Identification of Regional Dialects Using Mel Frequency Cepstral Coefficients (MFCCs) and Neural Network

Ayu Mawadda Warohma
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
ayumawaddawarohma@gmail.com

Puspa Kurniasari
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
puspakurniasari@gmail.com

Suci Dwijayanti
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
suci.dwijayanti@gmail.com

Irmawan
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
Irmawan@unsri.ac.id

Bhakti Yudho Suprapto
Department of Electrical Engineering
Faculty of Engineering, Universitas Sriwijaya
Ogan Ilir 30662 South of Sumatera, Indonesia
bhakti@ft.unsri.ac.id

Abstract*— Dialects may affect the voice recognition because they have an influence on the intonation and the pronunciation of a syllable. Every dialect shows the characteristic of a particular tribe and region. This paper addresses to develop a speech recognition system using Mel Frequency Cepstral Coefficients which are fed to a neural network based on backpropagation algorithm to recognize some particular dialects. The regional dialects used in this paper are Indonesian dialects, namely, Bataknese, Minang, and Javanese. The experiments are conducted to recognize the dialects from training and testing data. The testing data used in the experiments are from different people who have never been trained before so that it is expected that the recognition results will be more valid. The accuracy rate is used to evaluate the performance of the system. The accuracy of dialects recognition for Bataknese, Javanese and Minang are less than 20%, 50% - 80% and 70% - 90%, respectively. This method has succeeded in identifying the regional dialects and the most precise identification is on Minang and Javanese dialects.

Keywords : Backpropagation, Javanese dialect, Bataknese dialect, Minang dialect, MFCC, Neural Network.

I. INTRODUCTION

As an archipelagic state, Indonesia has a diversity of languages in each region. It is estimated that there are 700 variations of regional languages used in Indonesia. Most regions have their own local language which is very different from Bahasa Indonesia as an official language and natives tend to use the local language for daily conversation. Hence, the characteristic of local dialects could be heard when they speak Bahasa Indonesia. The dialect may indicate their origin or where they come from. The dialect is considered to be a major obstacle for voice recognition systems because a strong dialect may affect the intonation and pronunciation of syllables[1]. However, there are a few research on dialect identification in Indonesia so that it is still very possible to develop it in the future. Some studies on dialects have been shown in [2] that study of dialect on Bangladeshi, Persian accent[3], and the Tunisian dialect[4]. Rahmawati and Lestari [1] studied local dialects in Indonesia, Javanese and Sundanese. Some methods have been used to recognize such dialects. The spectral features and prosodic features were used as the input of Neural Network for Hindi dialect detection[5], Mel Frequency Cepstral Coefficients (MFCCs) and Support Vector Machine (SVM) were utilized for Pashto dialect[6], Gaussian Mixture Models (GMM) and MFCCs were for Bangladeshi dialect [2]. Results from various research have shown that MFCCs are reliable features to recognize the voice with the accuracy over 80%.

In this paper, the MFCCs are used in a feature extraction stage because of their sensitivity and capability to capture the main characteristic of the speaker's voice, the high accuracy of the voice recognition, and low complexity. A Backpropagation Neural Network algorithm (BPNN) is then utilized to recognize the dialects. The BPNN has adaptive learning capabilities for the training data provided and generalized to new conditions [1]. In BPNN, some statistical parameters are also altered to further enhance the accurate detection of the spoken words. Changing the number of hidden neuron units in each hidden layer of BPNNmay affect the recognition accuracy of the three dialects studied in the research.

Here, three dialects, namely Javanese, Bataknese and Minang, are used to be identified. These dialects are chosen because they have unique characteristics and the speakers may carry them while speaking Bahasa Indonesia. Dialect of Bataknese has a characteristic of a loud and high intonation, and t some emphasis after each sentence. While the Javanese dialect has a tendency to add the sound of [h] and silent sound, and the pronunciation of /b/j/d/ seems like spurted commonly known as “medok”. Then, dialect of Minang is very typical with the pronunciation of the letter /e/ which is thick and spoken in fast tempo. Each of these dialects has a distinctive intonation when they are spoken in some words.

