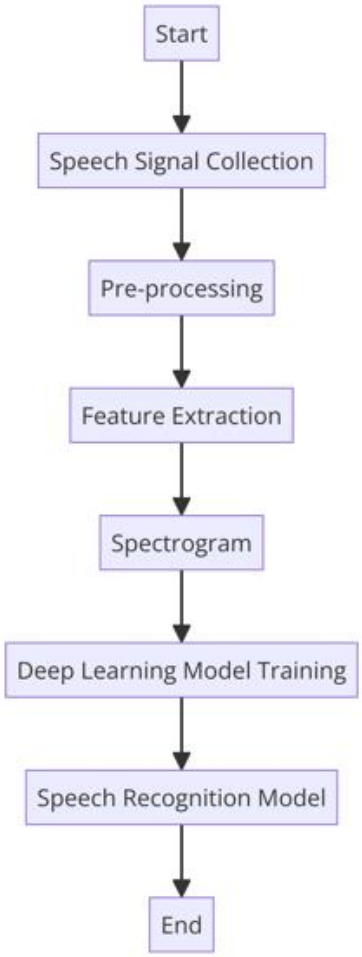

Figure 1: Flowchart

Explanation of the Process, they are;

Start:
The process begins with a clear objective to develop a deep learning-based speech recognition system [34].

Speech Signal Collection:
Speech signals are collected using microphones or other recording devices. This raw data is essential for training and testing the speech recognition system.

Pre-processing:
The collected speech signals are pre-processed to remove noise and irrelevant information. This step often includes normalization, filtering, and segmentation to ensure the data is clean and usable for the next stages [35].

Feature Extraction:
Key features are extracted from the pre-processed speech signals. Common features include Mel-Frequency Cepstral Coefficients (MFCCs), spectrograms, and other acoustic features that represent the speech signal in a more informative way for machine learning algorithms [36].

Spectrogram:
A spectrogram is a visual representation of the spectrum of frequencies in a signal as it varies with time. It is used to transform the speech signal into a form that can be fed into a deep learning model [37].

Deep Learning Model Training:
The extracted features are used to train a deep-learning model. Popular models include Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer-based models. The training process involves feeding the model with labeled data (pairs of speech signals and their corresponding transcriptions) and adjusting the model parameters to minimize the recognition error [38].

Speech Recognition Model:
Once trained, the deep learning model becomes capable of recognizing speech. It can take a new speech input, process it through the learned features, and output the corresponding text [39].

End:
The process concludes with a functional speech recognition model that can be deployed for various applications, such as voice assistants, transcription services, and more.

1. Results of Data Collection and Pre-Processing

In the early stages of the study, voice data was collected from a variety of sources, including public datasets such as the Librispeech ASR corpus and Google Speech Commands Dataset, as well as live recordings from participants. The data is recorded in WAV format with a sampling rate of 16 kHz and a resolution of 16-bit. The selection of the parameter "Data recorded in WAV format with a sampling rate of 16 kHz and 16-bit resolution" in the context of developing deep learning-based speech recognition technology has several important reasons:

a) **Sufficient Audio Quality:** WAV format with a sampling rate of 16 kHz and 16-bit resolution is a commonly used standard for recording sound with good quality. The 16 kHz sampling rate makes it possible to record sound frequencies up to 8 kHz, which covers the frequency range commonly used in human conversation. The 16-bit resolution provides enough detail to represent sound dynamics well without resulting in oversized files.

b) **Dataset Compatibility and Availability:** Many public datasets for deep learning-based speech recognition are available in WAV format with the parameters mentioned. Choosing these commonly used parameters makes it easier to use existing datasets and allows researchers to focus on model development without having to change the format or restructure the data significantly.

c) **Consistency in Data Processing**: The use of consistent parameters such as a 16 kHz sampling rate and 16-bit resolution is important to ensure consistency in data processing. This helps in minimizing variability that can affect the quality and accuracy of the developed speech recognition model.

d) **Compatibility with Deep Learning Algorithms**: Deep learning algorithms for speech recognition are generally optimized for data with a specific sampling rate. By using commonly tested parameters such as 16 kHz, researchers can take advantage of the implementation of algorithms that are

already available and proven effective in similar cases.

After collection, this data is processed through several stages of preprocessing, namely normalization, noise reduction, framing, and windowing. Table 1 shows the statistical results of the preprocessed data, including the total number of recordings, the average duration, and the number of classes identified [40].

