*Advances in Engineering Research, volume 208*
**Proceedings of the International Conference on Innovation in Science and Technology (ICIST 2020)**

# Accent Recognition Using Mel-Frequency Cepstral Coefficients and Convolutional Neural Network

Dwi Sari Widyowaty*, Andi Sunyoto, Hanif Al Fatta
Magister of Informatics Engineering
Universitas Amikom Yogyakarta
Yogyakarta, Indonesia
*dwi.1234@students.amikom.ac.id

***Abstract*—Everyone has a different accent, the environment and culture can influence the difference in accents. Utilization of the recognition of the speaker's accent can be used as a method to detect the speaker's country of origin. Accent recognition belongs to the field of Automatic Speech Recognition (ASR), research on accent recognition is a step towards a smarter and more sophisticated ASR. Recently, ASR has become a trend in technology, such as virtual assistants. This study aimed to classify accents from several countries, namely English, Arabic, French, Spanish, and Mandarin, where all speakers used the same English Script. Previous research on accent recognition achieved 48.24 % using Mel Frequency Cepstral Coefficients (MFCC) and 2-layer Convolutional Neural Network (CNN). This study increases the accuracy by improving the pre-processing and the CNN model, the methods resulting 51.96 % accuracy using the similar of datasets as the previous study namely 1231 speakers and the methods namely Mel Frequency Cepstral Coefficients (MFCC) and 2-layer Convolutional Neural Network (CNN). By splitting the audio segment at the pre-processing and improving the model of CNN, it turns out to produce better accuracy.**

***Keywords—accent recognition, speech recognition, MFCC, CNN***

## I. INTRODUCTION

English is a widely spoken language in parts of the world, almost all residents in the Country are required to learn English and use English every day. There are about 61 Countries that make English the official language. Each Country has a different accent, for example, the United States and India have different accents even though both speak English. This difference in accent can be influenced by the environment, culture, birthplace, age, and gender [1].

Recently, speech recognition has become a trend, for example in virtual assistants, such as Google assist, Siri, and Alexa. The author has tried to use one virtual assistant, but the virtual assistant cannot recognize the author's accent and the originated country. Therefore, research on accent recognition is one step toward smarter and sophisticated the virtual assistant [2].

Several studies have been conducted on speech recognition, the accent recognition experiment is feature extraction first, then classification. Some audio feature extraction based on frequency domain are MFCC, Spectral Centroid and spread, Spectral Entropy, Spectral Flux, Chroma Vector, and Spectral Roll Off [3]. Several methods have been used, namely SVM, Naïve Bayes, Softmax Regression, GDA, GMM, and K-Means [4]. In the previous studies, CNN has been used in accent research, but the accuracy obtained is low [5]. This paper aims to increase the accuracy of accent recognition and get better accuracy than previous research.

In this paper, our proposed methods are MFCCs and CNN, the output is to classify the Native languages. In this study, the author uses five classes of Native Language, there are Arabic, English, French, Mandarin, and Spanish. The rest of this paper is organized as follows. The detail of related literature is in section 2, our proposed method is described in section 3, the experimental results are shown in section 4, and the last concludes paper in section 5.

## II. RELATED LITERATURE

There have been several existing research on accent recognition, Ma and Fokoué [6] have used two classes of accents, namely the US and Non-US, and the proposed methods was KNN and SVM, those methods achieve high accuracy, KNN 96,05 % and SVM 95,09 %. Another research on two classes of accents is Kamarudin et al. [7], the proposed method is GMM where this research classified Quranic Accent namely Haf Al Asim and Al Kisaie. The proposed method achieves high accuracy 99,3 %.

The next research was Watanaprakornkul [8] where used three classes of accents, namely Cantonese, Hindi, Russian, the proposed methods are MFCC, PLP, SVM and GMM. For MFCC and GMM method achieve 51,47 % accuracy, while MFCC and SVM achieve 48,16 % accuracy.

Two experiments studied on five classes, Bryan et al. [4] and Singh et al. [5], the dataset used is Arabic, English, French, Mandarin, and Spanish. Morgan Bryant et al. concluded that GDA and Naïve Bayes was the best classification were achieved 42 % accuracy. While, Singh et al. concluded that when using the dataset with a segment length in the form of a paragraph, MFCC and CNN were able to get an accuracy of 48.24 %.

