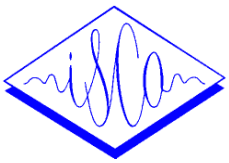

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

COMPARATIVE STUDY OF GMM, DTW, AND ANN ON THAI SPEAKER IDENTIFICATION SYSTEM

Chularat Tanprasert and Varin Achariyakulporn

Information Research and Development Division,
National Electronics and Computer Technology Center,
National Science and Technology Development Agency,
Ministry of Science, Technology, and Environment,
539/2 Gypsum Metropolitan Building, 22nd Fl., Sri-Ayudhya Road,
Rachathewi, Bangkok, 10400, THAILAND

ABSTRACT

This paper proposes a new investigation on Gaussian mixture model (GMM) by comparing it with some preliminary experiments on multilayered perceptron network (MLP) with backpropagation learning algorithm (BKP) and dynamic time warping (DTW) techniques on Thai text-dependent speaker identification system. Three major identification engines are conducted on 50 speakers with isolated digits 0-9. Training and testing utterances were recorded over a five week duration. Furthermore, three well-known speech features, namely linear predictive coding derived cepstrum (LPCC), postfiltered ceptrum (PFL), and Mel frequency cepstral coefficient (MFCC) were evaluated. From our previous experiments, the MFCC has given the highest identification rates on DTW and MLP. Therefore, GMM with MFCC feature was experimented and attained 87.54% average identification accuracy, as opposed to 86.74% of DTW and 82.34% of MLP. The results are the same with top-3 concatenated digits, the average identification rates are 99%, 98.70 %, and 97.30% for GMM, DTW, and MLP, respectively.

1. INTRODUCTION

Mostly, speaker recognition has an important role in the area of security systems (computer security, banking over the telephone, as well as access to the internet, etc.). It can be categorized into two groups: speaker verification and speaker identification [1]. Speaker verification task is to decide whether or not an unlabeled voice token belongs to a specific reference speaker, while speaker identification tries to determine who in the specific group of speaker's domain is speaking. To provide superior identification performance, a distinguished recognition engine must be developed. Many recognition engines have been proposed for this task. Some efficiency engines, specifically Dynamic Time Warping (DTW), Vector Quantization (VQ), and Hidden Markov Model (HMM) were comparatively performed in [2]. For text-dependent speaker identification systems, we can roughly assume that DTW is the most efficient approach, according to our previous works [3]. However, it takes a lot of processing time. So, other recognition engines were deeply evaluated.

In this paper, the Gaussian Mixture Model (GMM) is studied for Thai text-dependent speaker identification. GMM is evaluated and compared with MLP and DTW. The use of GMM for modeling speaker identity is motivated by the interpretation that Gaussian components represent some general speaker-dependent spectral shapes and the capability of Gaussian mixtures to model arbitrary densities [4]. GMM, DTW, and MLP were experimentally evaluated on 50 speakers of isolated digits 0-9 of Thai Language which contain a combination of long-vowels and short-vowels and different five tones. The experiments also mentioned about comparison on three speech features: linear predictive coding derived cepstrum (LPCC), postfiltered ceptrum (PFL), and Mel frequency cepstral coefficient (MFCC). Finally, concatenated digits of top-3 identification rate were obtained on three major recognition engines to increase an identification rate.

2. SPEAKER IDENTIFICATION SYSTEM

2.1 Proposed Speaker Identification System

Architecture of speaker identification system derived from paper [9] is shown in figure 1.

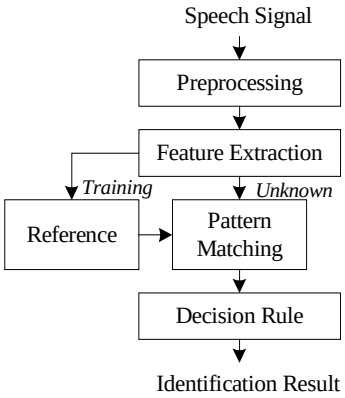

Figure 1: Speaker identification model.