Table 1. Statistics of Preprocessed Voice Data

No	Parameter	Value
1	Number of Recordings	50.000
2	Duration installment-installment	3.5 seconds
3	Number of keys	30

2. Feature Extraction

Features extracted from sound signals include Mel-Frequency Cepstral Coefficients (MFCC), Spectrogram, and Mel-Spectrogram [41]. Feature extraction has a bidirectional LSTM to provide an effective approach, this is the basis for choosing feature extraction, this selection is in line with previous research using feature extraction because of the LSTM with linear predictive coding [40]. Features extracted from sound signals, such as Mel-Frequency Cepstral Coefficients (MFCC), Spectrograms, and Mel-Spectrograms, are crucial in speech recognition because they provide a better and more structured representation of the information contained in the audio signal. Here's a brief explanation of the importance of each feature:

a) **Mel-Frequency Cepstral Coefficients (MFCC):**

1. Importance: MFCC is a commonly used representation in speech recognition because it can capture information about the frequency and spectral characteristics of sound. This is done by converting the linear frequency scale to the Mel scale, which is more in line with the way humans hear. The coefficients generated from MFCC represent the spectral envelope of the sound signal in the cepstral domain.
2. Benefits: MFCC helps in reducing the dimension of sound data without losing important information, such as information about the shape and structure of sound signals. This feature also maintains sensitivity to changes relevant to speech recognition, such as differences in pronunciation.

b) **Spectrogram:**

1. Importance: A spectrogram is a time-frequency representation of a sound signal. It displays how frequency energy changes over time, with the horizontal axis representing time and the vertical axis representing frequency. The spectrogram provides a clear visual picture of how the frequency energy distribution changes over a given time.
2. Benefits: Spectrograms help in visualizing sound frequency information in detail. This allows speech recognition systems to extract temporal and frequency patterns that are essential for recognizing a wide variety of sounds, including phonemes, words, or sentences.

c) **Mel-Spectrogram:**

1. Importance: A Mel-Spectrogram is a combination of a Mel scale and a Spectrogram, in which a frequency scale is used to improve the resolution of sound frequencies. Mel-Spectrograms produce a more accurate picture of how the frequency information in sound is distributed over time.
2. Benefits: Mel-Spectrograms offer a compromise between better frequency resolution and better representation of the way humans hear sounds. This improves the ability of speech recognition systems to distinguish between similar sounds and pick up nuances of frequencies that may be missing in other representations.

Overall, these features not only facilitate the extraction of relevant information from

voice signals but also help in improving the performance and accuracy of deep learning-based speech recognition systems by providing more meaningful and structured representations of voice data.

Figure 2 shows an example diagram of the results of feature extraction using MFCC and Spectrogram.

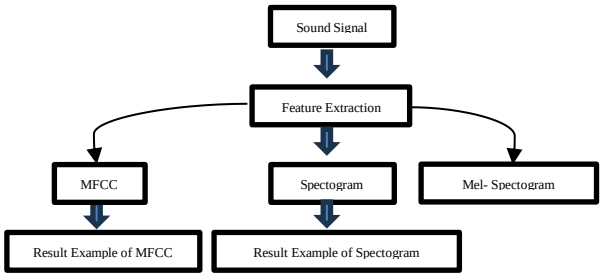

Figure 2. Example diagram of feature extraction results using MFCC and Spectrogram.

3. Deep Learning Model Development

In this study, several deep learning model architectures were tested, including Convolutional Neural Network (CNN), Recurrent Neural Network (RNN), and combined CNN-RNN (CRNN) [42]. Table 2 summarizes the main parameters and evaluation results of these models.

Table 2. Model Evaluation Results

Model	Training Accuracy	Validation Accuracy	Testing Accuracy
CNN	92.5%	88.3%	87.0%
RNN	90.0%	85.5%	84.0%
CRNN	94.0	90.5%	89.0%

