ADMI IJML
International Journal Multidisciplinary Science
IJML Vol 2 No. 2 June 2023 | ISSN: 2963-8119 (print), ISSN: 2963-7821 (online), Page 26-33

CONCEPTUAL REGIONAL ORIGIN RECOGNITION USING CNN CONVOUTION NEURAL NETWORK ON BANDUNG, BOGOR AND CIREBON REGIONAL ACCENTS

Adam Huda Nugraha[1], Achmad Benny Mutiara[2], Dewi Agushinta Rahayu[3]
Gunadarma University[1,2,3]

Article History
Received : May
Revised : June
Accepted : June
Published : June

Corresponding author*:
adam_huda@staff.gunadarma.ac.id

No. Contact:

DOI:
https://doi.org/10.56127/ijml.v2i2.696

Abstract: Sound detection is a challenge in machine learning due to the noisy nature of signals, and the small amount of (labeled) data that is usually available. The need for sound detection in Indonesia is quite important because there are many community organizations that form groups according to the land of their origin. Especially in big cities, where people from various tribes gather and exchange cultures. However, it has a disadvantage that affects these tribes, namely the loss of the original culture of certain areas. The Sundanese are the object of this research, including Bandung, Bogor and Cirebon. Voice data is divided into 2 types, namely male and female, each region consists of 50 respondents with 25 male and female voices with a maximum voting time of 1 minute. The method used is CNN architecture based on supervised learning, preprocessing using MFCC (Mel Frequency Cepstral Coefficients) to obtain feature extraction from voice data. CNN architecture is carried out 3 times convolution with max pooling and dropout on each convolution.

Keywords: Sound, MFCC, CNN architecture.

INTRODUCTION

Sound detection is becoming a challenge at this point. This is due to the lack of labeled data available (Dufaux et al., 2000; Li et al., 2017). One of the uses of voice detection is to minimize fraud or confessions. In real conditions faced by humans, for one speaker even though he utters the same sound or word, the values of this sound signal will not be exactly the same, the way of pronunciation (long or short in saying the word), the dialect (different language) of the speaker, expressions that produce high-low pressure intonation when speaking and emotions when speaking often change over time (Wolfram, 2017).

The same sound (word) uttered by different speakers also has its own uniqueness (for example, gender and age), because each sound signal is generated through the vocal tract and nasal passages which are different for each human being (Alkhawaldeh, 2019; Ertam, 2019). The higher the sampling rate, the more accurate the signal value is, so that in sound management, the sample rate is generally taken twice from the minimum frequency as specified in the Nyquist frequency (Jurafsy and Martin, 2008). The minimum sampling rate of the human voice is in the range below 4000Hz, as used in telephone lines, so that sound processing with a frequency of 8000Hz is a value with a sufficient sampling rate to represent the human voice (Jurafsky & Martin, 2019).

Feature extraction is converting amplitude signals or spectrograms into only a few coefficient vectors which are thought to contain important information. Voice recognition based on pronunciation can be classified as Isolated Words, Connected Words and Continuous (Legoh & others, 2019). Isolated words are recognizing the sound per word spoken, the recognition system waits for one word to finish speaking then the recognition process begins. Connected words are similar to isolated words, but are able to recognize more than one spoken word, whereas in continuous speech, the system continuously recognizes every word spoken without waiting.

Sound signals can also be mixed with other sound signals (noise) such as the sound of rain, the sound of vehicles, the sound of indoor air conditioners and others, depending on the environment in which they speak. This is what makes speech recognition a research topic that continues to grow today. Speech recognition basically requires voice database (training) and recognition process (classification). Feature extraction is converting amplitude signals or spectrograms into only a few coefficient vectors which are thought to contain important information. This research use Mel-Frequency Cepstrum Coefficient (MFCC) for recognized, takes the frequency coefficient of a voice signal, about 12 coefficient vectors (Deng et al., 2020; Winursito et al., 2018).