Following this section, the discussion of the paper is divided into several parts. Section 2 explains the theories and methods used in this study. In Section 3, we describe the design and implementation of the proposed method. Section 4 provides the experimental results and the discussion of the paper. In Section 5, the conclusion is presented.

II. Modelling of Dialect Identification

In the initial process of local dialect identification, silence is removed from the speech. Here, we use a speech corpus which are in the form of sentences spoken by different speakers. Those sentences have the same duration of 3 seconds for each speech sample. A feature extraction is conducted to obtain MFCCs with 13 coefficients. We consider utilizing MFCCs because of their ability to compactly represent the speech amplitude spectrum. These MFCCs are then fed to BPNN to recognize the dialects.

A. Silence Removal

The dialect recognition system generally begins with silence removal [3]. In this study, short time energy method is used to determine speech and non-speech parts of the signals. Only the speech parts are then considered to be processed to obtain MFCCs. Frame energy used in this step is 0.01 from the average energy of all waveforms.

B. Feature Extraction using MFCC

Feature Extraction is obtained by several steps as shown in Figure 1.

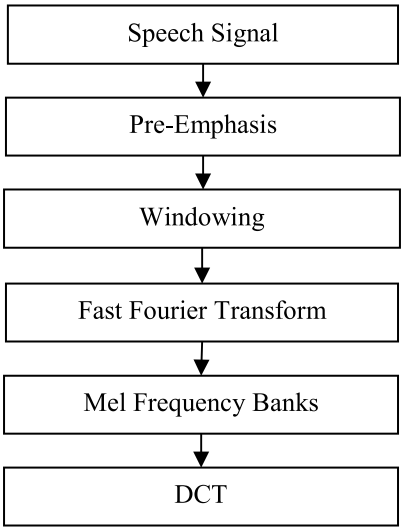

Fig. 1 MFCC Flow Chart

1) Pre-Emphasis

Pre-emphasis aims to allow the signal spectrum to be evenly distributed across all frequencies. It removes noise and maximizes its energy. In this process, the filter suppresses the lower frequency while leaving the higher frequency. Pre-emphasis filter is calculated as follows

$$H(Z) = 1 - \alpha z^{-1} \text{ },0.95 \leq \alpha \leq 1 \qquad (1)$$

Here, we use α of 0.97. The pre-emphasis output with the filter model as in (1) can be written as [6]

$$g(n) - x(n) - \alpha(n - 1) \qquad (2)$$

where *x(n)* is symmetrical window function.

2) Framming

The audio signals change over time and are generally non-linear which make the analysis be difficult. Thus, framing is needed to break down the digital signal into groups of a certain time which has smaller dimensions than the previous signal. To analyze it, the audio signal is divided into short frames with the lengths varying between 20-40ms. Frame size should not be less than 20 ms because it may cause each frame to lose its spectral estimated reliability and its differentiating power may decrease. In this study, the standard frame size for each signal is 25 ms and 16 kHz per second is taken as the sampling frequency. Thus, it has a 60% overlapping window between the frame and audio signal. The sound signal is then blocked into the frame with N sample. The adjacent frames are separated as far as M sample. Each frame has N sample and separated as far as M sample. If M << N, the spectral prediction of each frame becomes better.

3) Windowing

Windowing is utilized to minimize the effect of signal discontinuity at each beginning and end of the frame. It divides the signal into small frames so the analysis in the time domain can be calculated easily. It also reduces the signal to zero at the beginning and end of each frame. The window used for this study is Hamming Window as follows [6]

$$W(n) = \left(0.54 - 0.46\, cos\left[\frac{2\pi n}{N - 1}\right]\right) \qquad (3)$$

N represents the total number of points in the window and n is an integer.

