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

Sukmawati Nur Endah1,a, Satriyo Adhy1,b, Sutikno1,c and Rizky Akbar1,d

1Informatics Department, Faculty of Science and Mathematics, Universitas Diponegoro Prof. Soedharto Street, Kampus UNDIP Tembalang, Semarang, Indonesia

a sukmawati020578@gmail.com, b satriyo@undip.ac.id, c tik@undip.ac.id, d email.akbarrizky@gmail.com

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 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.

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

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](image)

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:

\[ \hat{p}(n) = x(n) - k_p * x(n - 1), \quad 0,9 \leq k_p \leq 1,0 \]  

(1)

2. **Frame Blocking**

The speech signal result from preemphasis \( \hat{p}(n) \) 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 \( L_m \leq L_f \), 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 signal will totally lost. The purpose from overlapping is so that there is no signal information that misses because of excision.

\[ j_f = \left( \frac{L_f}{L_m} \right) + 1 \]  

(2)

\[ f(v,w) = \hat{p}((u - 1) * L_m) + v \]  

(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:

\[ h_w = 0,54 - 0,46 * \cos \left( \frac{2 \pi p + \pi}{L_f - 1} \right) \]  

(4)

\[ \hat{w}(v,u) = f(v,u) * h_w \]  

(5)

4. **Autocorrelation analysis**

Autocorrelation analysis of every signal frame will be conducted after through windowing process, autocorrelation equation:
\[ r(m + 1, u) = \sum_{v=1}^{m} \tilde{w}(v, u) \ast \tilde{w}(v + m, u), \]
\[ m = 0, 1, \ldots, p \quad (6) \]

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:
\[ c(m, u) = k_j(m, u) + \sum_{e=1}^{m-1} \left( \frac{e}{m} \right) \ast c(e, u) \ast k_i(m - e, u), \quad 1 \leq m \leq p \quad (7) \]

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:
\[ \hat{c}(m, u) = w(m) \ast c(m, u), 1 \leq m \leq p \quad (8) \]
\[ w(m) = \left[ 1 + \frac{p}{2} \ast \sin \left( \frac{m+1}{m} \right) \right], 1 \leq m \leq p \quad (9) \]

7. Parameter weighting

Hidden Markov Model (HMM) is an approach that can 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 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
\[ \pi = P[q_1 = i], 1 \leq i \leq N \quad (11) \]

4. Transition probability distribution state \( A = (a_{ij}) \) where
\[ a_{ij} = P[q_u+1 = s_j | q_u = s_i], 1 \leq i, j \leq N \quad (12) \]

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
\[ b_j(k) = P[o_u = v_k | q_u = j], 1 \leq k \leq M \quad (13) \]

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 λ = P(A, B, π) in order to produce maximum P(O | λ)

2. Counting P(O | λ) if known an observation sequence O = O₁, O₂, ..., O_T and a model λ = P(A, B, π)

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.

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, 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.
experiment result is can be also seeing in Fig 3.

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

<table>
<thead>
<tr>
<th>HMM State</th>
<th>LPC Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>9 10 11 12 13 14 15 16</td>
</tr>
<tr>
<td>2</td>
<td>81.40% 83.20% 84.30% 84.90% 83.50% 83.90% 82.50% 83.50%</td>
</tr>
<tr>
<td>3</td>
<td>81.10% 84.30% 84.80% 85.20% 84.90% 85.30% 84.90% 85.00%</td>
</tr>
<tr>
<td>4</td>
<td>84.50% 86.70% 89.40% 90.30% 88.80% 89.00% 87.80% 88.90%</td>
</tr>
<tr>
<td>7</td>
<td>84.60% 86.70% 90.10% 90.70% 89.20% 90.00% 88.80% 87.80%</td>
</tr>
<tr>
<td>15</td>
<td>89.10% 91.00% 90.30% 92.20% 92.40% 92.60% 92.60% 92.60%</td>
</tr>
<tr>
<td>16</td>
<td>88.60% 91.00% 90.30% 92.20% 92.40% 92.60% 92.60% 93.60%</td>
</tr>
</tbody>
</table>

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**


