TELKOMNIKA, Vol.15, No.1, March 2017, pp. 292~298
ISSN: 1693-6930, accredited **A** by DIKTI, Decree No: 58/DIKTI/Kep/2013
DOI: 10.12928/TELKOMNIKA.v15i1.3605

Comparison of Feature Extraction MFCC and LPC in Automatic Speech Recognition for Indonesian

Sukmawati Nur Endah*[1], Satriyo Adhy[2], Sutikno[3]
[1,2,3]Informatics Department, Faculty of Science and Mathematics, Universitas Diponegoro
Jl. Prof. Sudharto, S.H. Kampus Tembalang UNDIP, Semarang, Jawa Tengah, Indonesia
Corresponding author, e-mail: sukma_ne@undip.ac.id*[1]; satriyo@undip.ac.id[2]; tik@undip.ac.id[3]

Abstract

Speech recognition can be defined as the process of converting voice signals into the ranks of the word, by applying a specific algorithm that is implemented in a computer program. The research of speech recognition in Indonesia is relatively limited. This paper has studied methods of feature extraction which is the best among the Linear Predictive Coding (LPC) and Mel Frequency Cepstral Coefficients (MFCC) for speech recognition in Indonesian language. This is important because the method can produce a high accuracy for a particular language does not necessarily produce the same accuracy for other languages, considering every language has different characteristics. Thus this research hopefully can help further accelerate the use of automatic speech recognition for Indonesian language. There are two main processes in speech recognition, feature extraction and recognition. The method used for comparison feature extraction in this study is the LPC and MFCC, while the method of recognition using Hidden Markov Model (HMM). The test results showed that the MFCC method is better than LPC in Indonesian language speech recognition.

Keywords: *Mel Frequency Cepstral Coefficients, linier predictive coding, speech recognition*

1. Introduction

Research on speech recognition started in the 1950s [1]. Speech recognition can be defined as the process of converting voice signals into the ranks of the word, by applying a specific algorithm that is implemented in a computer program. It is also often called Speech to Text. The research of speech recognition in Indonesia is relatively limited. An Indonesian-language speech recognition system is still limited to writing a simple message on a mobile device and for search engine in Google; it is far behind English speech recognition systems, of which several have already been applied to related fields [2-4]. For that, Indonesian-language speech recognition needs to be further developed, so it can be used in various fields, as in the automatic detection of infringements in audio broadcast programs [5].

There are two main processes in speech recognition: feature extraction and recognition. Various methods have been developed to produce a high level of accuracy. Feature extraction techniques that have been developed include Linear Predictive Coding (LPC), Cepstral Analysis [6], Mel Frequency Cepstral Coefficients (MFCC) [7], Wavelet Cepstral Coefficients (WCC) [8] and retrieval based prosodic features [9]. These feature extraction methods have already been reviewed by Anusuya [6] for speech in English. Considering the characteristics of each language different from each other, resulting in feature extraction method which is right for the English language is not necessarily appropriate for other languages, especially Indonesian. Selection of appropriate methods of extraction features can help improve the accuracy of recognition. Thus it need to be compared the methods of feature extraction for speech recognition in Indonesian language. Several speech recognition in Indonesian language have also used those extraction method, such as Sakti's [10] and Thiang's [11] research. Feature extraction methods used in Sakti's research is MFCC, whereas in Thiang's research is LPC. But until now no one has studied methods of feature extraction which is best between LPC and MFCC for speech recognition in Indonesian language with the same data. Hopefully this research can help accelerate the development of using automatic speech recognition for indonesian.