4) Fourier Transform

The next step is to perform a Fourier transform on each sample of the sound signal to measure the response of the signal magnitude. It is necessary to convert the audio signal from the time domain into the frequency domain. Short time Fourier transform can be calculated as follows [6]

$$X_1(k) = \sum_{n=0}^{N-1} x_n e^{-\frac{j2\pi nk}{N}}, \qquad k = 0,1, \ldots.., N - 1 \qquad (4)$$

k represents frequency variable and *N* represent number of discrete time sample

5) Mel Frequency Banks

Frequency resolution adjustment is then performed by a perceptual frequency scale that meets the human ear characteristics in accordance with Mel stage filter bank. However, the frequency limits which can be achieved by the FFT spectrum do not correspond to the linear scale from the phonetic signal. Therefore, the triangle filter bank, which is a filter that has the frequency of response at each triangle-shaped height, is used to overcome this problem [6]. . It is necessary to be used to calculate the values of component weights of the spectral filter so that its response can be converted to Mel scale as in the following equation [6]

$$mel = 2595 * log_{10}\left(1+\frac{f}{700}\right), \qquad (5)$$

where *f* denotes the signal frequency.

6) Discrete Cosine Transform (DCT)

DCT is performed to convert the frequency domain back into the original time domain so that the analysis can be carried out on it. The following equation is to perform the DCT [6]

$$c(n) = \sum_{m-1}^{M} Y(m) \cos\left[\frac{\pi n(m-\frac{1}{2})}{M}\right] \qquad (6)$$

where *C(n)* represents the MFCC, *m* is the number of the coefficients , *N* is the number of triangular bandpass filters, M is the total number of mel-scale cepstral coefficients and *Y(m)* results from spectrum multiplication with conjugate.

Based on the calculation results using DCT, 13 coefficients are obtained from the process of MFCCs characteristic extraction and they are used as input to Neural Network

C. Backpropagation Neural Network Algorithm

A Neural Network has a learning system that is able to recognize a system based on its characteristics. With such learning system, the Neural Network acquires knowledge even though there are disturbance and uncertainty. It can also represent knowledge flexibly by creating self-representation through self-regulation or learning ability (self-organizing). In addition, it is also able to learn from the tolerance of an error and process the knowledge efficiently with its parallel system [3][7].

In this paper, backpropagation which is the most widely used and stable algorithm, is used as the learning algorithm to \overcome problems. The structure of Neural Network can be seen in Figure 2. The structure of the Neural Network consists of input layers, hidden layers and output layers. Here, the number of hidden layers are two and the number of neurons in each layer are altered to find the best learning outcomes.

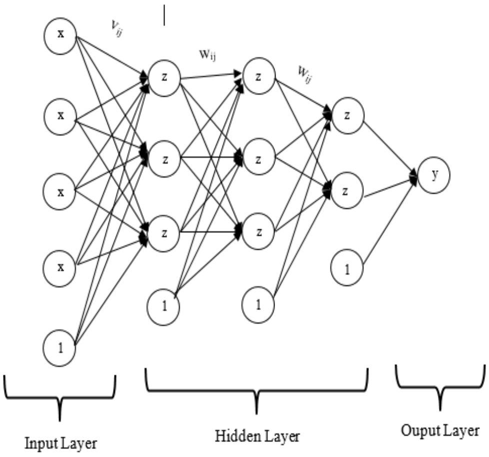

Fig .2 Neural Networks Architecture

III. Design and Implementation

The architecture design for the recognition of dialects being studied is shown in Figure 3. A training process is performed by the neural network followed by a testing process.

