# Automatic Speech Recognition for Indonesian using Linear Predictive Coding (LPC) and Hidden Markov Model (HMM)

Sukmawati Nur Endah[1, a], Satriyo Adhy[2,b], Sutikno[3,c] and Rizky Akbar[4,d]

[1,2,3,4] Informatics Department, Faculty of Science and Mathematics, Universitas Diponegoro

Prof. Soedharto Street, Kampus UNDIP Tembalang, Semarang, Indonesia

[a]1, [b]satriyo@undip.ac.id, [c]tik@undip.ac.id, [d]email.akbarrizky@gmail.com

**Keywords:** *Linear Predictive Coding* (LPC), *Hidden Markov Model* (HMM), Speech Recognition

## Abstract

Speech recognition is influential signal processing in communication technology. Speech recognition has allowed software to recognize the spoken word. Automatic speech recognition could be a solution to recognize the spoken word. This application was developed using Linear Predictive Coding (LPC) for feature extraction of speech signal and Hidden Markov Model (HMM) for generating the model of each the spoken word. The data of speech used for training and testing was produced by 10 speaker (5 men and 5 women) whose each speakers spoke 10 words and each of words spoken for 10 times. This research is tested using 10-*fold cross validation* for each pair LPC order and HMM *states*. System performance is measured based on the average accuracy testing from men and women speakers. According to the test results that the amount of HMM states affect the accuracy of system and the best accuracy is 94, 20% using LPC order =13 and HMM state=16.

## Introduction

Signal processing has important role in science and technology, especially in communication technology, both analog signal processing and digital signal processing. One of signal processing field that powerful in communication and technology is speech recognition.

Automatic speech recognition enables software to recognize and understand the spoken words by means digitalization of words and match digital signal with particular scheme. The spoken words are changed become digital signal by changed speech wave become batch of number then adjusted with particular codes to recognize those words. The outcome from spoken words recognition can be showed in text [1].

One of the methods that can be used in speech recognition is LPC as a speech feature signal and HMM as pattern recognition. LPC is one of feature technique that works properly in speech recognition. Meanwhile, LPC method is mathematically accurate and simple to be applied. LPC also provides accurate and efficient speech parameter for computation. The steps in LPC are preemphasis, frame blocking, windowing, autocorrelation analysis, LPC analysis, LPC parameter conversion to cepstral coefficients, parameter weighting, and temporal cepstral derivative.

HMM is a method that can classify the spectral characteristic from every speech in several patterns. Basic theory from HMM is classifying speech signal as parametric random process, and those parameter can be recognized precisely. HMM is popular method and mostly used in pattern identification for speech recognition system because HMM is reliable in several speech recognition system and it is integrated well into the system [2].

Researches in Indonesian speech recognition field with LPC and HMM have been conducted and among of them is "Words identification by means Hidden Markov Model (HMM) Method through Linear Predictive Coding (LPC) feature" by steps LPC until LPC analysis [3] and "Speech Recognition Application as Regulatory Cars with remote control" by steps LPC until autocorrelation analysis [4]. Both of researches have not used preprocessing step of speech signal (amplitude normalization and endpoint detection) and all LPC steps.

Thus, the research has been conducted is making speech recognition application by use preprocessing speech signal (amplitude normalization and endpoint detection), LPC with all steps and HMM.

## Research Method

Voice recognition divided into two types, speech recognition and speaker recognition. Automatic speech recognition is identification process in computer to recognize the spoken words by someone without seeing the identity his/her identity by conducting an acoustic signal conversion, which respond by audio device (input speech device), meanwhile speaker recognition is someone's identity recognition from his/her voice [5].

Speech recognition is divided into two categories, [4]:

1. According to input resources
   They are independent speech recognition resource and dependent speech recognition resource. The utilization of each type has pluses and minuses, depend on the application.

2. According to detection method
   There is isolated-word speech recognition and continuous speech recognition. The first type only detects one word in each operation; meanwhile, the second type detects continuous speech or several combined words in sentence that is spoken by resources.

Preprocessing that is conducted before feature extraction of speech signal is amplitude normalizes and endpoint detection [6]. LPC process has several steps that it shows in Fig.1.