Received November 4, 2016; Revised January 4, 2017; Accepted January 26, 2017

Besides feature extraction, another main process is recognition. Basically there are three approaches to speech recognition, namely [12] the acoustic-phonetic approach, pattern recognition approach and artificial intelligence approach. The speech recognition technique that is included in pattern recognition is the Hidden Markov Model (HMM) and Support Vector Machine (SVM) [6]. Singh et al, (2012) and O'Shaughnessy (2008) [12, 13] mentioned that the technique for speech recognition which consists of hundreds of thousands of words and that is still accepted up to now is the Hidden Markov Model (HMM), which appeared in 1975. According to Rabiner (1989), HMM is a stochastic process that occurs twice, with one of them being not a direct observation. A hidden stochastic process can be observed only through another set of stochastic processes that can produce the sequence of observation symbols. This is the reason that causes HMM to perform better than other methods [7]. In addition, the HMM technique is generally accepted in current speech recognition systems (state-of-the-art) in modern times because of two reasons, namely, its ability to model the non-linear dependence of each unit of the sound on the unit in question, and because it is a set of powerful analytical approaches that are available to estimate the model parameters [12]. This research will use HMM as the recognition method.

2. Research Method

The method used in this research can be seen in Figure 1.

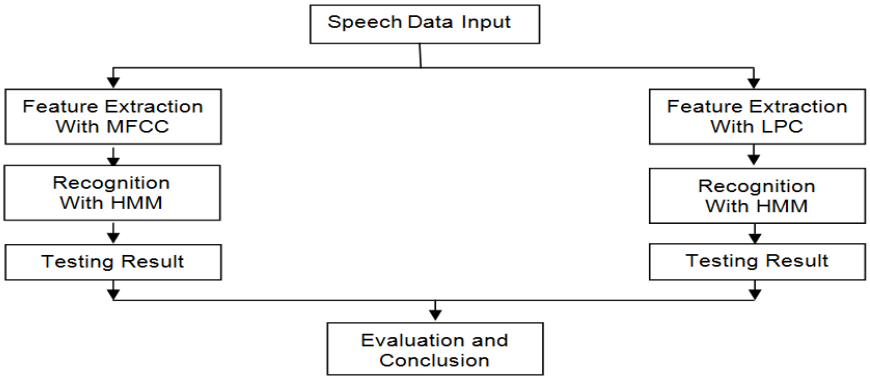

Figure 1. Research Method

Here is the explanation of each process.

1. Speech data input

Speech input in a speech file with .wav extension.

2. Feature extraction

There are two feature extraction will be compared in this research, that are MFCC and LPC. The stages of the process for the MFCC is as follows: DC-removal, Pre-emphasize, Frame Blocking, Windowing, FFT, Mel-Frequency Wrapping, Mel-Frequency Cepstrum and cepstral Filtering [5, 14]. DC-Removal is used to obtain a normalized value of the input data, by calculating the average of the sample data utterance. Preemphasize serves to reduce the signal noise ratio and balance the spectrum of the sound of a voice. Frame Blocking Process is used to cut the sound signal of long duration becomes shorter duration, in order to get characteristic of periodic signal. Windowing process aims to reduce spectral leakage or aliasing which is the effect of blocking frame which causes the signal becomes discontinue. FFT (Fast Fourier Transform) is a transformation method to get a signal in the frequency domain of the discrete signals exists. Filterbank conducted in order to determine the energy in the sound signal. The frequency of a signal is measured using mel scale. Mel-Frequency Cepstrum obtained from DCT process to get back the signal in domain time. The result is called the Mel-Frequency cepstral coefficient (MFCC). The results of MFCC have several drawbacks, namely the low-order which is very sensitive to spectral slope and the high-order which is very sensitive to noise. Therefore, the cepstral filtering has into one of these methods to minimize sensitivity.

For feature extraction using LPC has the following process steps Preemphasis, Frame Blocking, Windowing, autocorrelation analysis, LPC analysis, LPC Parameter Conversions Being cepstral coefficients, weighting parameters and cepstral Temporal derivatives [11, 14]. Preemphasis is a process to spread the utterance signal spectrum is aimed to reduce too extreme differences between a signal with a previous signal. Frame blocking is process dividing utterance signal become several frame. Windowing is used to reduce signal discontinuity at the start and end of frame. The window used is Hamming Window. The utility of autocorrelation process is to correlate the wave form. The next step is LPC analysis, which change every autocorrelation frame p+1 in the form of LPC parameters or commonly called the LPC coefficients. Then convert the LPC parameter into cepstral coefficient. Weighting is conducted for low-order cepstral coefficient is sensitive to the slope of the spectrum and higher order cepstral coefficients are sensitive to noise, then weighted cepstral coefficient with a window filter so as to minimize the those sensitivity. Temporal cepstral derivative (delta cepstral) is used to increase the representation of the properties spectral signal analyzed on parameters.