Speech signal - Speech signal in our experiments were recorded by a typical computer sound card in an office environment with 11.025 kHz sampling rate, 16-bit quantization level and a single channel. To avoid familiarity with the digit sequence, random number generation is conducted for speaking utterances with our automatic sound recording software.

Preprocessing – This step is implemented to modify an input speech signal in an appropriate format. Each utterance was passed through a high-pass filter with 200 Hz to eliminate surrounding noise. Later, an automatic energy-based approach for endpoint detection and preemphasis process were applied by

blocking the signal into 20-ms frames with a quarter of each frame overlapping. Finally, each frame of speech was smoothed with a Hamming smoothing window.

Feature extraction – This step extracts a set of essential characteristics that can identify or represent the whole speech signal. Numerous feature extraction algorithms have been proposed for speaker identification tasks. Three well-known speech features: LPCC, PFL, and MFCC were conducted and evaluated in this paper. Their details will be described in the next section.

Pattern matching – There are many identification engines to distinguish the owner of the speaking text. GMM, DTW, and MLP are well-known engines that are effective for the speaker identification task. Each engine will be explained in section 2.3.

Decision rule – The simplest way to decide what class an unknown pattern belongs to is to consider the minimum distance of this unknown pattern to each class referenced template. We used this method with GMM and MLP. This is because their models train reference utterances and return a single output. Another approach is to use the K-Nearest Neighbor (KNN) technique. The algorithm is to consider the K-minimum distances obtained from each pair of an unknown and each reference. The most frequent reference class within these K pairs is the answer of the unknown pattern. In our work, 5-NN was applied with DTW to get acceptable accuracy rate.

2.2 Speech Features

The feature extraction's step is designed to substitute the whole speech signal with a collection of predominant feature that still preserve the original signal's characteristic as much as possible. Proposed features can be classified into two groups: "high level" features such as dialect, context, speaking style, etc., and "low level" features such as spectral envelop-based features and prosodic features. The speech spectrum has been shown to be very effective feature for a speaker identification [5] because the spectrum reflects a person's vocal tract structure and it has quite a high likelihood of distinguishing one person's voice from others. In this paper, LPCC, PFL, and MFCC are chosen to experiment on Thai text-dependent speaker identification task.

Linear Prediction Coefficient Derived Cepstrum (LPCC) – To calculate cepstral coefficient, linear predictive coding (LPC) must first be calculated which is proposed by [6]. Then, we can easily compute the cepstral coefficient recursively from LPC [7] as described in equation (1).

$$c_{lpc}(n) = \begin{cases} a_n, n = 1 \\ \sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c_{lpc}(n-k) + a_n, 1 < n \le p \\ \sum_{k=1}^{n-1}(1-\frac{k}{n})a_k c_{lpc}(n-k), n > p \end{cases} \quad (1)$$

Postfiltered Cepstrum (PFL) – Postfiltered cepstrum is one of the most prominent derivations and has been proven to be strongly usable for speech and speaker recognition [8]. The concept of PFL was introduced to enhance noisy speech by emphasizing the spectral peaks (formant regions), which are rarely affected by noise. PFL can be practically computed from LPCC as shown in equation (2).

$$c_{pfl}(n) = c_{pfl}(n)[\alpha^n - \beta^n] \quad (2)$$

Mel-Frequency Cepstral Coefficient (MFCC) – Spectral representations have been used extensively for speaker recognition, however, these model-base representations can be severely affected by noise. A more efficient improvement is MFCC [9], which is more robust for noisy speech recognition with directly computed filterbank features by taking the shifted discrete cosine transformation of Mel-scale spectrum as shown in equation (3).

$$c_{mel}(n) = \sum_{k=1}^{K} \log(\tilde{S}_k)\cos(n(k-0.5)\frac{\pi}{k}) \quad (3)$$

2.3 Pattern Matching

Several identification engines have been tested for the speaker identification system. In this paper, three popular engines were tested. They are Gaussian mixture model, dynamic time wraping, and multilayered perceptron network with backpropagation learning algorithm.

