Proceedings of the 26th IASTED International Conference
ARTIFICIAL INTELLIGENCE AND APPLICATIONS
February 11–13, 2008, Innsbruck, Austria
ISBN Hardcopy: 978-0-88986-709-3 / CD: 978-0-88986-710-9

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

Agus Buono[a] , Benyamin Kusumoputro[b]
[a]Computational Intelligence Research Lab, Dept. of Computer Science, Bogor Agriculture University Dramaga Campus, Bogor-West Java, Indonesia
pudesha@yahoo.co.id
[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-Frequence 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-Frequence Cepstrum Coefficients, Self Organizing Map

1. Introduction

HMM has been widely applied for voice processing with promising results. However voice is a complex magnitude, which is influenced by many factors, i.e. duration, pressure, age, emotion, and sickness [1]. Due to these factors, until now, voice modeling has not reached a perfect result [2].

The research focuses on the problem of HMM application caused by inappropriate training data. Therefore four training data scenarios were proposed to analyze HMM performance, i.e. training data that has distance to its distribution centre: close, far, systematic, and random. The results of this study are expected to be used as the basic for advanced research in voice data modeling using HMM for different kind speaker physical conditions.

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, Π, transition probability matrix, A, and observation probability matrix, B. Hence, it is notated by $\lambda = (A, B, \Pi)$, where :

- A: NxN transition matrix with entries $a_{ij}=P(X_{t+1}=j|X_t=i)$, N is the number of possible hidden states
- B: NxM observation matrix with entries $b_{jk}=P(O_{t+1}=v_k|X_t=j)$, k=1, 2, 3, …, M; M is the number of possible observable states
- Π: Nx1 initial state vector with entries $\pi_i=P(X_1=i)$

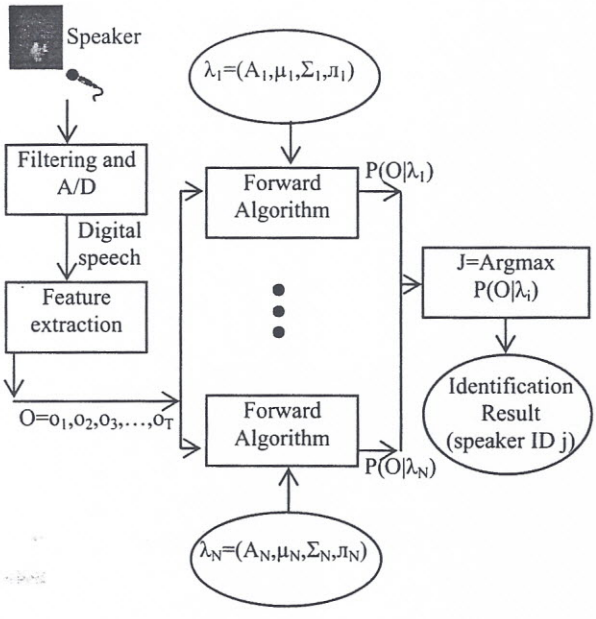

Figure 1. Diagram block for speaker identification system using HMM

In the Mixture Gaussian HMM, B consists of a mixture parameters, mean vectors and covariance matrixes for each hidden state, c_{ji}, μ_{ji} and Σ_{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|λ), decoding problem, P(Q|O, λ), 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 (α) as illustrated in **Figure 2**.

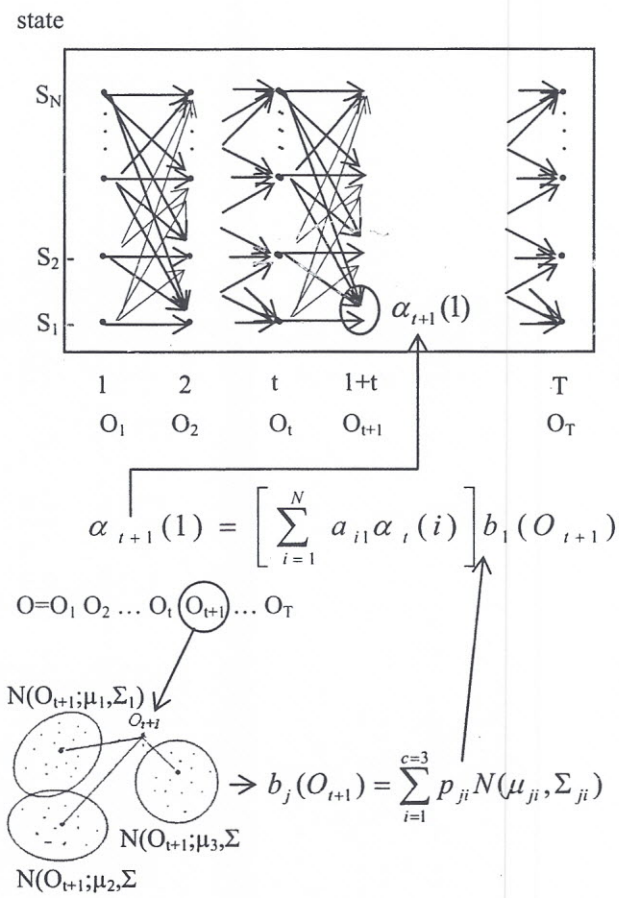

Figure 2. Illustration of forward variable computation

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.