3. Recognition

Hidden Morkov Model (HMM) is an approach that cans classify the characteristic of spectral from each part of sound in several patterns. Basic theory from HMM is with grouping sound signal as random parametric process, and this process parameter can be recognized (prediction) in precise accuration [14].

HMM have five components that are:

a. Amount of state (N)
   State is hidden parameter (hidden state). In application amount of this state become one of thus testing parameter. So, amount of state is set in such a way to obtain an optimal output. The number of states in the model Nstate labeled with $S = \{S_1, S_2, \dots, S_N\}$.
b. Model Parameter (M)
   Number of observation symbol that different in each state M. observation symbol correlates with physical output from modeled system. Individual symbols is denoted by $V = \{v_1, v_2, v_3, \dots, v_M\}$.
c. Early state distribution $\pi = (\pi_i)$ where:

$$\pi = P[q_1 = i], 1 \leq i \leq N \quad (1)$$

d. Transition probability distribution state $A = (a_{ij})$ where:

$$a_{ij} = P\left[q_{u+1} = s_j \middle| q_u = s_i\right], 1 \leq i, j \leq N \quad (2)$$

   That is probably an observation is in a state j when u+1 and when state i when u.
e. The observation symbol probability distribution $B = \{b_j(k)\}$ where:

$$b_j(k) = P[o_u = v_k | q_u = j], 1 \leq k \leq M \quad (3)$$

Represent symbol distribution in state *j, j = 1,2,3,…,N*

According to five component above, to plan HMM, needs two model parameters that is N and M, besides it also needs three possibility (π, A, B) that is modeled by use notation λ [λ = (A, B, π)].

According to Rabiner [14], problem can be solved by HMM are:

1. Arrange parameter $\lambda = \mathrm{P}(\mathrm{A}, \mathrm{B}, \pi)$ in order to produce maximum $P(O|\lambda)$
2. Counting $P(O|\lambda)$ if known an observation sequence $O = O_1, O_2, \dots, O_T$ and a model $\lambda = P(A, B, \pi)$.

Detail the process of speech recognition is developed from Endah, et al, (2015) [5]. It can be seen in Figure 2.

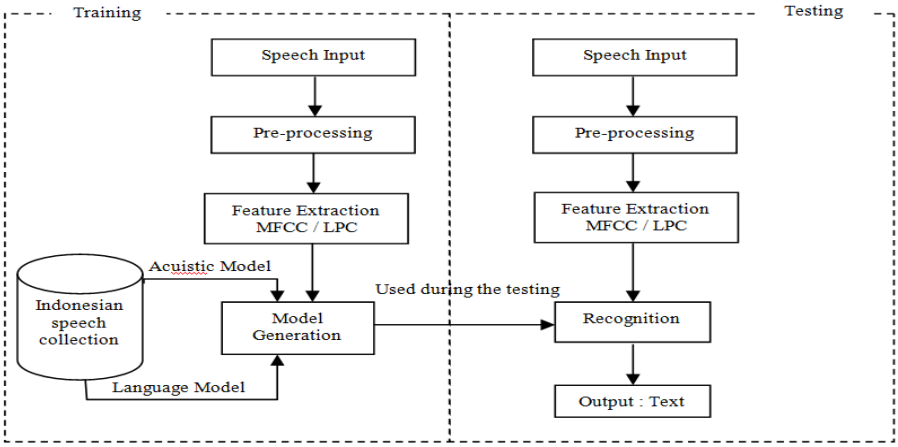

Figure 2. Detail Process of Speech Recognition

4. Testing Result

By using two scenarios, several data will be examined to get the level of accuracy in every feature extraction.

5. Evaluation and Conclusion

The test result will be evaluated to get a conclusion.