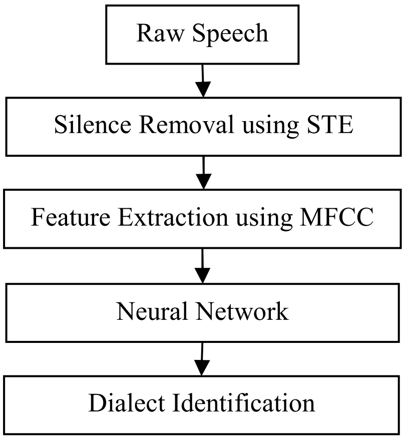

Fig. 3 Block Diagram of Dialect Identification Process

A. Database

Recording was done to obtain the speech signals which represent the dialects of Javanese, Minang, and Bataknese. Some local people with a strong local dialect were chosen as the representation of dialects. From this recording, we obtain database of dialects used in this study. A speech was recorded in a soundproof room and the sampling frequency is of 16 kHz. Since the objective of this study is to identify the regional dialect, all speakers speak the same clause in Bahasa Indonesia. It should be noted that some speakers who move from their place of origin or who do not speak their native language regularly may highly influence their dialect. Therefore, the selection of speakers affects the success rate of the dialect recognition significantly.

From the recording database, we have 120 voice samples which were divided into 40 voice samples of training data for each dialect area. Each dialect consists of 2 women and 2 men who speak the specific clause and repeat it 10 times. The test data are obtained from 2 different people for each dialect since this study also measures the level of accuracy in recognizing the dialects coming from people who are not in the training data.. All selected speakers are at the age of 20-25. In total, there were 180 data which consist of 120 training data and 60 testing data.

B. Feature Extraction

The spectral features used in this study are MFCCs which are fed to the neural network. 13 coefficients of MFCC obtained by reducing the speech signal frequency information into values that mimic the vocal cords in the ear basilar membrane. To obtain these features, the signal is divided into short frames of 25 ms with 10 ms of frame shift and 20 filters bank are used to collect energy information from each frequency band.

C. Parameter of Neural Netwok

The process of classification and the recognition of dialect proposed in this study use backpropagation neural network with 3 hidden layers. These layers are sufficient to handle complex pattern of classification problem. The number of neurons in each hidden layers are altered and the testing process is conducted 4 times with different number of hidden units. In the hidden layer, a log sigmoid activation function is utilized. The function is most commonly used in the classification of a pattern and has range of 0 and 1. The learning rate is 0.01, a large learning rate may affect the system performance and too low learning rate may cause the long duration of the learning process.

D. System Evaluation

To evaluate the performance in identifying local dialects, an accuracy rate is measured. The following standard formula is used to calculate the accuracy ,

$$Accuracy = \frac{\text{number of data known}}{\text{total number of data}} x100\% \qquad (7)$$

Besides the accuracy rate, mean square error (MSE) is also utilized to show the performance of the system.

IV. Experiment Results

MFCCs which are discussed in the previous section are used as the input to the neural network. The neural network us trained by using the BPNN algorithm with the obtained training data. Testing process is then performed to generalize the network by using the test data which are different from the training data.

In this study, four trials (tests) are conducted to evaluate the neural network performance. They were performed by changing the number of neurons in the hidden layer. Meanwhile, the number of input neurons was 13 neurons and the output was 1 neuron. The accuracy rate of each test is shown in Table 1. Table 2 shows the MSE for each test.

TABLE I. Performance of Dialect Recognition Accuracy

No	Dialek	Data	Accuracy			
			Testing 1	Testing 2	Testing 3	Testing 4
1	Bataknese	1	4.2%	6.9%	4.7%	10.5%
		2	8.0%	11.7%	14.3%	15.5%
		3	**8.6%**	**10.1%**	**12.1%**	**16.3%**
2	Javanese	1	52.6%	69.7%	73.4%	79.9%
		2	77.6%	82.4%	83.3%	83.4%
		3	**77.5%**	**70.3%**	**71.9%**	**76.7%**
3	Minang	1	71.2%	73.1%	77.4%	77.0%
		2	89.9%	93.0%	92.9%	93.2%
		3	**79.9%**	**81.4%**	**85.2%**	**83.9%**

This study consists of three experimental data for each dialect. Where the difference between the three experimental data is in the sample of voice used. The first to third data comes from different respondents' voice.