This time many organizations grouped from various regions. This is to increase solidarity and love for the homeland of the community. This is a form of caring for fellow citizens who were born in the same region, so that not a few people have several organizations that are similar to the organization in which they were born. Especially in big city areas, where people from various regions and tribes occupy the same area, this causes the mixing of all cultures.

The results of this study make a scientific contribution to the field of speech recognition research in Indonesian because there is still very little research on speech recognition in Indonesian languages. The voice database of the Sundanese tribe collected ± 150 minutes with each person speaking the local language for ± 1 minute.To prevent abuse and fraud in the form of impersonating certain native citizens. This can be detrimental to regional tribes or applicable regulations. With voice recognition coming from 3 regions (Bandung, Bogor and Cirebon), it can help and minimize individuals or people who want to commit fraud.

Materials and Methods

This study aims to obtain a CNN architectural model using Mel Frequency Cepstral Coefficients (MFCC). The MFCC method is used as an extraction of sound features which has a high level of accuracy and fast feature extraction time when compared to other feature extraction methods. For classifiers using supervised-based artificial neural networks such as Learning Vector Quantization (LVQ).

The stages of the research were carried out as shown in Figure 3.1, including the data collection stage, the data pre-processing stage which extracted sound features using a spectrogram or MFCC, the model/algorithm stage, the evaluation stage, the validation stage, and lastly model classification stage.

In this evaluation, we evaluate the performance of the Convolutional Neural Network (CNN) model in predicting classes or labels in data that have never been seen before. There are several evaluation metrics commonly used in this study, including: 1. Accuracy, 2. Precision, 3. Recall 4. F1-score, 5. Confusion Matrix.

Data collection

At the data collection stage, voice data is recorded. Voice data collection is in the form of regional languages of the Sundanese people from 3 regions, namely Bandung, Bogor and Cirebon. The collection was divided into 2 (two) based on gender, namely 25 men in each regional language and 25 women in each regional language with a total of 150 voice data.

The speech recognition system has 2 (two) process steps, the first stage is creating training data, and the second step is the testing step. In preparing the training data, the sound signal is pre-processed, the sound will be normalized and the noise will be removed. The next step after the pre-processing phase is the voice signal will be processed using MFCC to get the features of each voice/audio signal. MFCC will get the coefficient value of each sound signal. After getting the features of each sound signal, it will become a sound vector, and this vector becomes a label.

MFCC is a feature extraction method based on the principle of the character of human hearing. Sound signals are expressed on the MEL scale, a linear filter used for frequencies below 1000Hz and above 1000Hz using a logarithmic filter. The MFCC process block diagram can be seen in Figure 2.

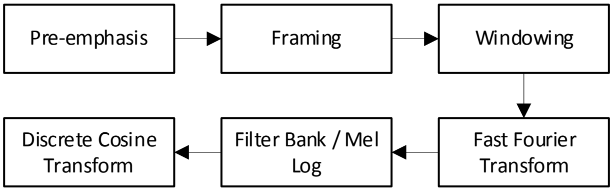

Figure 1. MFCC Process Diagram

Preprocessing

Preprocessing is a data processing technique before entering into a machine learning model. Feature data that has been transformed and collected into an array of each variable x. These variables are carried out in several approaches as follows:

1. Add 1 channel that represents the number of dimensions in the array.
2. Transforming each class/target variable into a multi-class variable.
3. Separating the dataset into training data and test data with a ratio of 80% for training data and 20% for testing data from 150 total data.

The CNN model was initially evaluated using data collection collected from 3 (three) Sundanese tribes in Indonesia. The voices collected are the voices of people who have different Sundanese tribes and genders and then collected in this study. This stage is carried out to generalize the system that has been created with the voices of people with varying voice characteristics. The evaluation process will be carried out as during the testing process with data of 25 male votes and 25 female votes in each tribe. The difference between this stage and the testing stage is that the votes collected have not yet been identified as the votes of 3 (three) Sundanese tribes, but all votes will be carried out at the previous stage. The first test of all data from data collection will have the sound identity of the Sundanese, namely the areas of Bandung, Bogor and Cirebon.