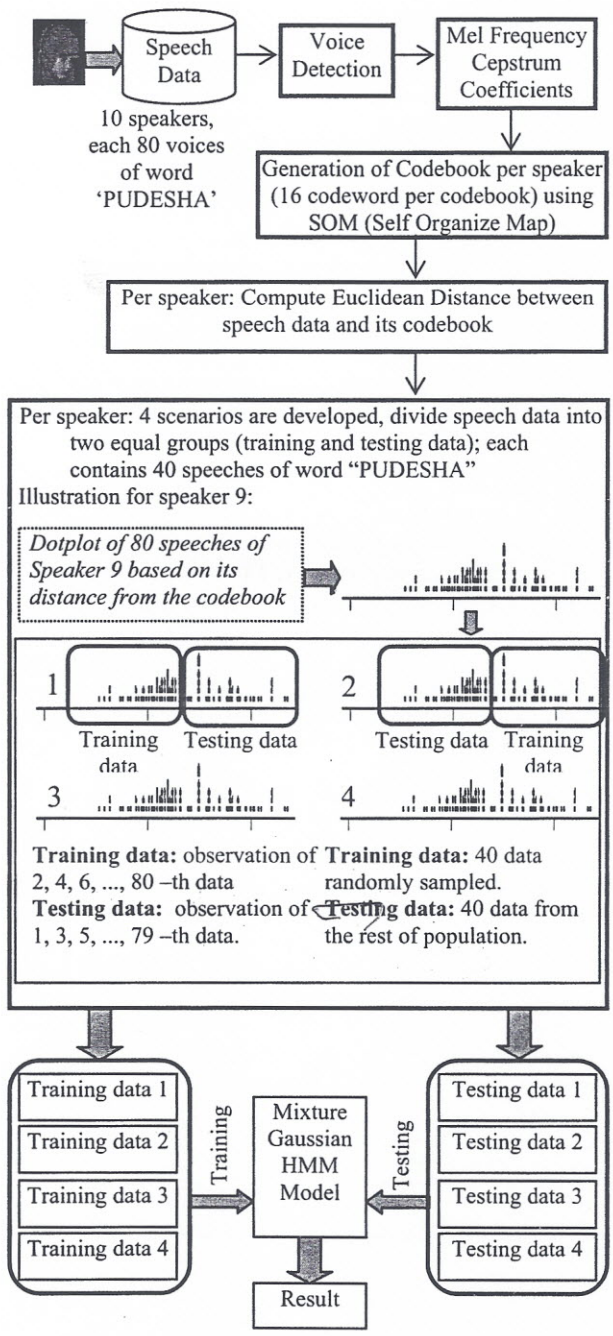

Figure 3. Research methodology block diagram

Each speech data is truncated for silence in the front and back recording as illustrated in **Figure 4**. Then this speech data is read frame by frame, which each frame has 256 data and 100 data overlapped between adjacent frames.

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].

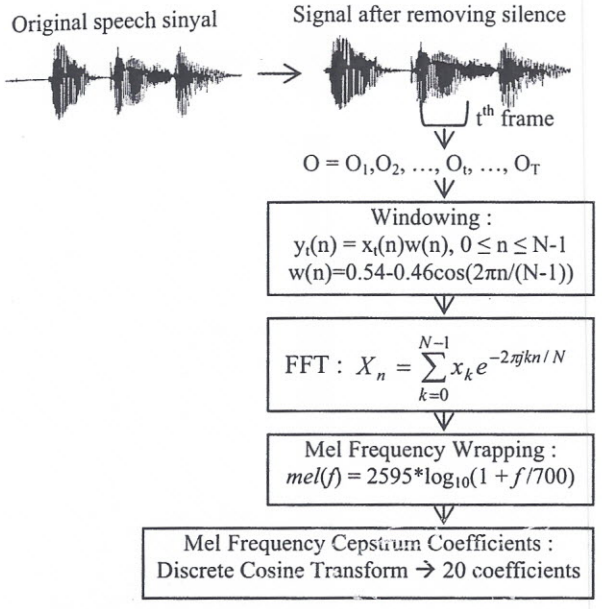

Figure 4. Signal feature extraction using MFCC

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**.