## III. PROPOSED METHODS

In our experiment, we used MFCCs as feature extraction and CNN to classify the class of native language. The proposed method is systematically shown in Figure 1.

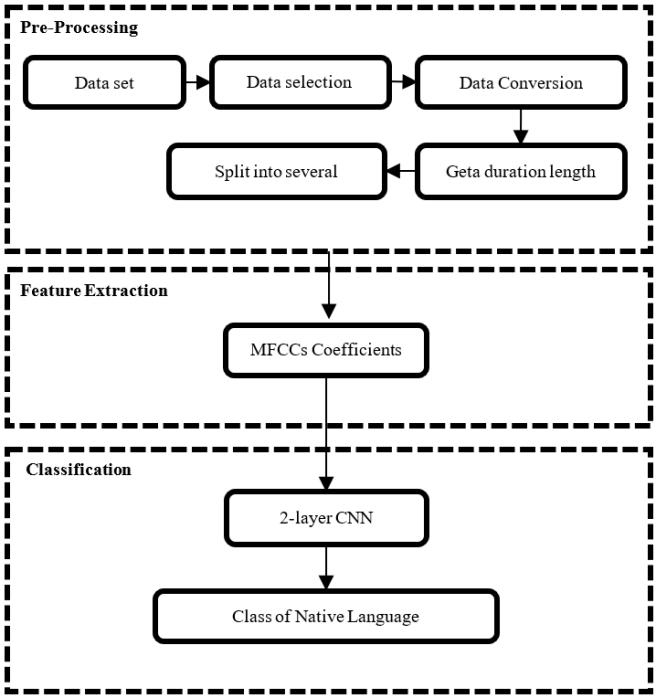



Fig. 1. Block diagram of proposed method.

### A. Dataset

This study uses a public dataset, namely The Speech Accent Archive from George Mason University [9]. Each audio is a recording from a speaker who has difference originated country, all speaker read the same script.

The Speech Accent Archive contains more than 2900 audios which are updated every moment. This dataset has been labelled Native Language, Country, age and gender. Nevertheless, in this study used 1231 audios only where the Native Language classes taken are English, Spanish, Mandarin, French, Arabic. The reason for dataset selection is based on previous research. Details of the dataset can be seen in Figure 2.

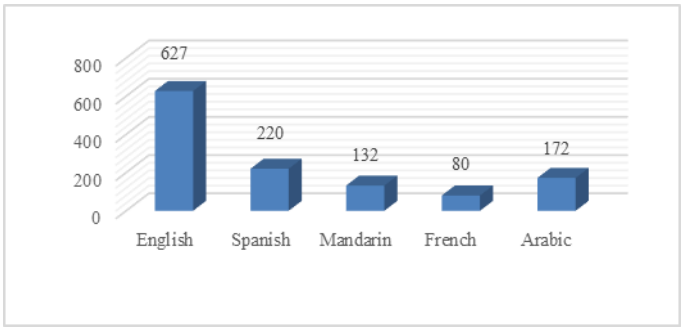



Fig. 2. Dataset selection.

### B. Mel-Frequency Cepstral Coefficients (MFCCs)

Mel-Frequency Cepstral Coefficients (MFCCs) is a very popular feature extraction method in Speech Recognition[3].MFCC is best for speech recognition because MFCC has auditory technique such as human perception concerning to frequency [10,11]. Several steps to extract MFCCs from the audio frame are Pre-Emphasis, Windowing, Fast Fourier Transform (FFT), Mel Filter Bank, Discrete Cosine Transform and Delta Feature [12]. The output of these processes is MFCC coefficients.

### C. Convolutional Neural Network (CNN)

Convolutional Neural Network (CNN) is a multi-layer arrangement consisting of convolutional layers, pooling layer and fully connected layer [13,14]. CNN is not only used in image recognition but also can be used for speech recognition [15,16].

## IV. RESULTS AND DISCUSSION

### A. Pre-processing Audio

In this step, the author observed that each audio has a different length. This study requires the same audio length, so the author takes 10- and 15-seconds audio length to do the research, in this case, we call (a)experiment A for 10 seconds and (b)experiment B for 15 seconds length. Then, each experiment is divided into several segments, A is divided into 10 segments, and B is divided into 15 segments where each segment becomes 1 second. Thus, segments become 12310 for A and segments becomes 18465 for B. The pre-processing step is systematically shown in Figure 3.