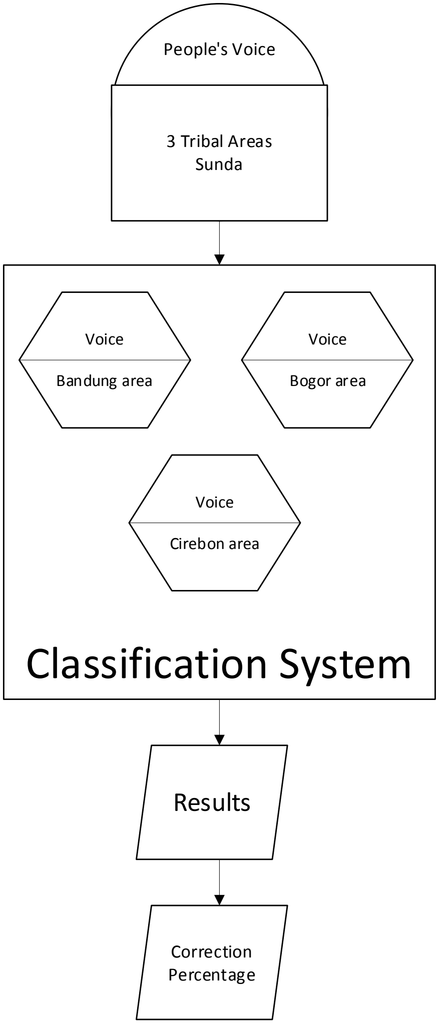

Figure 2. Evaluation Process

RESULTS AND DISCUSSION

Framework of the Proposed Research Method

This research aims to obtain a CNN architecture model using Mel Frequency Cepstral Coefficients (MFCC). The MFCC method is used as an extraction of voice features that have a high level of accuracy and fast feature extraction time when compared to other feature extraction methods. For classifiers using supervised artificial neural networks such as Learning Vector Quantization (LVQ).

Learning Vector Quantization (LVQ) is one of the supervised Artificial Neural Networks, which is a pattern classification method where each output unit represents a certain category or group. The category or group referred to in this study is a regional language with 3 (three) Sundanese tribal areas, identified based on accents and characters from Bandung, Bogor and Cirebon in Indonesia.

The research stages are carried out as shown in Figure 3.1, including the data colection stage, the data preprocessing stage which extracts voice features using spectrogram or MFCC, the model/algorithm stage, the evaluation stage, the validation stage, and finally the model classification stage.

There are several evaluation metrics that will be used in this research, including:

1. Accuracy: is the ratio between the number of correct predictions and the total number of predictions. Accuracy provides information on how often the model is correct in predicting the class on the test dataset.

2. Precision: is the ratio between the number of correct positive predictions and the total number of positive predictions. Precision provides information on how accurate the model is in predicting positive classes.
3. Recall: is the ratio between the number of correct positive predictions and the total number of positive classes in the actual data. Recall provides information on how many positive classes the model can predict.
4. F1-score: is the harmonic mean between precision and recall. F1-score provides information about the balance between precision and recall in predicting classes.
5. Confusion Matrix: a table used to evaluate the performance of a classification model on a test dataset by calculating the number of correct and incorrect predictions for each class in the actual data. Confusion matrix shows the number of correctly and incorrectly classified data in the form of a table consisting of four cells, namely True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN).