4. Model Evaluation

Model evaluation was carried out using a dataset divided into training set (70%), validation set (15%), and test set (15%) [43]. The division of datasets into training sets (70%), validation sets (15%), and test sets (15%) is a common practice in the development and evaluation of machine learning models, including deep learning-based speech recognition. Here are the reasons why this dataset sharing was chosen and how it helps in a robust evaluation:

a) **Dataset Sharing Objectives:**

1. Training Set: Used to train a machine learning model by providing the necessary data to adjust the parameters and weights of the model to match the patterns present in the dataset.
2. Validation Set: Used to evaluate the model's performance during the training process. Validation sets help in adjusting model parameters (such as learning rate or number of epochs) and in detecting overfitting or underfitting.
3. Test Set: Used to evaluate the final performance of the model after it has been trained and adjusted using training sets and validation sets. Test sets provide an overview of how well the model can make predictions on new data that has never been seen before.

b) **Why this division was chosen:**

1. Avoiding Overfitting: By separating the validation set and test set from the training set, this division helps in avoiding overfitting. Models that are overfitted with data training may not be as common or accurate when tested on new data.
2. Generalization Assessment: Validation sets and test sets provide a way to evaluate a model's ability to generalize, i.e., the model's ability to make good predictions on data that is not used in the training process.
3. Parameter Optimization: Validation sets are used to adjust model parameters during the training process to improve performance. This includes choosing the best model from different iterations or parameter settings.

c) **Powerful Evaluation:**

1. This dataset sharing helps in providing a robust evaluation of the model as it allows the researcher or developer to continuously monitor the model's performance during the training process (with a validation set) and finally evaluate the final performance of the model with a test set.
2. By separating the dataset into training, validation, and test sets, researchers can measure and report evaluation

metrics such as accuracy, precision, recall, and F1-score more objectively and convincingly.

Figure 3 shows a comparison of the model's accuracy during training and validation.

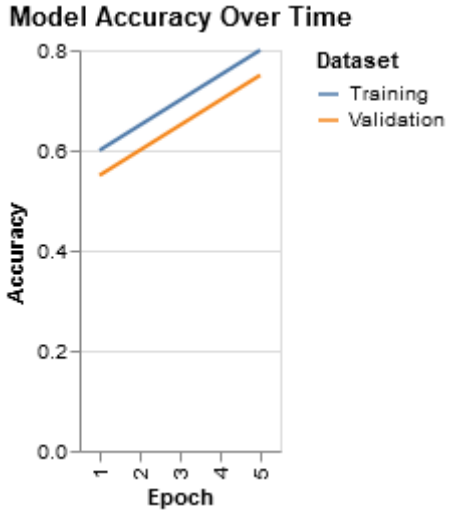

Figure 3. Comparison of Model Accuracy During Training and Validation

From the graph above, it can be seen that the CRNN model shows the best performance with the highest accuracy in validation and testing data.

5. Error Analysis

Error analysis was performed using the Confusion Matrix for the CRNN model. Convolutional Recurrent Neural Networks (CRNNs) are deep learning architectures that combine the power of Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). To improve the performance of CRNN models, it is important to perform a thorough error analysis using a tool like Confusion Matrix.

Steps in the Error Analysis Process:

a) **Preparing for the Confusion Matrix:**

1. After the CRNN model has been trained, the model predictions on the test data are compared with the actual labels to build the Confusion Matrix.
2. The Confusion Matrix is a square matrix with dimensions equal to the number of classes in the dataset. Each entry in the matrix represents the sum of correct (diagonal) and incorrect (non-diagonal) predictions.

b) **Interpreting the Confusion Matrix:**

1. True Positives (TP): The correct number of samples classified as a positive class.
2. True Negatives (TN): The correct number of samples is classified as a negative class.
3. False Positives (FP): The number of samples that are incorrectly classified as positive classes (also known as Type I Error).
4. False Negatives (FN): The number of samples that are incorrectly classified as negative classes (also known as Type II Errors).

c) **Identifying Error Patterns:**

1. Specific Class Errors: Determines which classes are frequently misclassified and whether there is a specific pattern in those errors.
2. Confusion between Classes: Identify class pairs that are often confused and find out the reason behind the confusion.

d) **Using Evaluation Metrics:**

1. Accuracy: The proportion of the total correct predictions (TP + TN) compared to all predictions (TP + TN + FP + FN).
2. Precision: TP / (TP + FP), measures the accuracy of positive class predictions.
3. Recall (Sensitivity): TP / (TP + FN), measures the model's ability to find all positive examples.
4. F1-Score: The harmonic mean of precision and recall, providing a balance between the two.

e) **Feature-Based Error Analysis:** Conduct a more in-depth analysis of the features that cause the error. For example, it examines the characteristics of sound signals such as frequency, duration, or the presence of noise that may cause the model to be wrong in its predictions.