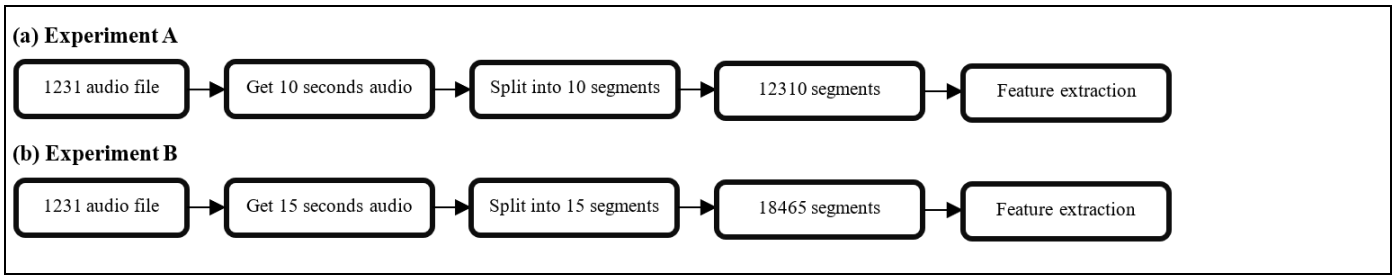



Fig. 3. Block diagram of pre-processing step.

### B. MFCCs Extraction

Each Sound has its characteristics, the spectrogram visualizes the frequency and time domain. Figure 4 displays the spectrogram for the example English Native, rows represent 13 features MFCC, and the colour represents the audible frequency or not, for example, the dark red indicates the highest frequencies that can be heard by humans, in contrast the blue colour indicates frequencies that cannot be heard by a human. In this extraction (MFCCs), the author makes arrangements 13 number MFCCs, 2048 FFT, 512 Hop Length, and 22050 sample rates. This feature extraction used librosa library [17].

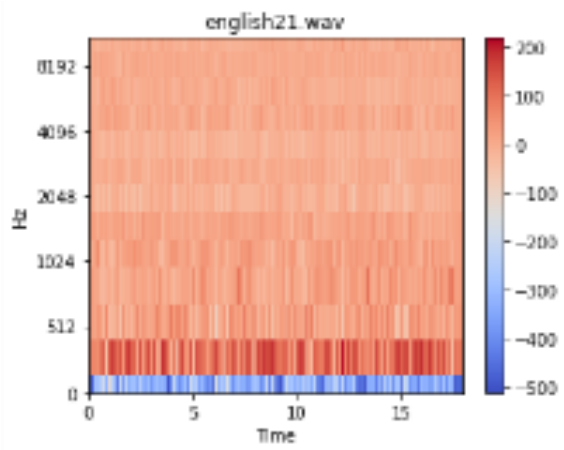



Fig. 4. Display MFCCs for English21.wav.

Generally, CNN is used for image classification [18]. So, in order for the sound dataset to become CNN Input, an array need to be prepared. The CNN input preparation can be seen in Figure 5.

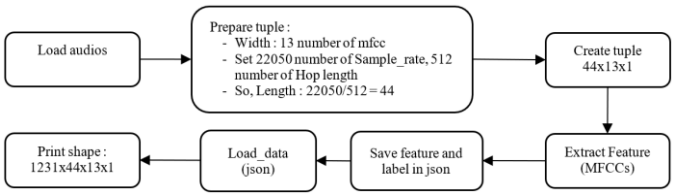



Fig. 5. Preparation for CNN input.

### C. Convolutional Neural Network (CNN)

In Figure 6, displays the model of 2-Layer CNN Model. Size of the input shape is 44x13x1, where 47 is calculated by sample_rate/ hop_length = 22050/512 = 44, while 13 is the number of MFCCs. First, Input shape will be processed by 1-layer CNN. On the first layer, we use 32 filters, kernel_size=3,3 , activation = 'relu' , data_formats= 'channels_last' , MaxPooling2D=2, 2 , strides= 2, 2 , padding='same'. While, on the second layer, we use 64 filters, kernel_size=3,3 , activation = 'relu' , MaxPooling2D=2, 2 , strides= 2, 2 , padding='same' , dropout=0.25.

We got this set from a long experimental process, and to avoid overfitting we added a dropout to the model. The dropout is proven to reduce the possibility of overfitting [19,20]. After completing the convolutional process, the next step is to flatten layer. We use flatten Dense = 128 , activation='relu' , Dropout=0.5. Last, flatten output Dense = 5 , activation='softmax'. Flatten output dense is determined based on the number of dataset classes. We use Keras library to implement this model.