Gaussian Mixture Model (GMM) – GMM as described in this paper is referred from [4]. A Gaussian mixture density is a weighted sum of M component densities given by equation (4)

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \quad (4)$$

where $\vec{x}$ is a D-dimensional random vector, $b_i(\vec{x})$ are the component densities and p_i are the mixture weights. Component density is a D-variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\sum_i|^{1/2}} \exp\left\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^t \sum_i^{-1}(\vec{x}-\vec{\mu}_i)\right\} \quad (5)$$

with mean vector $\vec{\mu}_i$ and covariance matrix $\sum_i$. The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities as shown in (6).

$$\lambda = \{p_i, \vec{\mu}_i, \sum_i\}, i = 1,..., M. \quad (6)$$

λ refers to each speaker in the GMM. In this paper, the GMM model has one covariance matrix per Gaussian component which is called the "nodal covariance".

Dynamic Time Warping (DTW) – DTW is a well-known algorithm to calculate the distance of each pair of an unknown sequence and a referenced sequence of the speech signal.

DTW can automatically adjust against pairs to achieve minimum cumulative distance. Figure 2 illustrates DTW algorithm, where A and B represent a pair of mapping sequence. DTW allows flexible matching points for interacting sequences, called time-alignment. A time-alignment window (r) is a

conditional value for DTW, which defines the allowable time warping in unit of frame. Changing the value of r will result in an effect on both processing time and system performance. Our previous experiment [3] has been evaluated with the value of r = 5, which is appropriated for Thai isolated digit utterance. Euclidean distance method is used to find a cumulative distance. The details of DTW algorithm are given in [10].

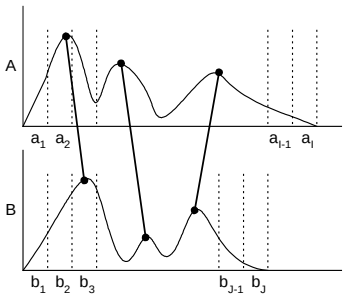

Figure 2: An illustration of the DTW algorithm

Multi-layered Perceptron (MLP) with Backpropagation Learning Algorithm (BKP) – MLP with BKP has been chosen to use in our experiment since it has been successfully applied to many pattern classification problems including speaker recognition. In our previous experiment [12], a windowing technique has been shown to collect speech characteristics more effectively than using time normalization method. Hence, the windowing technique was applied to MLP with BKP. Each MLP consists of four layers: one input layer, two hidden layers and one output layer. Each input vector is created for every four frames with three-frames overlapping. The ANN simulator software named SNNS [11] was used in the training and testing process.

3. EXPERIMENTAL RESULTS

Speech data was collected from 50 speakers (30 males and 20 females). In each of 5 weeks, a speaker utters each isolated digit 0-9 in Thai language 10 times. Speech files from week 1-3 are used as reference templates and the rest are used for testing process. Hence, each speaker will have 30 references and 20 testing sounds per isolated digit. Three interesting experiments were set up as follows.

3.1 Experiments on Speech Features

Three speech features: LPCC, PFL, and MFCC were compared with 50 speakers in our speaker identification system. According to our previous experiment [10], 15-order coefficients give better performance when compared against 10-order for LPC and LPCC. For this reason, 15-order coefficients are conducted for the three features. DTW with 5-NN is chosen to be a pattern matching system.

The results on this experiment are illustrated in Table 1. MFCC has the best performance with 86.74% average identification rate and LPCC gives the worst rate at 84.61%. In addition, MFCC achieves the best identification rate (92.30%) for digit "5". PFL can represent speech characteristics better than LPCC as ever proven in [8].