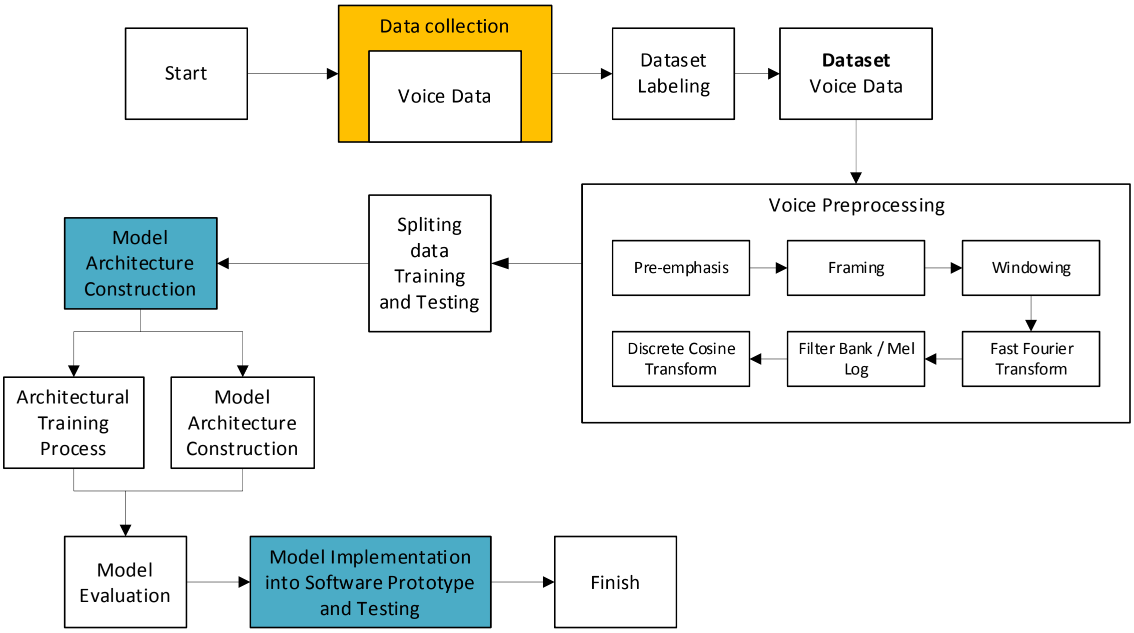
Start
Data collection
Voice Data
Dataset Labeling
Dataset Voice Data
Voice Preprocessing
Pre-emphasis
Framing
Windowing
Discrete Cosine Transform
Filter Bank / Mel Log
Fast Fourier Transform
Spliting data Training and Testing
Model Architecture Construction
Architectural Training Process
Model Architecture Construction
Model Evaluation
Model Implementation into Software Prototype and Testing
Finish

Figure 3. Research method

Data Collection

Data collection from the model that has been made in Chapter 3 is used for the training process and produces several outputs which are discussed and tested according to the method used. Testing of the model is divided into 3 parts to the sounds of the Sundanese from 3 regions. Identification test was conducted to find out whether the model can recognize every sound from each area of the Sundanese tribe. Testing is measured by accuracy, precision, and error. The data collected was 150 votes from the Sundanese in Indonesia, namely Bandung, Bogor and Cirebon. Each region has 50 votes, namely 25 male votes and 25 female votes.

Dataset Labeling

Dataset labeling in Convolutional Neural Network (CNN) is an important stage in the preparation of training data, because it can affect the accuracy and quality of the resulting model. Labeling is done manually by labeling each sound or training data manually with the appropriate label. This requires a lot of time and effort, but it is possible to get accurate labeling. In this research, there are 6 desired labels, namely Bandung Male Voice, Bogor Male Voice, Cirebon Male Voice, Bandung Female Voice, Bogor Female Voice, and Cirebon Female Voice. The Training Data has 120 Voices and the Testing Data has 30 Voices.