3. Results and Discussion

Testing is held by using a 10-fold cross validation, where validation is performed 10 times for each pair MFCC coefficients and HMM state or LPC-order and HMM state. Coefficients MFCC used is 8, 12 and 20, using the LPC Order 8-16, while the HMM state for MFCC is 3-15 and HMM state for LPC using 2, 3, 4, 7, 15 and 16. Decision pair coefficients MFCC and HMM state or LPC-order and HMM state is based on the results of several previous studies that have resulted in a high degree of accuracy. Data sets and composition of the 10-fold cross validation are shown in Table 1 and Table 2. Besides, this testing also used two scenarios for different sets of data to determine the level of accuracy of the system either by using MFCC feature extraction methods and LPC.

Table 1. Cross Validation Data Set

	Data Set
Male Recorder	A, B, C, D, E
Female Recorder	F, G, H, I, J

Table 2. Composition of 10-Fold Cross Validation

	Training Data	Test Data
1st Iteration	A2, B1, B2, C1, C2 ,D1 ,D2 ,E1 ,E2, F2, G1, G2, H1, H2, I1, I2, J1, J2	A1, F1
2nd Iteration	A1, B1, B2, C1, C2 ,D1 ,D2 ,E1 ,E2, F1, G1, G2, H1, H2, I1, I2, J1, J2	A2, F2
3rd Iteration	A1, A2, B2, C1, C2 ,D1 ,D2 ,E1 ,E2, F1, F2, G2, H1, H2, I1, I2, J1, J2	B1, G1
4th Iteration	A1, A2, B1, C1, C2 ,D1 ,D2 ,E1 ,E2, F1, F2, G1, H1, H2, I1, I2, J1, J2	B2, G2
5th Iteration	A1, A2, B1, B2, C2 ,D1 ,D2 ,E1 ,E2, F1, F2, G1, G2, H2, I1, I2, J1, J2	C1, H1
6th Iteration	A1, A2, B1, B2, C1 ,D1 ,D2 ,E1 ,E2, F1, F2, G1, G2, H1, I1, I2, J1, J2	C2, H2
7th Iteration	A1, A2, B1, B2, C1 ,C2 ,D2 ,E1 ,E2, F1, F2, G1, G2, H1, H2, I2, J1, J2	D1, I1
8th Iteration	A1, A2, B1, B2, C1 ,C2 ,D1 ,E1 ,E2, F1, F2, G1, G2, H1, H2, I1, J1, J2	D2, I2
9th Iteration	A1, A2, B1, B2, C1 ,C2 ,D1 , D2, E2, F1, F2, G1, G2, H1, H2, I1, I2, J2	E1, J1
10th Iteration	A1, A2, B1, B2, C1 ,C2 ,D1 , D2, E1, F1, F2, G1, G2, H1, H2, I1, I2, J1	E2, J2

The following are details of the results of the test data utterance that has been done.

1. 1st Scenario

The Data used to process this test is 600 Data consist of 15 words uttered by 10 different people (5 male and 5 female) that are uttered 4 times for each person. The words used in this test are “adik”, “ayah”, “botol”, “cerdas”, “dunia”, “ikan”, “jual”, “keluarga”, “lenyap”, “mimpi”, “minum”, “om”, “pasar”, “pergi”, dan “toko”. The words are recorded using sampling frequency 44100Hz, channel mono, 6 bit in 1 second.

Using cross validation dataset divide into 10 partitions for male and female recorder. Then it’s held 10 times iteration. Every iteration using 60 utterance data (30 male utterances and 30 female utterances) that consist of 15 words. While the result 540 utterance data are becomes training data. For 10 times iteration counted the number of wrong words recognized from male test data and female test data totaled 600 testing data (300 male testing data, 300 female testing data).

After held 10-foldcross validation test result using MFCC feature extraction for every couple of coefficients MFCC and HMM state can be seen in Table 3 below. While the test results with LPC feature extraction for all LPC orders and HMM state can be seen as illustrated in Table 4.

