Arabic Speech Recognition Using MFCC Feature Extraction and ANN Classification

Elvira Sukma Wahyuni
Department Of Electrical Engineering
Universitas Islam Indonesia
elvira.wahyuni@uii.ac.id

***Abstract*— This research addresses a challenging issue that is to recognize spoken Arabic letters, that are three letters of *hijaiyah* that have indentical pronounciation when pronounced by Indonesian speakers but actually has different makhraj in Arabic, the letters are *sa*, *sya* and *tsa*. The research uses Mel-Frequency Cepstral Coefficients (MFCC) based feature extraction and Artificial Neural Network (ANN) classification method. The result shows the proposed method obtain a good accuracy with an average acuracy is 92.42%, with recognition accuracy each letters (sa, sya, and tsa) prespectivly 92.38%, 93.26% and 91.63%.**

Keywords—voice recognition; signal processing; MFCC featur extraction; ANN classification;

I. INTRODUCTION

The pronunciation letters in Bahasa are very different from the pronunciation letters in Arabic (hijaiyah). Each letter of Hijaiyah has its own *makhraj*. *Makhraj* is the place of the letter out. These differences becomes obstacles for Indonesian speakers who are accustomed to using Bahasa, where the pronunciation of each letter was not paying attention to each *makhraj*. The next obstacle was a few of hijaiyah letters have similar sounds yet in their pronunciation of each sound has a very different. For example, the pronunciation of the letters sa (س), tsa (ث), sya (ش), most of indonesian speakers will pronounce with the same sound as "sa", whereas those letters have each sound differently. The letter of sa (س) sounded by placing tongue tip on two lower incisors, the letter of tsa (ث) sounded by placing tongue tip on two upper incisors, and the letter of sya (ش) sounded from the middle of the tongue and fixed the right palate above it.

Nowdays speech recognition is most interesting research, there are many studies have been done, i.e. speech recognition for isolated Malay digit 0-9 [1], isolated words recognition in Chinese to distinguish words of "ka shi", "ting zhi", "qian jin", "hou tui" and "ji shu" [2]. Isolated digit recognition in english[3]. Words recognition to compare the similarity of letters in words " forward", "Left", "Right", "Reverse", and "Control" [4].

Vowel or speech identification using learning algorithm method has not been able to deliver perfect recognition like a human brain. For example, same speech signal from one speaker then repeated at different times will have different speeds and sampling times, and that would be a problem in speech recognition system. Meanwhile, human brain has ability to identifying the process easily in a short time. The natural characteristic of speech signal is non stationer and *noise* become any other problems in speech recognition system.

Various learning algorithms have been developed and enhanced in speech recognition. In this study Mel-Frequency Cepstral Coefficients (MFCC) based feature extraction and Artificial Neural Network (ANN) classification method is applied to recognize letters of hijaiyah, and the letters are sa (س), tsa (ث), and sya (ش).

The rest of the article is organized as follows. Section 2 drescribes the detail of system model and methodology. The experiments result and evaluations are demonstrated in section 3. Finally, we give the conclusion in section 4.

II. METHODE

A. Data

The data used in this study is a record of human voice that pronounces letters of sa (س), tsa (ث), sya (ش) according to the *makhraj* pronunciation of hijaiyah. From the feature extraction result we have 13 feature and 738 data, where respectively 248 data of sa (س), 254 data of sya (ش), and 236 data of tsa (ث).

B. Feature extraction

MFCC feature extraction is a method that is widely used in speech recognition [5]. Figure 1 shows block diagram of MFCC feature extraction.

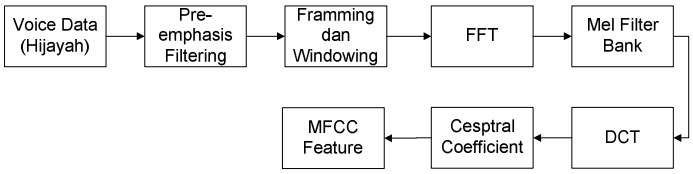

Fig. 1. Block diagram of MFCC feature extraction.

- Pre-emphasis Filtering

Pre-emphasis filtering was applied to increase the high frequency energy and decrease the low frequency energy, simply shown in equation (1).

$$y_t = \alpha x_t + (1-\alpha)x_{t-1} \quad (1)$$

- Framming dan Windowing

Speech signal is non stationary, but usually the speech signal has stationary at a certain time range (20-40 ms), namely short windows or frame. In the framing process speech signal will be divided into several frames and then processed.