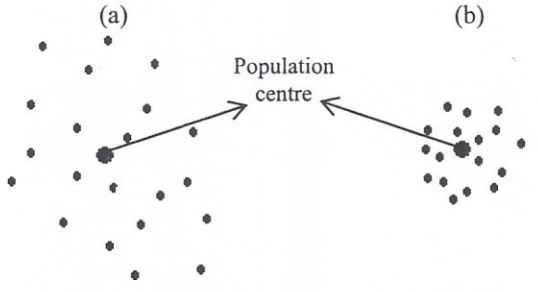

Figure 5. Illustration of population variability (a) higher variability (b) less variability

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

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

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

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

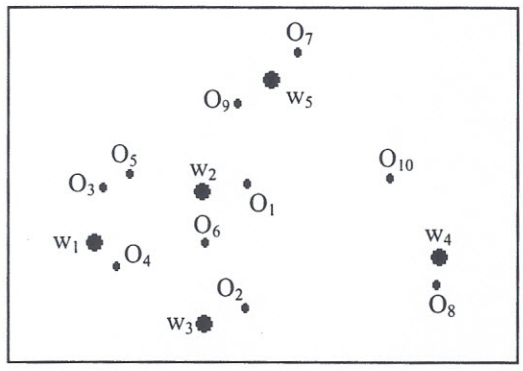

Figure 6. A frame distribution around 5 codewords

In this case:

Codebook = (w_1, w_2, w_3, w_4, w_5), where w_j : j^{th} codeword

O = O_1, O_2, O_3, …, O_{10}; where O_t : t^{th} frame

A distance d(O, codebook) between O and codebook is:

$$d(O, codebook) = \frac{d(w_1, O_3) + d(w_1, O_4)}{2} + \frac{d(w_2, O_1) + d(w_2, O_5) + d(w_2, O_6)}{3} + d(w_3, O_2) + \frac{d(w_4, O_8) + d(w_4, O_{10})}{2} + \frac{d(w_5, O_7) + d(w_5, O_9)}{2}$$

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, ..., 80 –th data
4. Scenario 4 : 40 training data are randomly selected.

Figure 7 illustrates the scenario.

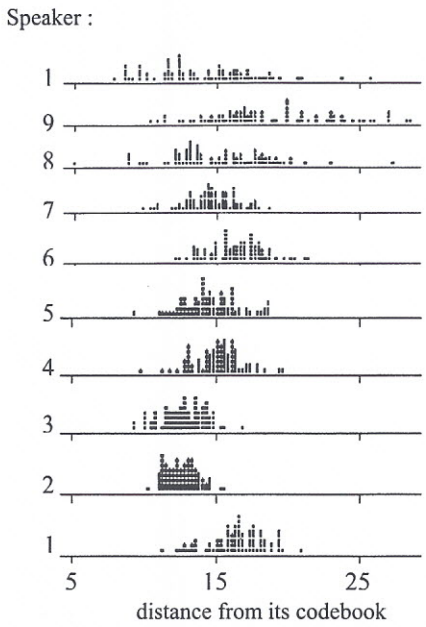

Figure 7. Dotplot of 80 speech data for 10 speakers

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**.

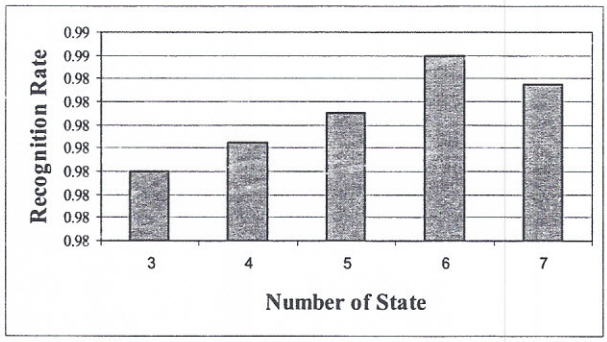

Figure 8. A comparison of recognition rates for various numbers of states

Table 1. Recognition rate per speaker for various numbers of states for four scenarios (Sc.).