Preprocessing

The speech recognition system has 2 (two) process steps, the first step is to create training data, and the second step is the testing step. In preparing the training data, the voice signal is carried out in the pre-processing phase, the voice will be normalized and remove the noise that occurs. The next step after the preprocessing phase is the sound signal will be processed using MFCC to get the features of each sound/audio signal. MFCC will get the coefficient value of each sound signal. After obtaining the features of each sound signal, it will become a sound vector, and this vector becomes a label.

MFCC is a feature extraction method based on the principle of human hearing characteristics. The sound signal is expressed in MEL scale, a linear filter is used for frequencies below 1000Hz and above 1000Hz a logarithmic filter is used. Linear audio is audio that from the beginning of the sound to the end of the sound is uninterrupted without any editing process, adding, deleting and inserting sounds so that converting analog signals into digital is easier.

The following are the stages in the MFCC method carried out in Preprocessing:

1. Sound's Signal
   The first stage is to get the sound signal in digital form. Sound signals usually consist of analog waves which are then converted into digital signals using hardware, for this study the sound signal was recorded using a smartphone.

2. Pre-emphasis
   The second stage is pre-emphasis, which is amplitude amplification at high frequencies in the sound signal to reduce noise and improve the clarity of the sound signal.

3. Framing
   The next stage is framing, which divides the speech signal into several short time frames, usually around 20-40 milliseconds. The purpose is to reduce the complexity of data processing and obtain frequency information within a certain time range.

4. Windowing
   The next stage is windowing, which is giving weight to each of the previously divided time frames, so that the spectral information generated from each time frame can be processed more accurately.

5. FFT
   The next stage is Fast Fourier Transform (FFT), which is the process of transforming signals from the time domain to the frequency domain to obtain spectral information on each time frame.

6. Filter Bank
   The next stage is the filter bank, which is the process of selecting the filters used to extract spectral information at each time frame. The filters used are usually mel-frequency filter banks, which are designed to mimic the frequency response of human hearing.

7. DCT
   The next stage is the Discrete Cosine Transform (DCT), which is a transformation process used to generate cepstral coefficients from the filtered power spectrum. DCT is used because it is more efficient in producing cepstral coefficients and has properties that are suitable for sound signal processing.

8. MFCC Coefficients
   The final step is to calculate the mel-frequency cepstral coefficients (MFCC), which are numerical coefficients that represent features of the speech signal, such as pitch, intonation, and timbre. MFCCs are usually generated by selecting a certain number of coefficients from the DCT results, which are then used as input features in CNN models. MFCC is very useful in speech recognition, speech transcription, and various other speech signal processing applications.

The method is carried out by loading the voice dataset and converting it into a collection of arrays that have previously been transformed into a collection of audio features with a total of 40 features.

Preprocessing is a data processing technique before entering the machine learning model. Data features that have been transformed and collected into a collection of arrays of each x variable. The variable is carried out several approaches as follows:

1. Adding 1 channel that represents the number of dimensions in the array.
2. Transforming each class / target variable into a multi-class variable.
3. Separating the dataset into training data and test data with a ratio of 80% for training data and 20% for testing data from 150 total data.

CONCLUSION

Some conclusions that can be drawn in this study are the results of training on the built model resulting in a loss value of 36.88% indicating that the Convolutional Neural Network model can be applied to identify sounds although it is quite difficult to identify the characteristics of each regional sound because it has an abstract form of data on each image. Classification is based on spatial or spatial image conditions so that it produces an accuracy value of 85% during training. Based on the test results using the confusion matrix method, the recall calculation produces an average value of 100% and precision which also produces an average value of 100%. This value indicates that there is a very good harmony between precision and recall values.

Based on the model used to identify and verify sounds in the Sundanese area of Indonesia, the accuracy value for measuring the performance of the model is high, which is 83.33% in the 86th iteration validation accuracy. Some of the factors that might make this happen are the layers in the model that are not very good at being implemented in the form of abstract data, the filters on each layer are too small or too many, the weights of the pre-trained are not suitable for the type of sound classification in the Sunda region, or the lack of the amount or feature of the data used to train the model.