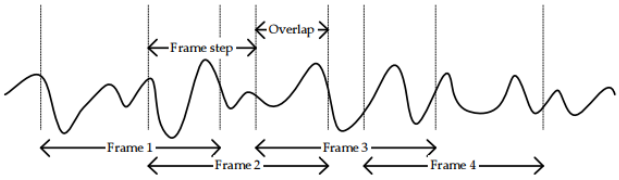

Fig. 2. Ilustration of framing process.

- Fast Fourier Transform (FFT)

At this stage, each frame was converted from time domain form into frequency domain form. FFT is a computational algorithm of Discrete Fourier Transform (DFT).

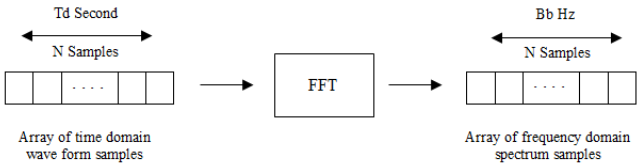

Fig. 3. Ilustration of FFT.

Let *f(x,y)* for *x=0,1,2,. . ., M-1* and *y = 0,1,2,. . ., N*-1 denote a digital image of size *MxN* pixels. The 2-D DFT of *f(x,y)*, denoted by *F(u,v)*, is given by the equation (2).

$$F(u,v) = \sum_{x=0}^{M-1}\sum_{y=0}^{N-1} f(x,y)e^{-j2\pi(ux/M+uy/N)} \quad (2)$$

For *u=0,1,2, . . ., M-1* and *v=0,1,2, . . ., N-1*.

- Mel Filter bank

At this stage bank-filter analysis was used to perform linear predictions. In this study using mel-scale computation provided by HTK, is shown in the equation (3) [6].

$$Mel(f) = 2595\log_{10}(1+\frac{f}{700}) \quad (3)$$

Figure 4 shows an illustration of mel-scale filter bank.

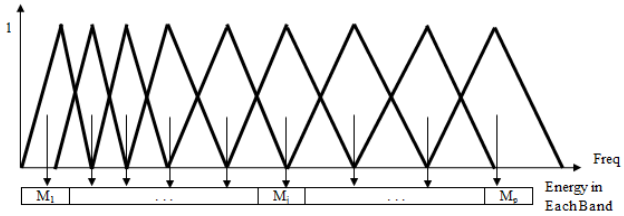

Fig. 4. Mel scale filter bank.

- Descrete Cosine Transform (DCT)

To bring back the signal in time domain form, then at this stage is done by Descrete Cosine Transform (DCT). DCT transforms the cosine component only. DCT 2-D for image *f(x,y)* with size *NxM* can be expressed in equation (4).

$$C(u,v) = \frac{2}{\sqrt{MN}}\alpha(u)\alpha(v)f(x,y)\cos\left[\frac{\pi(2x+1)u}{2N}\right]\cos\left[\frac{\pi(2y+1)u}{2M}\right] \quad (4)$$

For *u=0,1,2, . . ., M-1* and *v=0,1,2, . . ., N-1*.

- Cepstral Coefficient

The final result is 13 Cepstral Coefficient that will be used as a feature in speech recognition.

C. Classification

Figure 5 shows block diagram of ANN classification.

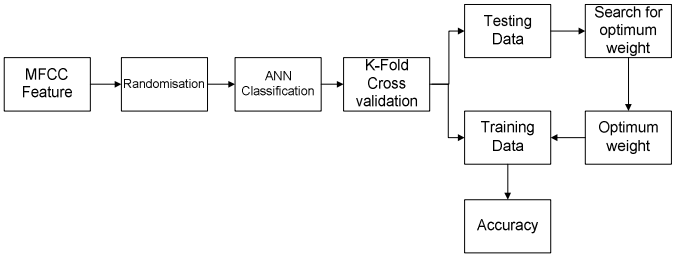

Fig. 5. Block diagram of ANN classification.

At the classification stage, the letter to be recognize was given as class 1 and the others was given as class 0.