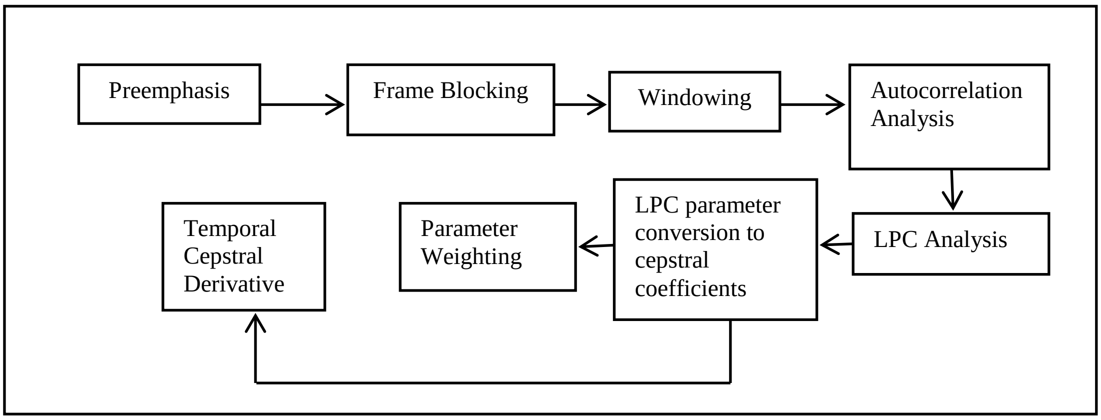

Fig.1. LPC Process

1. Preemphasis
   Preemphasis is a process to flatten speech signal spectrum intend to decrease extreme difference between one signal with previous signal. Outputs from preemphasis series are:

$$\tilde{p}(n) = x(n) - k_p * x(n-1), \qquad 0{,}9 \le k_p \le 1{,}0 \tag{1}$$

2. Frame Blocking
   The speech signal result from preemphasis $\tilde{p}(n)$ process is blocked or divided into several frames that is consist of $L_f$ – speech sample, with adjacent distance between frames is separated by $L_m$ sample. If $L_m \le L_f$, several adjacent frame will overlap each other and result estimation spectral LPC will correlates from frame to frame. Whereas, if $L_m > L_f$, there will no overlap between adjacent frame so that several speech sign signal will totally lost. The purpose from overlapping is so that there is no signal information that misses because of excision.

$$J_f = \left(\frac{L - L_f}{L_m}\right) + 1 \tag{2}$$

$$f(v,u) = \tilde{p}(((u-1) * L_m) + v) \tag{3}$$

3. Windowing
   Windowing is used to decrease signal discontinuity in beginning and end of frame. A type of window that is used is Hamming Window in general form:

$$h_w = 0{,}54 - 0{,}46 * \cos\left(\frac{2 * pi * v}{L_f - 1}\right) \tag{4}$$

$$\tilde{w}(v,u) = f(v,u) * h_w \tag{5}$$

4. Autocorrelation analysis
   Autocorrelation analysis of every signal frame will be conducted after through windowing process, autocorrelation equation:

$$r(m+1,u) = \sum_{v=1}^{L_f - m} \tilde{w}(v,u) * \tilde{w}(v+m,u), \quad m = 0,1,..,p \tag{6}$$

   Where the highest value from those autocorrelation, p is order from LPC analysis that will be conducted. General value for this LPC analysis order is 8 to 16.

5. LPC Analysis
   The next step is LPC analysis, which change every autocorrelation frame p+1 into LPC parameters or usually called LPC coefficient. The method usually used in this LPC analysis is Levinson-Durbin Method.

6. LPC parameter conversion to cepstral coefficients
   The series of parameter that can be derived directly from a series of LPC coefficient is cepstral coefficient c (m), which is determined recursively as follow:

$$c(m,u) = k_l(m,u) + \sum_{e=1}^{m-1} \left(\frac{e}{m}\right) * c(e,u) * k_l(m-e,u), \quad 1 \leq m \leq p \tag{7}$$

7. Parameter weighting
   Parameter weighting is conducted because cepstral coefficient order is low sensitive against the slope of the spectrum cepstral coefficient order high sensitive against the noise, thus coefficient cepstral weighting is conducted with window filter so that to minimize that sensitivity. The form of cepstral coefficient after weighting is:

$$\hat{c}(m,u) = w(m) * c(m,u), 1 \leq m \leq p \tag{8}$$

$$w(m) = \left[1 + \frac{p}{2} * sin\left(\frac{pi * m}{p}\right)\right], 1 \leq m \leq p \tag{9}$$

8. Derivatives temporal cepstral
   Derivatives temporal cepstral (delta cepstral) increase representation of the spectral characteristic of the signal that is analyzed in parameter. Derivatives temporal cepstral can be written as follows:

$$\frac{\partial c(m,i)}{\partial t} = \Delta c(m,u) \approx m_u * \sum_{e=-K}^{K} e * c(m,u) * (u + e) \tag{10}$$

   With (2K+1) is amount of frame where the calculation of the first derivative of the temporal cepstral conducted.

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 [2,7].