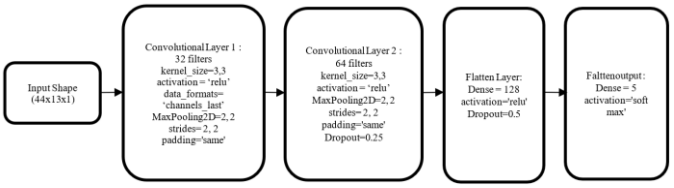



Fig. 6. Block diagram 2-Layer CNN model.

### D. Results

In this experiment, dataset splits are test data 25 % and Validation data 20 % with epoch 30. We found that high epoch did not yield good accuracy, so we took 30 epoch only. The result of this experiment, compared to the previous result can be seen in table 1.

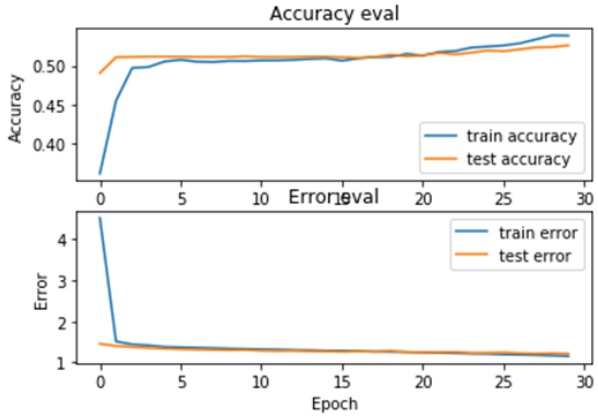
TABLE I. EXPERIMENTAL RESULT COMPARED TO THE PREVIOUS STUDY

| **Previous Study** | | | **Experimental Result** | | | | |
|---|---|---|---|---|---|---|---|
| ***Segment Length*** | ***Test Accuracy*** | ***Loss*** | ***Name of Experiment*** | ***Duration Length*** | ***Segment Split*** | ***Test Accuracy*** | ***Loss*** |
| Paragraph | 0,4824 | 1,3271 | A | 10 seconds | 10 | 0,5075 | 1,2675 |
| | | | **B** | **15 seconds** | **15** | **0,5196** | **1,1628** |

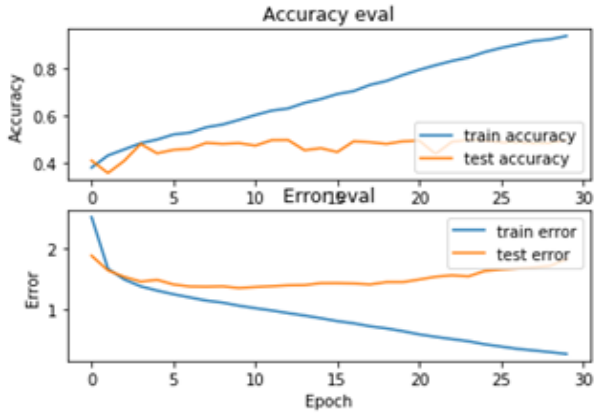
Table 1 explains that Experimental Results achieve the best accuracy, Experiment A achieves 0.5075 accuracy and loss 1.2675. In comparison, Experiment B achieves0,5196 accuracy and loss 1,1628. Table 1 compared the experimental result to the previous study, the previous study achieved 0,4824 accuracy and loss 1,3271. The experiment proves that our methodology has increased the accuracy especially best in experiment B.

Figure 7 shows that the system can reduce overfitting. In Figure 7(a)the audio split into several segments and adding a dropout regularization. While in Figure 7 (b) the author tries to experiment without segment split and dropout, the results

achieve small accuracy and overfitting. So, by doing segments split and adding dropouts, the performance of CNN can be better.




(a)




(b)

Fig. 7. Plot (a) With segment split and dropout (Experiment B) and (b) without segment split and dropout.

## V. CONCLUSION

Accent recognition is included in the scope Automatic Speech Recognition (ASR), by adding accent recognition technology can increase the sophistication of ASR, such as currently trend technology namely virtual assistant.