- ANN Classification

ANN is classification algorithm that works by adopting the working system of human biological nervous system. In this study ANN model used is back propagation. Back propagation have hidden layer between input layer and output layer to handle ANN vulnerability in pattern recognition. The activation function used in back propagation model is sigmoid function, shown in equation (5).

$$f(x) = \frac{1}{1+e^{-x}} \quad (5)$$

Here are three phases of back propagation training:

Phase 1: Forward propagation

During forward propagation, the input signal (x_i) is propagated to a hidden layer using specified activation function. Then output from the hidden layer (z_j) is propagated to the next hidden layer and so on, until we have network output (y_k). Then the output (y_k) is compared to the target (t_k).

Phase 2: Backward propagation

Based on error (t_k-y_k) calculated the error factor (δ_k). So on for the hidden layer.

Phase 3: weight updating

After all the error factors (δ) are calculated, all of line weights are modified together.

All three phases are repeated until the stopping conditions are met.

- K-Fold Cross Validation

In each classification process was evaluated using 10-fold cross validation. K-fold cross validation was chosen becouse it is more accurate in estimating performance. Cross-validation is an evaluation method that divides data into two segments, one used for learning (training model) and the other is used as a validation (testing model). Stratified random data is divided into *K* partition with the same size, then it will be done iteration as much as K. At each iteration one partition is used for testing and *K*-1 partition for training. The advantage of this method is that every training data has become testing data and vice versa.

- Accuracy

The success rate of the classification is measured by calculating the number of correct classification devided by the total number of classifications, as shown in the equation (6).

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \times 100\% \qquad (6)$$

Where:

TP = True Positive
TN = True Negative
FP = False Positive
FN = False Negative

Each value of TP, TN, FP and FN was obtaned based on the confusion matrix, shown in figure 6.

	Predicted : No	Predicted : Yes
Actual : No	True Negative	False Positive
Actual : Yes	False Negative	True Positive

Fig. 6. confusion matrix

AUC calculates the area under ROC curve. AUC has a value with intermediate range 0.0–1.0, if the AUC value approches 1 then it shows the higher classification accuracy.

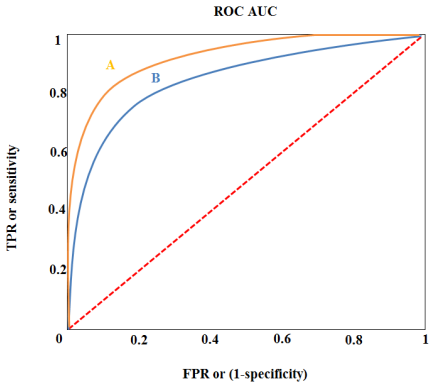

Fig. 7. Ilustratuion of ROC AUC

III. Result and Analisys

Figure 8 Shows speech waveform, filterbank energies and mel frequency ceptrum from speech sample letter of sa (س).

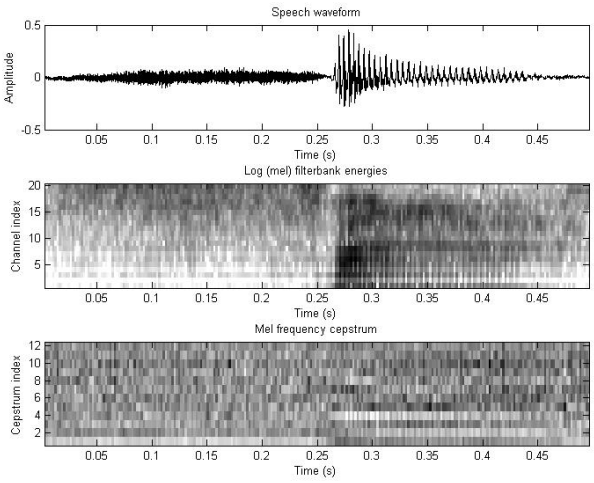

Fig. 8. Speech sample of sa

Figure 9 Shows speech waveform, filterbank energies and mel frequency ceptrum from speech sample letter of sya (ش).