The first test was conducted by using 3 hidden layers with 7 neurons of first hidden layer, 4 neurons of the second hidden layer, and 7 neurons of the third hidden layer. The percentage of recognition accuracy in the first test is low. It may be caused by the number of hidden layer neurons which are small so the network is not able to resolve the problem of pattern recognition. The second trail, the number of neurons on the hidden layer are increased to 25 neurons for the first hidden layer, 15 neurons for the second hidden layer and 10 neurons for the third hidden layer. Thus, the percentage of accuracy improved. Then, in the third trial, the number of neurons in the hidden layer was changed into 26 neurons for the first hidden layer, 52 neurons for the second hidden layer, and 104 neurons on the third hidden layer. As the result, the percentage of the accuracy decreased in some test data. The dialect of Minang can be well recognized in this trial, but Javanese and Bataknese dialects were not yet well recognized. In the last trial, the numbers of neurons in the hidden layer were increased to 50 neurons for the first hidden layer, 75 neurons for the second hidden layer, and 100 neurons for the third hidden layer. In the fourth trial, the level of dialect accuracy increased significantly and an optimum value was obtained in which each dialect can be recognized. These results indicate that increasing the number of hidden layers may affect the accuracy of dialect recognition because the computing process becomes more complex. Thus, this experiment is done by using the number of neurons on the hidden layer as performed in the fourth test.

Table 2 shows the mean squares error (MSE) values of each trials. In the training process, we used 30,000 epochs. This training stops when the value of MSE was convergent and no significant change. The number of neurons in the hidden layer affects the time of the training process. The addition of neurons in the hidden layers make the learning process will be more thoroughly so that the number of error is getting smaller as indicated by the value of MSE at each test. The MSE value is propotional to the level of recognition accuracy.

TABLE II. MSE of Dialect Recognition

No	Dialek	Data	MSE			
			Testing 1	Testing 2	Testing 3	Testing 4
1	Bataknese	1	0.399	0.282	0.251	0.224
		2	0.394	0.293	0.226	0.188
		3	0.442	0.292	0.229	0.214
2	Javanese	1	0.406	0.292	0.238	0.214
		2	0.398	0.293	0.227	0.193
		3	0.418	0.275	0.241	0.200
3	Minang	1	0.396	0.300	0.236	0.195
		2	0.399	0.307	0.222	0.188
		3	0.390	0.290	0.251	0.197

Figure 4 shows the performance of the regional dialect recognition expressed as the percentage of the accuracy level.

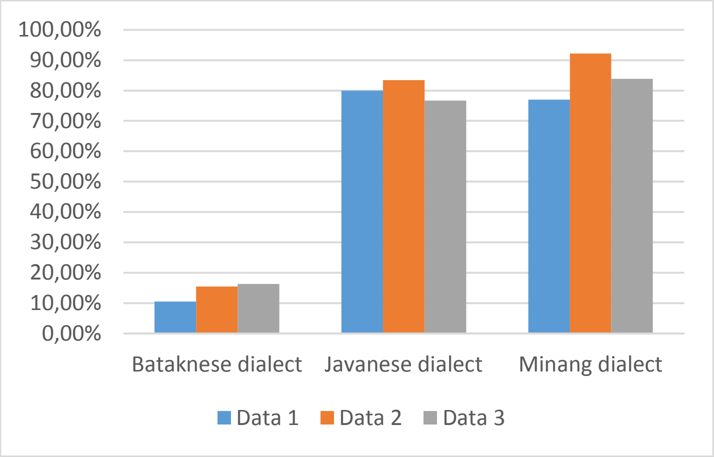

Fig.4 The Comparison of Recognition Accuracy Level

In the Fig. 4, the percentage of Bataknese dialect recognition for the first to the third data is in the range of 4% - 16%. This range of percentage is considered to be very low, it seems the BPNN gets difficulty to recognize this dialect well due to the level of Bataknese dialect variation and the difference of variation between the testing data and training data. It might be caused by some factors such as the dialect of Bataknese in the urban areas is different from the rural areas called as Karo Bataknese and Nias Bataknese dialects. For this reason, we need to consider a specification or uniformity between the test data and training data in the recognition of Bataknese dialect for a particular area.