To identify and verify sound in Sundanese areas in Indonesia, data collection is carried out for 60 seconds per person. In analyzing also adjust to the provision of time in accordance with data collection. However, in implementing the test on the web application for testing, it is permissible to retrieve voice data only once the pronunciation of the sentence is given according to the information from the website.

Model Implementation into a Web Server

In implementing the CNN model into a web server, there are several steps that must be taken to ensure the model can run well and provide accurate voice prediction results. Implementation of the CNN model into a web server can be done with the following steps:

1. Model training
   The first step is to train the CNN model to predict different voice labels. You can use the right training dataset to train your model. Once the training is complete, the CNN model should be saved into a format that is compatible with the web server.
2. Creating an API endpoint
   After training the CNN model, you need to create an API endpoint on your web server. The API endpoint will receive HTTP requests and will provide responses in the form of prediction results from the CNN model. In this research, we use the Python programming language to create the API endpoint.
3. Model integration into the API endpoint
   The next step is to integrate the CNN model into the API endpoint. You need to import the trained CNN model into your API endpoint code. After that, you can use the CNN model to process requests and provide voice predictions.
4. Test the API endpoint
   Once the CNN model is integrated into the API endpoint, you need to test the API endpoint to ensure that everything is working properly. You can use API testing tools like Postman or cURL to send HTTP requests to your API endpoint and check the results.
5. Integrate the API endpoint into the web application
   Once the API endpoint is working properly, you can integrate it into your web application. This can be done by calling the API endpoint from JavaScript code or other code in your web application.

By using the stages above, implementing the CNN model into a web server and creating a web application so that it can recognize voices with high accuracy. By using this web application, it is hoped that this research will be useful for the purposes of voice recognition in the Sundanese tribe for the 3 specified regions.

REFERENCES

[1] Aggarwal, A, Sahay, T,, dan Chandra, M 2015, Performance Evaluation of Artificial Neural Networks for Isolated Hindi Digit Recognition with LPC And MFCC, International Conference on Advanced Computing and Communication Systems, 2015, pages 1-6, IEEE.

[2] Al-Haddad, S, A, R,, Samad, S, A,, Hussain, A,, Ishak, K, A, dan Mirvaziri, H 2007, Decision Fusion for Isolated Malay Digit Recognition Using Dynamic Time Warping (DTW) And Hidden Markov Model (HMM), SCORED 2007, 5th Student Conference on Research and Development, pages 1-6, IEEE

[3] Ali, H,, Jianwei, A, dan Iqbal, K 2015, Automatic Speech Recognition of Urdu Digits with Optimal Classification Approach, International Journal of Computer Applications, 118(9)

[4] Alkhawaldeh, R. S. (2019). DGR: gender recognition of human speech using one-dimensional conventional neural network. Scientific Programming, 2019.

[5] Anna, N & Santoso, CL 1997, Pendidikan anak, edk 5, Family Press, Jakarta.

[6] Azis, A., Wardhono, W. S., & Afirianto, T., 2020, Pengembangan Media Pembelajaran Holografis (Studi Kasus: Bab Indera Pendengaran Manusia),Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer e-ISSN, 2548, 964X.

[7] Chapaneri, S, V, dan Jayaswal, D, J 2013, Efficient Speech Recognition System for Isolated Digits, International Journal Computer Science and Engineering Technologies, 4(3):228–236

[8] Chavan, M, R, S, & Sable, G, S 2013, An Overview of Speech Recognition Using HMM, International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE), 2(6):233,238.

[9] Chu, W, C 2003, Speech Coding Algorithms: Foundation and Evolution of Standardized Coders, A John Wiley & Sons Inc.

[10] Computer Graphics Inter-Facing 1996, 3rd edn, Modern technology Corporation, Minnepolis.

[11] Conley, D 2002, The daily miracle: an introduction to journalism, Oxford University Press, New York.