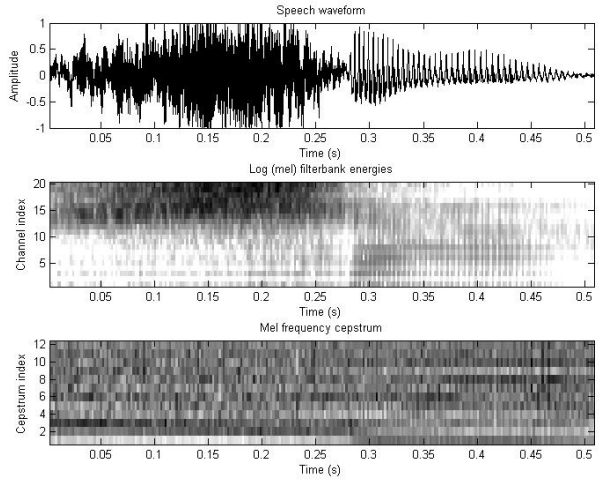

Fig. 9. Speech sample of sya

Figure 10 Shows speech waveform, filterbank energies and mel frequency ceptrum from speech sample letter of tsa (ث).

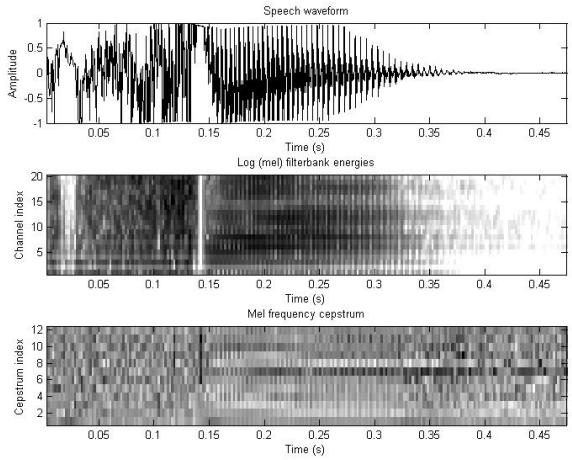

Fig. 10. Speech sample of tsa

Table 1 shows number of TP, TN, FP and FN speech recognation of sa (س), tsa (ث) and sya (ش).

TABLE I. NUMBER TP, TN, FP AND FN

Voice rcognition of	*True Positve*	*True Negative*	*False Positive*	*False Negative*
س	21.90	46.28	2.72	2.90
ش	22.81	46.01	2.39	2.59
ث	20.21	47.42	2.78	3.39
Average	21.64	46.57	2.63	2.96

Table 2 shows percent accuracy, percent incorect, percent unclassified and ROC AUC speech recognition of sa (س), tsa (ث) and sya (ش).

TABLE II. PERCENT ACCURACY, PERCENT INCORECT, PERCENT UNCLASSIFIED AND ROC AUC

Voice rcognition of	*Percent Accuration*	*Percent Incorect*	*Percent Unclassified*	*ROC AUC*
س	92.38	7.62	0.00	0.97
ش	93.26	6.74	0.00	0.97
ث	91.63	8.37	0.00	0.96
Average	92.42	7.57	0.00	0.96

IV. CONCLUSION

The speech recognition of hijaiyah becomes a challenge issue in learning algorithm. Capability of Feature extraction method in producing features that have contributed to speech recognition as well as capability of classification method in learning process are two key to successfull speech recognition. In this study MFCC based feature extraction and ANN classification method is able to make a better recognize with avarage accuracy of 92.42%.

ACKNOWLEDGMENT

The authors would like to thank to departement of Electrical Engineering, Islamic University of Indonesia for supporting this project.

REFERENCES

[1] S. C. Sajjan and C. Vijaya, "Comparison of DTW and HMM for Isolated Word Recognition," *IEEE*, no. 1, pp. 466–470, 2012.

[2] W. Jian, Y. Li, J. Le, and Y. Yang, "Improvement Algorithm of DTW on Isolated-word Recognition," *IEEE*, pp. 1–4, 2011.

[3] I. Technology, "Isolated Digit Recognition Using MFCC AND DTW," *IJAEEE*, no. 1, pp. 59–64, 2012.

[4] B. J. Mohan, R. B. N, and A. R. Module, "Speech Recognition using MFCC and DTW," *IEE Mohan Adv. Electr. Eng. (ICAEE), 2014 Int. Conf. Publ. Year*, pp. 1–4, 2014.

[5] S. Sremath, S. Reza, A. Singh, and R. Wang, "Speaker identification features extraction methods : A systematic review," *Expert Syst. Appl.*, vol. 90, pp. 250–271, 2017.

[6] J. Huang, S. Xiao, Q. Zhou, F. Guo, X. You, H. Li, and B. Li, "A Robust Feature Extraction Algorithm for the Classification of Acoustic Targets in Wild Environments," 2015.