A PROBLEM IN DATA VARIABILITY ON SPEAKER IDENTIFICATION SYSTEM USING HIDDEN MARKOV MODEL

Agus Buono\textsuperscript{a}, Benyamin Kusumoputro\textsuperscript{b}

\textsuperscript{a}Computational Intelligence Research Lab, Dept. of Computer Science, Bogor Agriculture University Dramaga Campus, Bogor-West Java, Indonesia
pu desha@yahoo.co.id

\textsuperscript{b}Computational Intelligence Research Lab, Faculty of Computer Science, University of Indonesia, Depok Campus, Depok 16424, PO.Box 3442, Jakarta, Indonesia
nynykusumo@yahoo.com

ABSTRACT
The paper addresses a problem on speaker identification system using Hidden Markov Model (HMM) caused by the training data selected far from its distribution centre. Four scenarios for unguided data have been conducted to partition the data into training data and testing data. The data were recorded from ten speakers. Each speaker uttered 80 times with the same physical (health) condition. The data collected then pre-processed using Mel-Frequency Cepstrum Coefficients (MFCC) feature extraction method. The four scenarios are based on the distance of each speech to its distribution centre, which is computed using Self Organizing Map (SOM) algorithm. HMM with many number of states (from 3 up to 7) showed that speaker with multi-modals distribution will drop the system accuracy up to 9% from its highest recognition rate, i.e. 100%.

KEY WORDS
Hidden Markov Model, Mel-Frequency Cepstrum Coefficients, Self Organizing Map

2. Speaker Identification Using HMM

HMM is a Markov chain, where its hidden state can yield an observable state. A HMM is specified completely by three components, i.e. initial state distribution, $\Pi$, transition probability matrix, $A$, and observation probability matrix, $B$. Hence, it is notated by $\lambda = (A, B, \Pi)$, where:

$A$: $N \times N$ transition matrix with entries $a_{ij} = P(X_{t+1} = j | X_t = i)$, $N$ is the number of possible hidden states
$B$: $N \times M$ observation matrix with entries $b_{ik} = P(O_{t+1} = v_k | X_t = j)$, $k = 1, 2, 3, ..., M; M$ is the number of possible observable states
$\Pi$: $N \times 1$ initial state vector with entries $\pi_i = P(X_1 = i)$

Figure 1. Diagram block for speaker identification system using HMM

29
In the Mixture Gaussian HMM, $B$ consists of a mixture parameters, mean vectors and covariance matrices for each hidden state, $c_{ji}$, $\mu_{ji}$ and $\Sigma_{ji}$, respectively, $j=1, 2, 3, ..., N$. Thus value of $b_j(O_{t+1})$ is formulated as follow:

$$b_j(O_{t+1}) = \sum_{i=1}^{c} c_{ji}N(O_{t+1},\mu_{ji},\Sigma_{ji})$$

There are three problems in HMM [1], i.e. evaluation problem, $P(O|\lambda)$, decoding problem, $P(Q|O, \lambda)$, and training problem [3]. Diagram block for speaker identification system using HMM is presented in Figure 1. Probability for each speech data belong to a certain model is computed recursively using forward variable $\alpha_i$ as illustrated in Figure 2.

![Figure 2. Illustration of forward variable computation](image)

3. Methodology

3.1 Research Block Diagram

Figure 3 depicted the research methodology.

3.2 Data and Feature Extraction

Data used in the research are recorded from ten speakers using Matlab with 11KHz sampling rate and 1.28 second duration. Each speaker utters word “PUDESHA” 80 times, which resulted total of 800 speech data. In this case, a speaker has not to follow certain instruction in uttering the required word. Having in mind that these 80 data were recorded during the same physical condition, hence there is no difference in age, healthiness, and emotion.

![Figure 3. Research methodology block diagram](image)
On each frame, 20 coefficients of MFCC are computed. Thus, a speech data having T frames will be converted into a T by 20 data matrix. The following Figure 4 depicted the process, [4].