Table 3. Testing Results 1st Scenario with the MFCC Feature Extraction

HMM State	MFCC Coefficient		
	8	12	20
3	50,83%	74,67%	86,67%
4	60,67%	77,50%	87,83%
5	59,33%	79,67%	87,50%
6	61,50%	82,50%	89,83%
7	65,17%	83,83%	89,50%
8	65,83%	84,67%	91,00%
9	66,67%	85,33%	91,17%
10	69,83%	86,00%	91,50%
11	67,83%	87,17%	93,00%
12	70,33%	89,50%	91,83%
13	72,17%	90,00%	92,83%
14	74,17%	90,33%	91,67%
15	72,17%	89,67%	93,50%

Table 4. Testing Results 1st Scenario with the LPC Feature Extraction

HMM State	LPC Order								
	8	9	10	11	12	13	14	15	16
2	40,83%	45,50%	48,17%	52,50%	51,50%	52,50%	55,17%	54,17%	56,50%
3	42,83%	49,67%	51,00%	54,50%	55,67%	55,83%	60,17%	63,00%	64,17%
4	43,00%	53,17%	54,67%	58,50%	60,33%	62,17%	61,67%	62,67%	65,17%
7	52,17%	58,83%	61,83%	64,67%	65,83%	66,83%	68,17%	70,50%	69,83%
15	67,83%	69,33%	73,83%	76,00%	79,83%	77,33%	78,00%	80,17%	81,17%
16	66,00%	71,67%	71,50%	75,67%	78,33%	78,50%	79,17%	80,17%	80,17%

2. 2nd Scenario

The data used for this testing process is 1000 words data. It consists of 10 words uttered 10 times by 10 different people (5 male and 5 female). The words used in this study are “dan”, “diponegoro”, “fakultas”, “informatika”, “jurusan”, “matematika”, “sains”, “semarang”, “teknik”, “universitas”. By using cross validation, dataset is divided into 10 partitions for male recorder and female recorder. Then do for 10 times iterations. Each iteration tested using 100 words data (50 words male and 50 words female) consisting of 10 words. The remaining 900 words data used as training data. For 10 iterations, counted the number of false words recognized from male and female test data in numbered 1000 testing data (500 male testing data and 500 female testing data). The word is recorded by using sampling frequency of 8000Hz, mono channel, 16 bit and carried for 1 second. After 10-foldcross validation test results using MFCC feature extraction for each pair MFCC coefficient and HMM state can be seen in Table 5 below. While the test results with LPC feature extraction for all LPC order and HMM state can be seen in Table 6.

Table 5. Test Results 2nd Scenario with MFCC Feature Extraction

HMM State	MFCC Coefficient		
	8	12	20
3	40,30%	54,80%	71,10%
4	45,40%	58,60%	75,90%
5	50,10%	61,60%	75,30%
6	49,00%	58,70%	78,20%
7	51,10%	66,80%	78,80%
8	53,90%	69,70%	79,70%
9	50,20%	70,90%	81,00%
10	53,40%	70,80%	82,80%
11	54,90%	71,30%	84,60%
12	58,30%	73,80%	85,10%
13	56,60%	74,70%	84,60%
14	57,00%	76,60%	83,50%
15	58,00%	76,50%	84,90%

Table 6. Test Results 2nd Scenario with LPC Feature Extraction

HMM State	LPC Order								
	8	9	10	11	12	13	14	15	16
2	81,40%	83,50%	83,10%	84,80%	83,70%	84,90%	82,10%	82,50%	83,90%
3	81,10%	84,30%	84,80%	85,60%	87,40%	85,00%	86,60%	83,80%	86,30%
4	83,40%	85,40%	85,50%	87,40%	87,40%	88,50%	87,80%	88,50%	87,60%
7	84,50%	86,70%	89,40%	90,00%	91,70%	92,30%	91,70%	91,10%	90,00%
15	88,64%	91,50%	90,30%	91,30%	93,00%	92,10%	92,90%	91,70%	92,60%
16	89,10%	91,00%	90,30%	92,20%	92,60%	94,20%	92,60%	92,90%	93,60%

Based on the test results above shows that the higher of MFCC coefficient, so the level of accuracy also higher. Otherwise the size of the HMM state did not significantly affect the accuracy of the results. This is because the greater value of the coefficient used, the representation of speech signal characteristics become more detailed. On the contrary, the size of the LPC order significant did not affect the results, while the greater of its HMM state produce higher accuracy value.