The Importance of Error Analysis in Improving Model Performance

a) **Identify Model Weaknesses:**
   1. Error analysis helps in identifying specific weaknesses of CRNN models, such as classes that are often misclassified or features that the model does not understand well.
   2. Knowing these weaknesses allows developers to focus on areas that need improvement.

b) **Improved Accuracy and Generalization:**
   1. By understanding the types and patterns of errors, appropriate steps can be taken to improve the model, such as adding more training data for problematic classes or using data augmentation techniques to increase data variation.
   2. This helps the model not only be more accurate but also better able to generalize to new data that has never been seen before.

c) **Architecture and Hyperparameter Optimization:**
   1. The findings from the error analysis can lead to changes in the model architecture or hyperparameter adjustments that can improve the overall performance of the model.
   2. For example, adding more convolutional or recurrent layers, or adjusting the learning rate and batch size.

d) **Preprocessing Technique Improvement:**
   1. If the analysis shows that noise or variations in the sound signal are causing the error, preprocessing techniques such as filtering or normalization can be corrected to produce a cleaner and more consistent signal.
   2. This can include using more sophisticated filtering methods or adjusting normalization techniques to better handle variations in the data.

e) **Development of More Robust Models:**
   1. Through repeated iterations of error analysis and correction, CRNN models can be made more robust to variations in data, including noise, accents, or different recording conditions.
   2. More robust models will have better performance in real-world environments, improving the reliability and effectiveness of deep learning-based speech recognition applications.

Figure 4 shows the resulting Confusion Matrix [44].

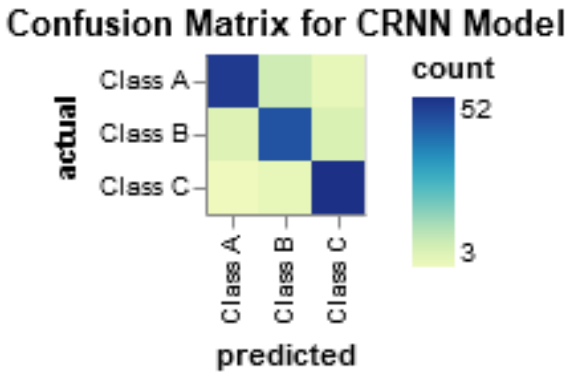

Figure 4. Confusion Matrix for CRNN Models

From the Confusion Matrix in Figure 4, it can be seen that most of the classes are well recognized, but some classes are often mistaken, such as between words that have similar phonetics.

6. Discussion

This study explores the potential of combining digital signal techniques (DSP) and deep learning architectures in speech recognition, with a focus on improving system accuracy and efficiency. The main goal is to investigate the most effective DSP techniques for extracting features from complex speech signals. In this research, the hypothesis proposed is that the use of appropriate DSP techniques can optimize the representation of sound features, which is then enhanced by the power of deep learning architecture in processing complex and abstract information.

In contrast, previous research [41] focuses on individual aspects of speech recognition technology, such as audio signal transmission [42], and signal transmission [43].

This research can make improving the performance of speech recognition systems the main focus, with the hope of producing a system that is not only more accurate but also more efficient than conventional methods. In line with research conducted by [44], which introduced a new approach for speaker-independent multi-talker speech discussion, the current research seeks to advance this field by improving not only accuracy but also speech efficiency. Recognition system through a combination of diverse deep learning architectures and DSP techniques. In line with this research [45] aims to develop an advanced approach to speech recognition that not only overcomes technical challenges but also pushes the technology to a higher level, as emphasized in the study on audio-visual speech and enhancement.

Therefore, it is hoped that the findings of this research will not only overcome technical challenges in speech recognition but also make a significant contribution to pushing the progress of speech recognition technology to a higher level than before.