HMM have five components that are:

1. 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 $S = \{S_1, S_2, \ldots, S_N\}$.
2. 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 $V = \{v_1, v_2, v_3, \ldots, v_M\}$
3. Early state distribution $\pi = (\pi_i)$ where

$$\pi = P[q_1 = i], 1 \leq i \leq N \tag{11}$$

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

$$a_{ij} = P[q_{u+1} = s_j | q_u = s_i], 1 \leq i,j \leq N \tag{12}$$

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

$$b_j(k) = P[o_u = v_k | q_u = j], 1 \leq k \leq M \tag{13}$$

   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, problem can be solved by HMM are:

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

Automatic speech recognition application is a desktop-based application that serves as speech recognition and speech recognize as text. The main process in the application is speech recognition. The process starts by entering speech to be recognized. Speech input in the application is in the form of speech file that you want to identify, or via a live recording from the microphone. Furthermore, the applications process the speech recognition from the spoken word inputted and display text from the speech word. Speech recognition application process flow is shown in Fig. 2.

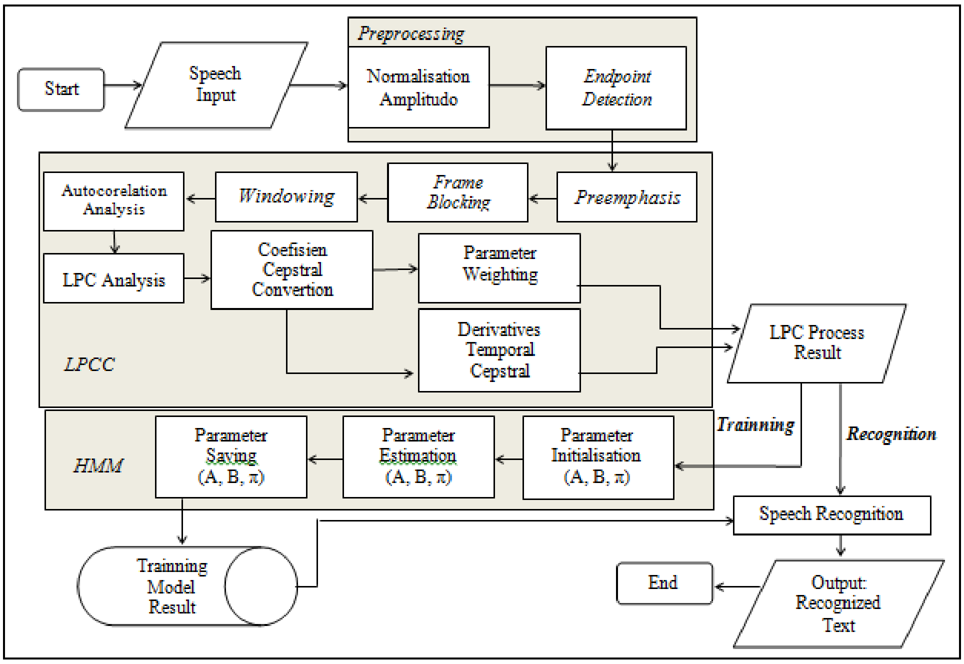

Fig 2. Speech recognition application process flow

The first process, speech input in the form of a set of words that will be trained. Speech is through the process of preprocessing and feature extraction using LPC. Preprocessing consists of normalized amplitude and endpoint detection. LPC consists of pre-emphasis filter, blocking frame, windowing, autocorrelation analysis, LPC analysis, LPC parameter conversion into cepstral coefficients, parameter weighting, and cepstral temporal derivative. LPC process results were used as observation for HMM process is a combination of parameter weighting and cepstral temporal derivative. The next stage is training done using HMM with speech that have undergone a preprocessing process and LPC feature extraction. The training phase produces a model that will be used for speech recognition.

In the speech recognition process, the user determines speech which wants to be recognized, the speech can be a file that have extension * .wav or speech is done directly through the microphone recording. Then speech passed the stage of preprocessing and LPC. The next stage, the application searches for the greatest probability value for each word based on the model that has been formed at the training stage. Word that has the greatest probability value is the speech recognized by the application.

## Experiment

Parameter testing is done by using a 10-fold cross validation, where validation is performed 10 times for each pair of HMM state. LPC order used is 8 to 16 and HMM state used is 2, 3, 4,7,15 and 16. The data used for these testing process as much as 1000 words data. Data consists of 10 words spoken as much as 10 times by 10 different people (5 men and 5 women). The words used in this study is “dan”, “diponegoro”, “fakultas”, “informatika”, “jurusan”, “matematika” , “sains” , “semarang”, “teknik”, and “universitas”.

By using cross validation, dataset is divided into a number of 10 partitions for man recorder and woman recorder. Then iteration is done a number of 10 iterations. Each iteration tested using 100 words data (50 man speech and 50 woman speech) consisting of 10 words. The remaining 900 words of data become training data. For 10 iterations counted the number of false words which recognized from the test data of man and woman as much as 1000 testing data (500 man testing data and 500 woman testing data). The accuracy level is measured based on the accuracy average of man and woman. Table 1 shows the experiment result after 10-fold cross validation for each pair of LPC order and HMM state. The graph of experiment result is can be also seeing in Fig 3.