The author has conducted experiments on a similar number of the dataset as previous studies [5]. In the previous studies, a paragraph has been tested and achieved 0.4824 accuracy and loss 1.3271. While in this paper improved the pre-processing step and model 2-layer CNN. These improvements can produce better accuracy, and this methodology achieved 0.5196 accuracy and loss 1.1628. Besides, improving the 2-layer CNN model by adding a dropout can reduce overfitting. Further potential research, increase the number of CNN layers to 3 or more layers and increase the number of dataset classes to get a better result.

## REFERENCES

[1] B.D. Barkana and A. Patel, “Analysis of Vowel Production in Mandarin/Hindi/American- Accented English for Accent Recognition Systems,” Appl. Acoust., vol. 162, p. 107203, 2020.

[2] D. Honnavalli and Shylaja, “Supervised Machine Learning Model for Accent Recognition in English Speech Using Sequential MFCC Features,” AIDE 2019, 2019.

[3] T. Giannakopoulos and A. Pikrakis, Introduction to Audio Analysis: a Matlab approach. 2014.

[4] M. Bryant, A. Chow, and S. Li, “Classification of Accents of English Speakers by Native Language,” pp. 1–5, 2014.

[5] Y. Singh, A. Pillay, and E. Jembere, “Features of Speech Audio for Deep Learning Accent Recognition,” pp. 4–6, 2019.

[6] Z. Ma and E. Fokoué, “A Comparison of Classifiers in Performing Speaker Accent Recognition Using MFCCs,” Open J. Stat., vol. 04, no. 04, pp. 258–266, 2014.

[7] N. Kamarudin, S.A.R. Al-Haddad, S.J. Hashim, M.A. Nematollahi, and A.R. Hassan, “Feature Extraction Using Spectral Centroid and Mel Frequency Cepstral Coefficient for Quranic Accent Automatic Identification,” 2014 IEEE Student Conf. Res. Dev. SCOReD 2014, pp. 0–5, 2014.

[8] P. Watanaprakornkul, C. Eksombatchai, and P. Chien, “Accent Classification,” Mach. Learn. Final Proj., 2010.

[9] G.M. University, G.M. University, “The Speech Accent Archive.” .

[10] K. Chakraborty, A. Talele, and S. Upadhya, “Voice Recognition Using MFCC Algorithm,” Int. J. Innov. Res. Adv. Eng., vol. 1, no. 10, pp. 158–161, 2014.

[11] 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.

[12] M.A. Imtiaz and G. Raja, “Isolated Word Automatic Speech Recognition (ASR) System Using MFCC, DTW & KNN,” Proc. - APMediaCast 2016, pp. 106–110, 2017.

[13] Y.S. Hariyani, S. Hadiyoso, and T.S. Siadari, “Deteksi Penyakit Covid-19 Berdasarkan Citra X-Ray Menggunakan Deep Residual Network,” ELKOMIKA J. Tek. Energi Elektr. Tek. Telekomun. Tek. Elektron., vol. 8, no. 2, p. 443, 2020.

[14] E.Y. Sari, Kusrini, and A. Sunyoto, “Optimization of Weight Backpropagation with Particle Swarm Optimization for Student Dropout Prediction,” 2019 4th Int. Conf. Inf. Technol. Inf. Syst. Electr. Eng. ICITISEE 2019, vol. 6, pp. 423–428, 2019.

[15] K. Chionh, “Application of Convolutional Neural Networks in Accent Identification.”

[16] K. Choi, D. Joo, and J. Kim, “Kapre: On-GPU Audio Preprocessing Layers for a Quick Implementation of Deep Neural Network Models with KERAS,” arXiv, 2017.

[17] B. Mcfee, C. Raffel, D. Liang, D.P.W. Ellis, M. Mcvicar, E. Battenberg, and O. Nieto, “Librosa - Audio Processing Python Library,” PROC. 14th PYTHON Sci. CONF, no. Scipy, pp. 18–25, 2015.

[18] C.U. Khasanah, E. Utami, and H. Al Fatta, “Pengaruh Dimensi Gambar Pada Data Training Terhadap Prediksi Kepribadian Menggunakan Convolutional Neural Network,” vol. 5, pp. 3–8, 2019.

[19] S. Park and N. Kwak, “Analysis on the Dropout Effect in Convolutional Neural Networks,” vol. 1, no. March 2017, pp. 368–383, 2017.

[20] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: A Simple Way to Prevent Neural Networks from Overfitting,” J. Mach. Learn. Res., vol. 299, no. 3–4, pp. 345–350, 2014.