[12] Cucu, H,, Caranica, A,, Buzo, A, dan Burileanu, C 2015, On Transcribing Informally-Pronounced Numbers In Romanian Speech, 38th International Conference on Telecommunications and Signal Processing (TSP) 2015, pages 372–376, IEEE.

[13] Darabkh, K, A, Khalifeh, A, F,, Bathech, B, A,, dan Sabah, S, W 2013, Efficient DTW-Based Speech Recognition System for Isolated Words of Arabic Language, Proceedings of International Conference on Electrical and Computer Systems Engineering (ICECSE 2013), pages 689–692, Citeseer

[14] Davis, S, B, dan Mermelstein, P 1990, Comparison of Parametric Representations for Monosyllabic Word Recognition In Continuously Spoken Sentences, Readings in Speech Recognition, pages 65–74, Elsevier.

[15] Deng, M. et al. (2020). Heart sound classification based on improved MFCC features and convolutional recurrent neural networks. Neural Networks, 130, 22–32.

[16] Dewi, I, N, Firdausillah, F,, dan Supriyanto, C 2013, Sphinx-4 Indonesian Isolated Digit Speech Recognition, Journal of Theoretical & Applied Information Technology, 53(1).

[17] Dhandhania, V, Hansen, J, K,, Kandi, S, J, dan Ramesh, A 2012, A Robust Speaker Independent Speech Recognizer for Isolated Hindi Digits, International Journal of Computer and Communication Engineering, 1(4):483.

[18] Dixit, A,, Vidwans, A,, dan Sharma, P 2016, Improved MFCC And LPC Algorithm for Bundelkhandi Isolated Digit Speech Recognition, International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pages 3755–3759, IEEE

[19] Dufaux, A. et al. (2000). Automatic sound detection and recognition for noisy environment. 2000 10th European Signal Processing Conference, 1–4.

[20] Ertam, F. (2019). An effective gender recognition approach using voice data via deeper LSTM networks. Applied Acoustics, 156, 351–358.

[21] Graves, A,, Mohamed, A, R,, dan Hinton, G 2013, Speech Recognition with Deep Recurrent Neural Networks, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 6645–6649, IEEE,

[22] Gulić, M,, Lučanin, D,, dan Šimić, A 2011, A Digit And Spelling Speech Recognition System for The Croatian Language, Proceedings of the 34th International Convention MIPRO, pages 1673–1678, IEEE,

[23] Hachkar, Z,, Farchi, A,, Mounir, B,, dan El-Abbadi, J 2011, A Comparison Of DHMM And DTW for Isolated Digits Recognition System of Arabic Language International Journal on Computer Science and Engineering, 3(3):1002–1008,

[24] Hochreiter, S,, dan Schmidhuber, J, 1997, Long Short-Term Memory, Neural Computation, 9(8):1735–1780,

[25] Jurafsky, D,, dan Martin, J, H 2008, Speech and Language Processing (Prentice Hall Series in Artificial Intelligence), Prentice Hall,

[26] Jurafsky, D., & Martin, J. H. (2019). Vector semantics and embeddings. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, 270–285.

[27] Kotler, P, Adam, S, Brown, L & Armstrong, G 2003, Principles of marketing, 2nd edn, Pearson

[28] Lamere, P,, Kwok, P,, Gouvea, E,, Raj, B,, Singh, R,, Walker, W,, Warmuth, M,, dan Wolf, P, 2003, The CMU Sphinx-4 Speech Recognition System, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2003), Hong Kong, volume-1, pages 2–5,

[29] Legoh, K., & others. (2019). Speaker Independent Speech Recognition System for Paite Language using C\# and Sql database in Visual Studio. 2019 2nd International Conference on Innovations in Electronics, Signal Processing and Communication (IESC), 34–38.

[30] Li, J. et al. (2017). A comparison of deep learning methods for environmental sound detection. 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 126–130.