Although, MFCC provided the best identification rate, some disadvantages are that more parameters must be set in order to optimize MFCC feature such as a number of filters in the bank, shapes of filters, frequency scale, and filter bandwidth. These parameters must be verified with experiments.

Digit	Identification Rate (%)		
	LPCC	PFL	MFCC
0	89.60	90.70	89.10
1	86.30	89.40	89.40
2	84.80	86.70	87.80
3	87.40	87.90	87.60
4	85.80	86.90	85.70
5	87.80	89.10	92.30
6	81.80	79.70	85.30
7	82.60	83.60	84.20
8	74.90	75.50	79.90
9	85.10	85.50	86.10
Average	84.61	85.50	86.74

Table 1: Identification rate results of LPCC, PFL, and MFCC.

3.2 Comparison among GMM, DTW, and MLP

As shown, MFCC is the most effective method of representing speech signals for our task. To continue the next experiment, MFCC is therefore defined as a fixed feature for comparing identification rate of GMM, DTW, and MLP. Similarly, the speakers, referenced sounds and testing sounds are the same as in the experiment in section 3.1. The aim is to compare the performance of these different identification engines using the same data and front-end processing.

For GMM experiment, each speaker was modeled by a 50 component GMM with nodal variances using 15-dimensional Mel frequency cepstral feature vectors. The initial model means and an identity matrix for the starting covariance matrix is consisted by randomly choosing 16 vectors (16 mixtures) from a speaker's training data. A variance limit of $\delta^2_{\min} = 0.01$ was used in training. A 50 iteration process was used with the EM algorithm for convergence of the likelihood function, which came from evaluating the summation values of 16 component densities weighted in each iteration. Because many parameters must be calculated, initial methods and parameters in this experiment were studied and referred from a previous successful paper [4]. DTW and MLP parameters have been evaluated with some of our preliminary experiments [3, 12] and succeeded for text-dependent speaker identification systems.

Figure 3 illustrates a line graph of the comparative results of GMM, DTW, and MLP for each digit with 50 speakers. The detailed results on each isolated digit and average identification rate are shown in Table 2. The GMM attains 87.54% average identification accuracy which is better than 86.74% for DTW and 82.34% for MLP on the isolated 0-9 digits. The identification rate result for almost all digits with GMM is higher than the ones obtained from the other two techniques. On the other hand, MLP gave the lowest identification rate of all digits, which can be explained that MLP tries to learn all training patterns even though some patterns have low

characteristics to represent their classes. These patterns certainly pull down the capability of recognition. However, with DTW almost all digits that have higher identification rates than GMM are short-vowel (their phonetics not contain ":" symbols), e.g. digit "6" and "7".

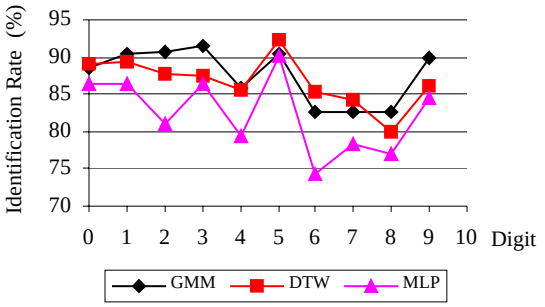

Figure 3: Comparative result for GMM, DTW, and MLP.

Digit	Phonetics	Identification Rate (%)		
		GMM	DTW	MLP
0	/su:n4/	88.60	89.10	86.30
1	/nvng1/	90.40	89.40	86.30
2	/s@:ng/	90.80	87.80	80.90
3	/sa:m4/	91.60	87.60	86.50
4	/si:1/	85.80	85.70	79.40
5	/ha:2/	90.40	92.30	90.10
6	/hok1/	82.60	85.30	74.20
7	/cet1/	82.70	84.20	78.20
8	/pa:t1/	82.60	79.90	77.00
9	/kao:2/	89.90	86.10	84.50
Average		87.54	86.74	82.34

Table 2: Identification rate results for GMM, DTW, and MLP

It also can be noted that the relative identification rate of each digits are the same for all 3 methods. Hence, the components of the speech signal (alphabet, tone, vowel, etc.) have an influence on the identification rate.