Sc.	Numb er of State					Rec. Rate
	3	4	5	6	7	
1	40	40	40	40	40	1.000
	40	39	39	40	40	0.990
	40	40	40	40	40	1.000
	40	40	40	40	40	1.000
	38	37	38	38	37	0.940
	39	40	39	39	39	0.980
	40	40	40	40	40	1.000
	38	38	38	38	38	0.950
	39	40	39	37	39	0.970
	40	40	40	40	40	1.000
2	39	39	40	40	40	0.990
	40	40	40	40	40	1.000
	36	37	36	36	36	0.905
	38	38	38	38	38	0.950
	40	40	40	40	40	1.000
	40	40	40	40	40	1.000
	40	40	40	40	40	1.000
	37	36	37	37	36	0.915
	36	36	38	38	37	0.925
	39	39	39	39	40	0.980
3	39	39	40	40	40	0.990
	40	40	40	40	40	1.000
	40	40	40	40	40	1.000
	40	40	39	39	39	0.985
	38	38	38	38	38	0.950
	40	40	40	40	40	1.000
	40	40	40	40	40	1.000
	36	36	36	38	36	0.910
	39	40	38	39	37	0.965
	40	40	40	40	40	1.000
4	40	40	40	40	40	1.000
	40	40	40	40	40	1.000
	40	40	40	40	40	1.000
	40	40	40	40	40	1.000
	40	39	39	39	39	0.980
	40	40	40	40	40	1.000
	40	40	40	40	40	1.000
	39	39	39	39	39	0.975
	40	40	40	39	40	0.995
	39	39	39	39	40	0.980
Rec. rate	0.9825	0.9831	0.9838	0.9850	0.9844	0.984

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**.

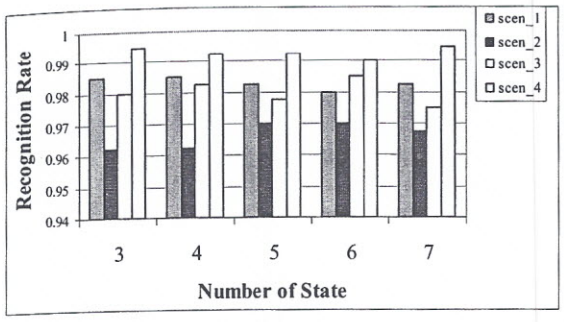

Figure 9. A comparison of recognition rates for various numbers of states for each scenario

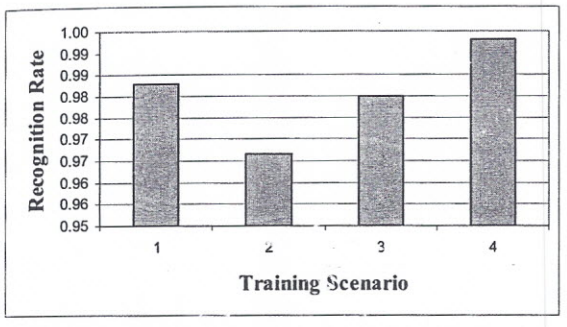

Figure 10. A comparison of recognition rates per scenario

These results indicate that if further objects were sampled for training data then recognition rate will significantly drop.

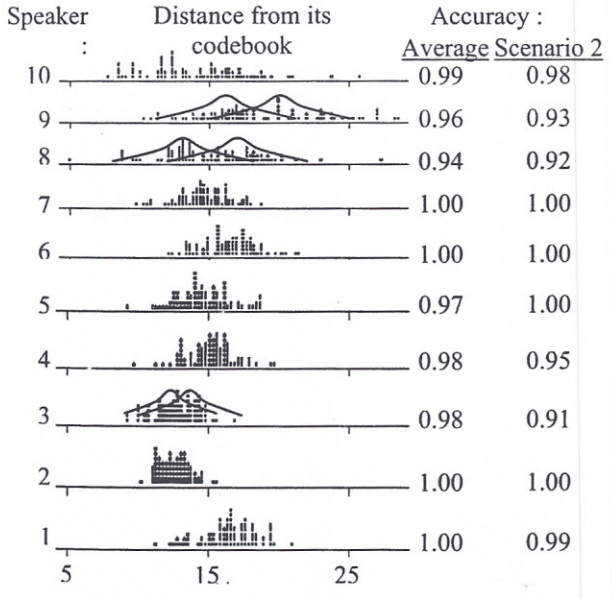

Figure 11. Recognition rate per speaker using Scenario 2

However, this scenario does not affect much HMM performance to recognize the speaker, i.e. 96.65% is a tolerable rate, due to variability in duration and pressure 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

[1] J. Campbell, "Speaker Recognition: A Tutorial", *Proc. of the IEEE*, Vol 85, No. 9, 1997, 1437-1462.

[2] L.R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", *Proceeding IEEE*, Vol 77 No. 2, 1989, 257-289.

[3] Rakesh D. "A Tutorial on Hidden Markov Model. Technical Report, Departement of Electrical Engineering, Indian Institute of Technology, Bombay", 1996

[4] Todor D. Ganchev. *Speaker Recognition.* PhD Dissertation, Wire Communications Laboratory, Department of Computer and Electrical Engineering, University of Patras Greece. 2005