[31] Limkar, M,, Rao, R,, dan Sagvekar, V, 2012, Isolated Digit Recognition Using MFCC And DTW, Mumbai University, India, 1:59–64,

[32] Lina, Qolbiyatul. 2019, Apa itu Convolutional Neural Network? https://medium.com/@16611110/apa-itu-convolutional-neural-network-836f70b193a4 (diakses 14 September 2021).

[33] McLoughlin, I. 2009, Applied Speech and Audio Processing: with Matlab Examples, Cambridge University Press.

[34] Mukhedkar, A, S,, dan Alex, J, S, R, 2014, Robust Feature Extraction Methods for Speech Recognition In Noisy Environments, First International Conference on Networks & Soft Computing (ICNSC), 2014, pages 295–299, IEEE,

[35] Ningthoujam N, dan Prathima V, R 2016, A Survey On Feature Extraction Algorithm for The Speech Recognition System, International Journal of Computer Science and Mobile Computing, 5(4),

[36] Pandit, P,, dan Bhatt, S, 2014, Automatic Speech Recognition of Gujarati Digits Using Dynamic Time Warping, International Journal of Engineering and Innovative Technology, 3(12)

[37] Prakoso, H,, Ferdiana, R,, dan Hartanto, R, 2016, Indonesian Automatic Speech Recognition System Using CMU-Sphinx Toolkit and Limited Dataset, International Symposium on Electronics and Smart Devices (ISESD), pages 283–286, IEEE,

[38] Rabiner, L, R, & Juang, B, H 1986, An Introduction to Hidden Markov Model, IEEE ASSP Magazine 0740-7467/86/0100-0004$01,00©1986 IEEE

[39] Sakoe, H,, dan Chiba, S, 1978, Dynamic Programming Algorithm Optimization for Spoken Word Recognition, IEEE Transactions on Acoustics, Speech, and Signal Processing, 26(1):43–49,

[40] Sakoe, H,, Isotani, R,, Yoshida, K,, Iso, K,, dan Watanabe, T, 1990, Speaker Independent Word Recognition Using Dynamic Programming Neural Networks, Readings in Speech Recognition, pages 439–442, Elsevier,

[41] Saksono, MT,Widyanto, H Achmad & A ZAjub 2008, Aplikasi Pengenalan Ucapan Sebagai Pengatur Mobil Dengan Pengendali Jarak Jauh, http://eprints.undip.ac.id/4310/1/ mar08_t05_ucapan_ayub.pdf.

[42] Silvester D, Gusti S, Budi A, Yaddarabullah A, Wahyu C, Robbi Rahim, 2021, Classification of bird sounds as an early warning method of forest fires using Convolutional Neural Network (CNN) algorithm, Journal of King Saud University–Computer and Information Sciences.

[43] Stevens, K, N, 2000, Acoustic phonetics, volume-30, MIT press,

[44] Terissi, L, D,, dan Gómez, J, C, 2005, Template-Based and HMM-Based Approaches for Isolated Spanish Digit Recognition, Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial, 9(26),

[45] Winursito, A. et al. (2018). Improvement of MFCC feature extraction accuracy using PCA in Indonesian speech recognition. 2018 International Conference on Information and Communications Technology (ICOIACT), 379–383.

[46] Wolfram, W. (2017). Dialect in society. The Handbook of Sociolinguistics, 107–126.

[47] Yi, J,, Ni, H,, Wen, Z,, Liu, B,, dan Tao, J, 2016, CTC Regularized Model Adaptation For Improving LSTM RNN Based Multi-Accent Mandarin Speech Recognition, 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), pages 1–5, IEEE.

[48] Zhichao Z, Shugong X, Shunqing Z, Tianhao Q, Shan Cao, 2020, Attention based convolutional recurrent neural network for environmental sound classification, Neurocomputing, Elsevier.