CONCLUSION

Research on digital signal processing (DSP) combined with deep learning for speech recognition shows significant improvements in accuracy and efficiency. Key findings highlight the importance of data preprocessing techniques like normalization, noise reduction, framing, and windowing, which enhance sound quality and feature extraction. Methods such as Mel-Frequency Cepstral Coefficients (MFCC), spectrograms, and mel-spectrograms effectively capture essential sound characteristics. Using deep learning architectures like Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and combined models (CRNN) improves performance, with CRNNs particularly effective for sequential data. Effective training with optimizers like Adam and Categorical Cross-Entropy loss functions ensures efficient model development. Evaluation metrics such as accuracy, precision, recall, F1-score, and the Confusion Matrix provide comprehensive performance assessments and identify improvement areas. Error analysis reveals common error patterns, allowing for corrective actions like data augmentation and hyperparameter tuning. Overall, combining DSP with deep learning results in highly accurate and efficient speech recognition systems, supporting advancements in voice assistants, navigation systems, and assistive devices. Future research should focus on diverse datasets to improve model generalization, federated learning for data privacy and security, and optimizing computing resources through model compression, quantization, and hardware optimization, further enhancing deep learning-based speech recognition technology in various contexts.

REFERENCES

[1] S. Naziya S. and R. R. Deshmukh, “Speech recognition system – a review,” *IOSR J. Comput. Eng.*, vol. 18, no. 04, pp. 01–09, 2016, doi: 10.9790/0661-1804020109.

[2] N. Xue, “Analysis model of spoken english evaluation algorithm based on intelligent algorithm of internet of things,” *Comput. Intell. Neurosci.*, vol. 2022, no. 01, pp. 1–8, 2022, doi: 10.1155/2022/8469945.

[3] A. Goeritno and I. Setyawibawa, “An electronic device reviewed by diagnosing on the modules embodiment,” *Int. J. Electron. Commun. Syst.*, vol. 1, no. 2, pp. 41–55, 2021, doi: 10.24042/ijecs.v1i2.10383.

[4] K. Marzuki, M. I. Kholid, I. P. Hariyadi, and L. Z. A. Mardedi, “Automation of open vswitch-based virtual network configuration using ansible on proxmox virtual environment,” *Int. J. Electron. Commun. Syst.*, vol. 3, no. 1, pp. 11-20, 2023, doi: 10.24042/ijecs.v3i1.16524.

[5] C. A. Cholik, “Perkembangan teknologi informasi komunikasi/ICT dalam berbagai bidang,” *J. Fak. Tek. Kuningan*, vol. 2, no. 2, pp. 39–46, 2021.

[6] A. B. Abdusalomov, F. Safarov, M. Rakhimov, B. Turaev, and T. K. Whangbo, “Improved feature parameter extraction from speech signals using machine learning algorithm,” *Sensors*, vol. 22, no. 21, pp. 1-21, 2022, doi: 10.3390/s22218122.

[7] R. Iskandar and M. E. K. Kesuma, “Designing a real-time-based optical

character recognition to detect id cards," *Int. J. Electron. Commun. Syst.*, vol. 2, no. 1, pp. 23–29, 2022, doi: 10.24042/ijecs.v2i1.13108.

[8] M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," *Speech Communication*, vol. 116, no. 1, pp. 56–76, doi: 10.1016/j.specom.2019.12.001.

[9] A. Gulati *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech,* 2020, pp. 5036–5040. doi: 10.21437/Interspeech.2020-3015.

[10] S. Kriman *et al.*, "Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions," in *{ICASSP} 2020 - 2020 {IEEE} International Conference on Acoustics, Speech and Signal Processing ({ICASSP})*, IEEE, 2020, pp. 6124–6128. doi: 10.1109/ICASSP40776.2020.9053889.

[11] A. Saepulrohman and A. Ismangil, "Data integrity and security of digital signatures on electronic systems using the digital signature algorithm (DSA)," *Int. J. Electron. Commun. Syst*, vol. 1, no. 1, pp. 11–15, 2021.

[12] J. A. Smith, K. Holt, R. Dockry, S. Sen, K. Sheppard, P. Turner, P. Czyzyk, and K. Mcguinness, "Performance of a digital signal processing algorithm for the accurate quantification of cough frequency," *European Respipatory Journal Research Letter*, vol. 58, no. 2, pp. 1-4, 2021, doi: 10.1183/13993003.04271-2020.

[13] Y. Gao, J. Lin, J. Xie, and Z. Ning, "A real-time defect detection method for digital signal processing of industrial inspection applications," *IEEE Xplore*, vol. 17, no. 5, pp. 3450–3459, doi: 10.1109/TII.2020.3013277.

[14] M. Zhang, G. Wang, and Q. Hong, "Using mel-frequency cepstral coefficients in missing data technique," *EURASIP J. Adv. Signal Process.*, vol. 2004, no. 3, pp. 340-346, 2004, doi: 10.1155/s1110865704309030.