3.3 Experiment on concatenated digits

The last experiment is to enhance the identification rates with longer speaking-text duration. The top-3 identification rate isolated digits, which are chosen from the highest rate three digits indicated in table 2, are concatenated to form a new speaking-text. Evaluation is also performed for 50 speakers and again compared between GMM, DTW, and MLP with MFCC. The result is shown in Table 3. Each engine achieves a very high identification rate as expected, and the results still indicate the same relative effectiveness of each identification engine as seen in isolated digit experiments.

Recognition Engine	Top-3 Digit	Identification Rate (%)
GMM	"325"	99.00
DTW	"510"	98.70
ANN	"530"	97.30

Table 3: Identification rate results for GMM, DTW, and MLP on top-3 concatenated digits.

4. CONCLUSION

This paper has compared the three popular pattern matching on Thai text-dependent, closed-set speaker identification system. GMM attains the highest identification rate by comparing against DTW and MLP. Therefore, it can be concluded that the Gaussian components of a GMM represent some general speaker-dependent spectral shapes with effectively for modeling speaker identity. Furthermore, higher identification rates can be obtained with the longer speaking texts.

5. REFERENCES

1. Campbell, J.P., Jr. "Prolog to Speaker Recognition: A Tutorial", Proceedings of *IEEE*, Vol. 85, No. 9, p.1436-1462, September 1997.
2. K. Yu, J. Mason and J. Oglesby. "Speaker Recognition using Hidden Markov Modes, Dynamic Time Warping and Vector Quantisation", IEE Proc.-Vis. Image Signal Process, Vol. 142, No. 5, October 1995
3. Wutiwiwatchai, C., Achariyakulporn, V., and Tanprasert, C. "Text-dependent Speaker Identification using LPC and DTW for Thai Language", 1999 *IEEE* 10th Region Conference (TENCON'99), Vol.1, 1999.
4. Douglas, A. R., Richard, C. R. "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", IEEE Trans., speech, audio processing, Vol. 3, No. 1, p.72-83, January 1995.
5. Dodingtion, G.R. 1985. "Speaker Recognition-Identifying People by their Voices", Proceedings of *IEEE*, Vol. 73, No. 11, 1651-1664.
6. O'Shaughnessy, D. 1988. "Linear Predictive Coding. *IEEE* Pontentials", 29-32.
7. Furui, S. "Cepstral Analysis Technique for Automatic Speaker Verification. *IEEE* Trans. Acoust., Speech, Signal Processing, Vol. ASSP-29, 254-272, 1981.
8. Zilovic, M.S., Ramachadran, R. P., and Mammone, R. J. "Speaker Identification Based on the Use of Robust Cepstral Features Obtained from Pole-Zero Transfer Functions", *IEEE* Trans. Speech, Audio Processing, Vol.6, No.3, 260-267, 1998.
9. Furui, S. "Digital Speech Processing, Synthesis, and Recognition", New York and Basel: Marcel Dekker, Inc., 1989.
10. Sintupinyo, W., Dubey, P., Achariyakulporn, V., Sae-tang, S., Wutiwiwatchai, C., and Tanprasert, C. "LPC-based Thai Speaker Identification using DTW" Proceedings of 1999 NSTDA Annual Conference, Thailand, p. 238-246, 1999.
11. SNNS (Stuttgart Neural Network Simulator) User Manual, Version 4.1, University of Stuttgart, Institute for Parallel and Distributed High Performance Systems (IPVR), Report No.6/95
12. Sae-Tang, S., Tanprasert, C. "Feature Windowing-Based for Thai Text-Dependent Speaker Identification Using MLP with Backpropagation Algorithm", Proceeding of 2000 International Symposium on Circuits and Systems.