**Figure 4. Signal feature extraction using MFCC**

3.3 Formulation of Speech Data Variability

Variability indicates how objects positions distributed around its population centre. A population in which its objects scattered close to the centre having less variability as described in Figure 5.

![Population Variability Illustration](image)

**Figure 5. Illustration of population variability**

(a) higher variability (b) less variability

Using SOM, a codebook consists of 16 codewords from each speaker is generated. The codebook represents the centre of population.

A distance between a speech data and a certain codebook is calculated using the following formula:

\[ d(O, \text{codebook}) = \sum_{j=1}^{16} \text{average}_{t, j} \left( \forall_\text{t}, d(t, j) \right) \]

Where \( d(t, j) \) is the smallest Euclidean distance between \( t^{th} \) frame and \( j^{th} \) codeword, as illustrated in Figure 6.

![Frame Distribution Around 5 Codewords](image)

**Figure 6. A frame distribution around 5 codewords**

In this case:

- Codebook = \((w_1, w_2, w_3, w_4, w_5)\)
- \(O = O_1, O_2, O_3, \ldots, O_{10}\)

A distance \( d(O, \text{codebook}) \) between \( O \) and codebook is:

\[ d(O, \text{codebook}) = \frac{d(w_1, O_3) + d(w_1, O_4)}{2} + \frac{d(w_2, O_1) + d(w_2, O_5) + d(w_2, O_6)}{3} + \frac{d(w_3, O_2) + d(w_4, O_8) + d(w_4, O_10)}{2} + \frac{d(w_5, O_7) + d(w_5, O_9)}{2} \]

3.4 Training Data Scenario

For each speaker, its speech data is sorted according to the distance to its population centre in ascending order and divided into two equal groups, i.e. training data and testing data, where each having 40 data. The four scenarios proposed to investigate the system performance are described as follow:

1. **Scenario 1**: first 40 near data to the centre are grouped as training data
2. **Scenario 2**: first 40 near data to the centre are grouped as testing data
3. **Scenario 3**: 40 training data are sampled systematically, i.e. \(2, 4, 6, \ldots, 80\)th data
4. **Scenario 4**: 40 training data are randomly selected.

Figure 7 illustrates the scenario.
4. Experiments and Results

A mixture Gaussian HMM with three components is developed for each scenario with various numbers of states, i.e. from 3 up to 7 states, and observed its recognition rate. The experimental results are presented in Table 1.

Table 1 shows that HMM is able to recognize speaker with unguided utterance at 98.4% rate on the average. HMM with 6 numbers of states resulted the best recognition rate as described in Figure 8.

As expected from the scenario design, Scenario 2 yielded the worst recognition rate, i.e. 96.65% on the average as depicted in Figure 9 and Figure 10.
has an ignorable effect. In other word, HMM is a robust method to overcome these variations.

Detailed observation in Scenario 2, indicates a multi-modals speech data distribution of a speaker has a contribution in lowering the system recognition rate (Speaker 3, 8, and 9) as shown in Figure 11.

These findings predict that the system accuracy will drop further if speaker physical condition (healthiness, age, and emotion) is different. The prediction is most likely supported by the fact that a speaker physical condition will affect on the voice distribution pattern.

5. Conclusion

Experimental results proved the robustness of HMM speaker identification system for unguided utterance (duration and pressure) by resulting satisfaction recognition rate (on the average, 96.65% up to 99.3%) using four different scenarios were applied.

The optimal number of states in the HMM is 6 with 98.4% accuracy on the average. Furthermore, sampling technique in partitioning training and testing data also affects the system accuracy. In this case, random sampling gives the best result.

A multi-modals speech data of a speaker will significantly drop the system recognition rate.

6. Future Research

Future research will focused in handling a multi-modals speech data of a speaker in noisy environment using modified HMM as classifier. Also, seeks an appropriate feature extraction to handle the problems, i.e. multi-modals and noisy data.

References