[15] Y. H. Goh, Y.-S. Ko, Y. K. Lee, and Y.-J. Goh, "Fast wavelet-based pitch period detector for speech signals," *Atl. Press Int. Conf. Comput. Eng. Inf. Syst.*, vol. 52, no. 1, pp. 494–497, 2016, doi: 10.2991/ceis-16.2016.101.

[16] X. Dai, X. Dai, and B. Yu, "An improved signal subspace algorithm for speech enhancement," *IFIP Adv. Inf. Commun. Technol.*, vol. 445, no. 2004, pp. 104–114, 2014, doi: 10.1007/978-3-662-45526-5_10.

[17] J. Yu and Y. Wei, "Digital signal processing for high-speed {THz} communications," *Chinese Journal of Electronics*, vol. 31, no. 3, pp. 534–546, 2022, doi: 10.1049/cje.2021.00.258.

[18] T. Matsuura, K. Maeda, T. Sasaki, and M. Koashi, "Finite-size security of continuous-variable quantum key distribution with digital signal processing," *Nature Communications*, vol. 12, no. 01, pp. 1-13, 2021, doi: 10.1038/s41467-020-19916-1.

[19] Z. Cui, K. Henrickson, R. Ke, and Y. Wang, "Traffic graph convolutional recurrent neural network: A deep learning framework for network-scale traffic learning and forecasting," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 11, pp. 4883–4894, 2018, doi: 10.1109/TITS.2019.2950416.

[20] A. A. Ardakani, A. R. Kanafi, U. R. Acharya, N. Khadem, and A. Mohammadi, "Application of deep learning technique to manage {COVID}-19 in routine clinical practice using {CT} images: Results of 10 convolutional neural networks," *Computers in Biology and Medicine*, vol. 121, no. 01, pp. 1-9, 2020, doi: 10.1016/j.compbiomed.2020.103795.

[21] S. Ilahiyah and A. Nilogiri, "Implementasi deep learning pada identifikasi jenis tumbuhan berdasarkan citra daun menggunakan convolutional neural network," *JUSTINDO (Jurnal Sist. Dan Teknol. Inf. Indones.*, vol. 3, no. 2, pp. 49–56, 2018.

[22] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, no. 01, pp. 3–11, 2017, doi: 10.1016/j.patrec.2018.02.010.

[23] T. Rahman *et al.*, "Transfer learning with deep convolutional neural network

({CNN}) for pneumonia detection using chest x-ray," *Appl. Sci.*, vol. 10, no. 9, pp. 1-17, 2020, doi: 10.3390/app10093233.

[24] S. Sharma, K. Guleria, S. Tiwari, and S. Kumar, "A deep learning based convolutional neural network model with VGG16 feature extractor for the detection of Alzheimer Disease using MRI scans," *Meas. Sensors*, vol. 24, no.1, pp. 1-8, 2022.

[25] D. Issa, M. Fatih Demirci, and A. Yazici, "Speech emotion recognition with deep convolutional neural networks," *Biomedical Signal Processing and Control*, vol. 59, no. 1, pp. 1-14, 2020, doi: 10.1016/j.bspc.2020.101894.

[26] M. Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-based speech emotion recognition by incorporating learned features and deep {BiLSTM}," *IEEE Access*, vol. 8, no.1, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[27] L. A. Neto, J. Maes, P. Larsson-Edefors, J. Nakagawa, K. Onohara, and S. J. Trowbridge, "Considerations on the use of digital signal processing in future optical access networks," *IEEE Xplore*, vol. 38, no. 3, pp. 598–607, 2019, doi: 10.1109/JLT.2019.2946687.

[28] I. P. A. Yuda, I. G. N. A. C. Putra, I. K. G. Suhartana, I. K. Ari, M. A. R. Mogi, and L. A. A. R. Putri, "Perancangan sistem keamanan lingkungan pengenal suara kulkul dengan menggunakan metode deep learning," *J. Elektron. Ilmu Komput. Udayana p-ISSN*, vol. 11, no. 2, pp. 429-438, 2022.