Table 1. The experiment result for every pair of LPC order and HMM state

| **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% |

| HMM State | LPC Order | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 |
| 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% |

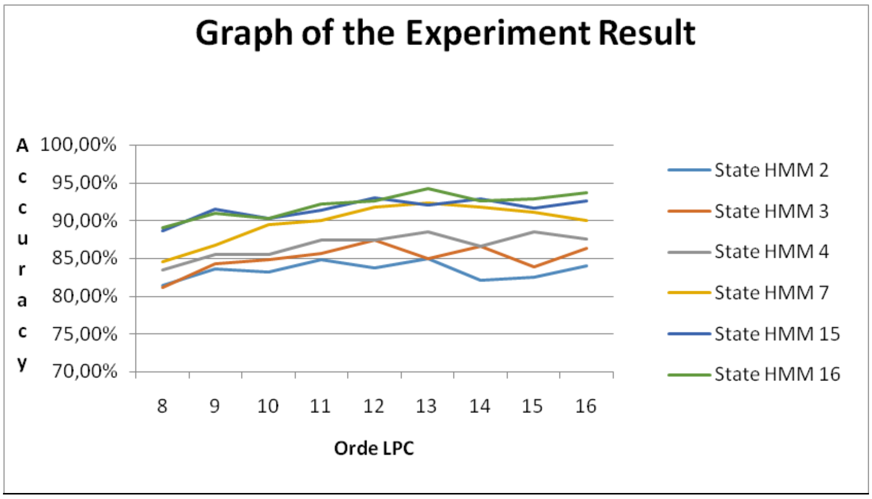

Fig 3. Accuracy Level Comparison Graph of Each Pair of LPC Order and HMM State

## Result Analysis

From Figure 3 shows that correspondence between the spoken words with the application recognition results obtained highest suitability percentage up to 94.20% in LPC order 13 and HMM state 16 based on the average of accuracy rate recorded man and woman.

Based on the degree of accuracy chart shown in Figure 3 can be seen that the average level of accuracy is influenced by the HMM state. In this research, higher value of HMM state, the accuracy rate is also higher, this is because the greater value of HMM state, the size of the HMM state matrix parameters are also getting bigger so it’s possible to generate a more optimum probability. While the magnitudes of the LPC order give no significant impact on the accuracy level. By observing the test parameters in each HMM state, LPC order value which produces maximum accuracy did not correlate with the increasing value of the LPC order used.

## Conclusion

The conclusion that can be drawn from this research is LPC and HMM can be used in speech recognition because it produces a fairly good level of accuracy, which reached 94.20% in LPC order = 13 and HMM state = 16. The amount of state were used in this study influence on the level of accuracy, but the value of the LPC order used does not affect the level of accuracy.

## References

[1]. Lestary, J., 2012. Aplikasi Pengenalan Ucapan Bahasa Inggris Menggunakan Linear Predictive Coding (LPC) dan Hidden Markov Model (HMM). [Online] Available at: http://publication.gunadarma.ac.id/bitstream/123456789/1082/1/50406418.pdf [Accessed 28 Agustus 2014].

[2]. Rabiner, L. & Juang, B.-H., Fundamentals Of Speech Recognition. Englewood Cliffs, New Jersey: PTR Prentice-Hall, Inc.(1993)

[3]. Munawar, B., Pengidentifikasi Kata Dengan Menggunakan Metode Hidden Markov Model (HMM) Melalui Ekstraksi Ciri Liniear Predictive Coding (LPC). Tugas Akhir. Bandung: Universitas Komputer Indonesia, (2010).

[4]. Saksono, M.W.T., Aplikasi Pengenalan Ucapan Sebagai Pengatur Mobil dengan Pengendali Jarak Jauh. Majalah Transmisi, 10(1), pp.21-26 (2008).

[5]. Nugraha, K., Aplikasi Perintah Suara Dengan Metode Fast Fourier Transform dan Divide And Conquer pada Simulasi Rumah Pintar. Tugas Akhir. Bandung: Teknik Informatika Unikom, (2011).

[6]. Saha, G., n.d. A New Silence Removal and Endpoint Detection Algorithm for Speech and Speaker Recognition Applications. [Online] Available at: http://citeseerx.ist.psu.edu/viewdoc/ summary? doi=10.1.1.138.623 [Accessed 5 Agustus 2014].

[7]. Syarief, Y., Simulasi Pengenalan Suara Menggunakan Model Hidden Markov. Tugas Akhir. Depok: Universitas Indonesia, (2000).