The highest accuracy result in each scenario can be summarized as shown in Table 7.

Table 7. The highest accuracy result in each scenario

Scenario	Method	
	MFCC	LPC
1	93,50%	81,17%
2	84,9%	94,20%
Average	89,2%	87,68%

With the same data, the above table showed that the MFCC feature extraction is still better than the LPC in Indonesian language speech pattern recognition. This is in line with the results of research conducted by Sakti [10] and Thiang [11]. Although many of the parameters involved, Sakti's research which uses MFCC produces an accuracy of 92,47%, while Thiang's research that uses LPC produces an accuracy of 91,4%. In addition, computing time at the training process for MFCC feature extraction takes longer than LPC.

4. Conclusion

The conclusion of this research is that the MFCC method is better than the LPC method in Indonesian language speech recognition. The greater value of MFCC coefficient, the values of accuracy will be higher.

References

[1] Patel I, Rao YS. Speech Recognition using HMM with MFCC-AN Analysis using Frequency Spectral Decomposion Technique*. Signal & Image Processing*: *An International Journal (SIPIJ).* 2010; 1(2): 101-110.

[2] Jadhav A, Patil A. A Smart Texting System for Android Mobile Users. *International Journal of Engineering Research and Applications (IJERA)*. 2012; 2 (2): 1126-1128.

[3] Jadhav A, Patil A. Android Speech to Text Converter for SMS Application. *IOSR Journal of Engineering*. 2012; 2(3): 420-423.

[4] Sharma FR, Wasson SG. Speech Recognition and Synthetis Tool: Assistive Technology for Physically Disabled Persons. *International Journal of Computer Science and Telecommunications*. 2012; 3(4): 86-91.

[5] Endah SN, Adhy S, Sutikno. Integrated System Design for Broadcast Program Infringement Detection. *Telecommunication Computing Electronics and Control (TELKOMNIKA)*. 2015; 12(2): 571-577.

[6] Anusuya MA, Katti SK. Speech Recognition by Machine: A Review. *International Journal of Computer Science and Information Security (IJCSIS)*. 2009; 6(3): 181-205.

[7] Kumar K, Aggarwal RK. Hindi Speech Recognition System Using HTK. *International Journal of Computing and Business Research*. 2011; 2(2): 2229-6166.

[8] Adam TB, Salam MS, Gunawan TS. Wavelet Cepstral Coefficients for Isolated Speech. *Telecommunication Computing Electronics and Control (TELKOMNIKA)*. 2013; 11(5): 2731-2738.

[9] Wang Y, Yang X, Zou J. Research of Emotion Recognition Based on Speech and Facial Expression. *Telecommunication Computing Electronics and Control (TELKOMNIKA)*. 2013; 11(1): 83-90.

[10] Sakti S, Kelana E, Riza H, Sakai S, Markov K, Nakamura S. *Development of Indonesian Large Vocabulary Continuous Speech Recognition System within A-STAR Project.* Proceeding of Workshop on Technologies and Corpora for Asia-Pasific Speech Translation*.* 2008: 19-24

[11] Thiang, Wijoyo S. *Speech Recognition Using Linear Predictive Coding and Artificial Neural Network for Controlling Movement of Mobile Robot*. Proceeding of International Conference on Information and Electronics Engineering. 2011; 6: 179-183.

[12] Singh B, Kapur N, Kaur P. Speech Recognition with Hidden Markov Model: A Review. *International Journal of Advanced Research in Computer Science and Software Engineering.* 2012; 2(3): 400-403.

[13] O'Shaughnessy D. Invited paper: Automatic Speech Recognition: History, Methods and Challenges. *Pattern Recognition Science Direct.* 2008; 41: 2965-2979.

[14] Rabiner L, Juang BH. Fundamentals Of Speech Recognition. Englewood Cliffs, New Jersey: PTR Prentice-Hall, Inc. 1993.