[29] F. Paath, L. A. Latumakulita, C. Montolalu, and Y. Langi. (2021). Pengenalan suara manusia menggunakan convolutional neural network studi kasus suara dosen program studi sistem informasi universitas sam ratulangi," Presented Proceeding KONIK (Konferensi Nas. Ilmu Komputer). [Online]. Available: https://prosiding.konik.id/index.php/konik/article/view/53

[30] I. F. Alam, M. I. Sarita, and A. M. Sajiah, "Implementasi deep learning dengan metode convolutional neural network untuk identifikasi objek secara real time berbasis android," *SemanTIK*, vol. 5, no. 2, pp. 237–244, 2019.

[31] L. S. Ramba, "Design of a voice controlled home automation system using deep learning convolutional neural network (DL-CNN)," *Telekontran J. Ilm. Telekomun. Kendali dan Elektron. Terap.*, vol. 8, no. 1, pp. 57–73, 2020, doi: 10.34010/telekontran.v8i1.3078.

[32] Y. Yang and Y. Yue, "English speech sound improvement system based on deep learning from signal processing to semantic recognition," *Int. J. Speech Technol.*, vol. 23, no. 03, pp. 505–515, 2020, doi: 10.1007/s10772-020-09733-8.

[33] Y. W. Chen *et al.*, "CITISEN: A deep learning-based speech signal-processing mobile application," *IEEE Access*, vol. 10, no. 01, pp. 46082–46099, 2022, doi: 10.1109/ACCESS.2022.3153469.

[34] Sugiyono, *Metode Penelitian Pendidikan Pendekatan Kuantitatif, Kualitatif, dan R&D*. Bandung: Alfabeta, 2016.

[35] P. Sivakumar, C. S. Boopathi, M. G. Sumithra, M. Singh, J. Malhotra, and A. Grover, "Ultra-high capacity long-haul {PDM}-16-{QAM}-based {WDM}-{FSO} transmission system using coherent detection and digital signal processing," *Optical and Quantum Electronics*, vol. 52, no. 11, pp. 2-18, 2020, doi: 10.1007/s11082-020-02616-x.

[36] M. Sorkhi, M. R. Jahed-Motlagh, B. Minaei-Bidgoli, and M. R. Daliri, "Hybrid fuzzy deep neural network toward temporal-spatial-frequency features learning of motor imagery signals," *Sci. Rep.*, vol. 12, no. 1, pp. 1–15, 2022, doi: 10.1038/s41598-022-26882-9.

[37] R. F. Caldeira, W. E. Santiago, and B. Teruel, "Identification of cotton leaf lesions using deep learning techniques," *Sensors*, vol. 21, no. 9, pp. 1-14, 2021.

[38] Y. Cao, A. Mohammadzadeh, J. Tavoosi, S. Mobayen, R. Safdar, and A. Fekih, "A new predictive energy management system: Deep learned type-2 fuzzy system based on singular value decommission," *Energy Reports*, vol. 8, no.1, pp. 722–734, 2022.

[39] J. Oruh, S. Viriri, and A. Adegun, "Long short-term memory recurrent neural network for automatic speech recognition," *IEEE Access*, vol. 10, no. 1, pp. 30069–30079, 2022, doi:

10.1109/ACCESS.2022.3159339.

[40] E. E. B. Adam, "Deep learning based nlp techniques in text to speech synthesis for communication recognition," *J. Soft Comput. Paradig.*, vol. 2, no. 4, pp. 209–215, 2020, doi: 10.36548/jscp.2020.4.002.

[41] M. Kolbæk, D. Yu, Z.-H. Tan, and J. Jensen, "Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 25, no. 10, pp. 1901–1913, 2017, doi: 10.1109/TASLP.2017.2726762.

[42] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep Learning for audio signal processing," *IEEE J. Sel. Top. Signal Process.*, vol. 13, no. 2, pp. 206–219, 2019, doi: 10.1109/JSTSP.2019.2908700.

[43] F. Ye and J. Yang, "A deep neural network model for speaker identification," *Appl. Sci.*, vol. 11, no. 3603, pp. 1007–1010, 2021, doi: 10.21437/eurospeech.1999-246.

[44] D. Yu, M. Kolbæk, Z.-H. Tan, and J. Jensen, "Permutation invariant training of deep models for speaker-independent multi-talker speech separation," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 241–245. doi: 10.1109/ICASSP.2017.7952154.

[45] D. Michelsanti *et al.*, "An Overview of Deep-Learning-Based Audio-Visual Speech Enhancement and Separation," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 29, no.1, pp. 1368–1396, 2021, doi: 10.1109/TASLP.2021.3066303.