For the Javanese dialect, the recognition percentage varies by 50% to 83%. As the case of Bataknese dialect, the Javanese dialect is also varied in each regions of Java so that the variation of the recognition becomes different. However, the Javanese dialect has a characteristic called "*medok*" or "*ngapak*" of which pronunciation in several letters such as /d/, /j/, and /b/ is stressed strongly. This typical characteristic makes the algorithm be able to recognize the Javanese dialect although it has variations.

The percentage of Minang dialect is not significantly different from the accuracy percentage of 70% - 90%. The fundamental difference of Minang dialect is in emphasizing intonation on letters like /e/ and spoken in a fast tempo. This characteristic may be recognized well by the BPNN algorithm as Minang dialect.

V. Conclusions

This paper discusses the dialects identification using MFCC and Neural Network. The experiments performed to three dialects in Indonesia, namely, Javanese, Minang and Bataknese dialects. The results show the sequence of the dialect recognition percentage. Javanese and Minang have higher accuracy level than Bataknese. It might be because Javanese and Minang have particular characteristics even though they are varied in each regions. Thus study shows that the characteristic of the dialects may affect the recognition. The success rate of recognition is also affected by the number of neuron. The more complicated recognition model, the bigger number of neuron is needed.

Further study can be explored to compare this proposed method with other recognition methods so that the advantages of each method can be highlighted. In addition, the recognition of a more specific dialect in one region that has a diverse dialect can also be studied. A dialect recognition system that involves gender as the factor of recognition can also be developed as Indonesia has ethnical, linguistic and cultural diversity

References

[1] R. Rahmawati ; D. P. Lestari, "Java and Sunda Dialect Recognition from Indonesian Speech using GMM and I-Vector," in *2017 11th International Conference on Telecommunication Systems Services and Applications (TSSA)* , 2017.

[2] P. P. Das, S. M. Allayear, R. Amin, and Z. Rahman, "Bangladeshi dialect recognition using Mel Frequency Cepstral Coefficient, and Gaussian Mixture Model," in *Proceedings of the 8th International Conference on Advanced Computational Intelligence, ICACI 2016*, 2016, pp. 359–364.

[3] A. Rabiee and S. Setayeshi, "Persian Accents Identification Using an Adaptive Neural Network," *Educ. Technol. Comput. Sci. ETCS 2010 Second Int. Work.*, vol. 1, pp. 7–10, 2010.

[4] M. Hassine, L. Boussaid, and H. Massaoud, "Tunisian dialect recognition based on hybrid techniques," *Int. Arab J. Inf. Technol.*, vol. 15, no. 1, pp. 58–65, 2018.

[5] S. S. A. S. Sinha, A. Jain, "Speech Processing for Hindi Dialect Recognition," in *Advances in Signal Processing and Intelligent Recognition Systems*, pp. 161–169.

[6] S. Khan, H. Ali, and K. Ullah, "Pashto language dialect recognition using mel frequency cepstral coefficient and support vector machines," in *ICIEECT 2017 - International Conference on Innovations in Electrical Engineering and Computational Technologies 2017, Proceedings*, 2017.

[7] J. Gonzalez-Dominguez, I. Lopez-Moreno, P. J. Moreno, and J. Gonzalez-Rodriguez, "Frame-by-frame language identification in short utterances using deep neural networks," *Neural Networks*, vol. 64, pp. 49–58, 2015