Voice Articulator for Thai Speaker Recognition

Shutinun Limpanakorn and Duangkaew Sawamiphakdi
Department of Computer Science, Faculty of Science and Technology, Thammasat University, Rangsit Campus, Klongluang, Pathum-thani 12121, Thailand.

Chularat Tanprasert
National Electronics and Computer Technology Center (NECTEC)
Ministry of Science, Technology and Environment
Bangkok Thai Tower Building, 11th floor,
108 Rangnam Road, Phyathai, Rachathewi, Bangkok 10400, Thailand

Abstract

Standard speaker recognition system employs a pre-processed form of an acoustic signal, which provides information about the distribution of signal energy across time and frequency. However, different signal representations may be employed, either as genuine alternatives to the acoustic representation, or as additional sources of information. Voice articulator encourages a viability and a potential of the speech signal representation especially in a Thai speaker recognition system. Applying the biometrical voice articulator additionally with a Backpropagation multilayered perceptron attains a high recognition accuracy. LPC and MFCC with several coefficient orders have been performed comparatively. The highest percentage of recognition accuracy with an efficient computational time is 97.24% belonging to the Bilabial articulator from the 16th coefficient order of MFCC.

Keywords: Thai speaker recognition system, Voice articulator, LPC, MFCC, MLP

1. Introduction

Speaker recognition in a recognition area of speech processing, is one of be biometric identification systems using voiceprint as a key. This process automatically recognizes a speaker who is speaking by using speaker-specific information included in speech waves [10,12]. This knowledge can be of benefit especially in the business area. Over the last few years, the concept of e-commerce has captured the attention of every major organization around the world. The ability to enable persons to complete transactions unattended while enabling organizations to process more transactions at lower cost, holds obvious appeal. While the adoption of e-commerce has been fueled by the rapid growth of the internet, data and security are also important issues of concern simultaneously in this information era, especially in the business world. Most business applications require people to identify themselves before completing a secure transaction which often uses a digital signature as a secure key. Since digital signature performs as a principle key to access valuable data, unauthorized persons try to imitate this kind of significant key. And although this traditional security key has revolutionized authentication techniques, it is still not quite safe in the current world.

With the addition of biometrics technology [6,11] using voice, retina, face or fingerprints, which are unique personal characteristics, security fears can be overcome. Immigration and Naturalization Service's Passenger Accelerated Service System (INSPASS), Canadian Passenger Accelerated Service System (CANPASS) and Port Passenger Accelerated Service System (PORTPASS) are current applications used in the United States of America which implement biometrics as identification characteristics.

Articulator capability, a significant addition to voice recognition has not been researched in Thai language but can also increase the identification performance by simply using personal voice and allows speaker

recognition to be widely used in the real world for security purpose.

Due to a realization of the principles of voice characteristics in the speaker identification system, the fundamental purpose of this paper is to research on which voice articulator is the best suitable organism for speaker recognition in Thai language.

2. Fundamental of voice articulator

Voiced speech is generated from air flow in the expiration phase leading from the lungs through the trachea, outward to the larynx, the pharynx and the mouth as shown in Figure 1.

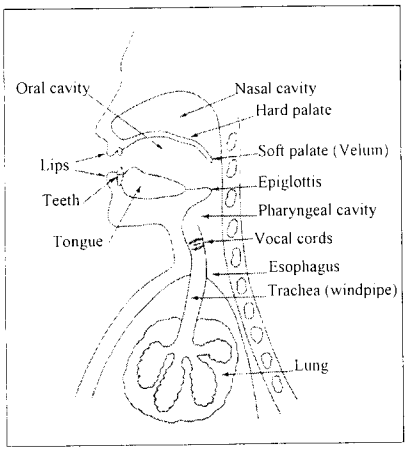

Figure 1. Speech producing organs

Linguistic information is interpreted from vibrations of the larynx. The tensed vocal cords with in the larynx are caused to vibrate by the air flow passing from the lungs, generated pulses go to the pharynx cavity, the mouth cavity and the nasal cavity depending on the position of the various articulators. Changing the larynx's way of vibration entails a change of voice quality. Individual speakers differ from each other in the formation of larynx mechanism whose vibration activity plays a major role in enabling listeners to identify individual voices. However various moods and emotions, affection or pleasure for example, can establish a different force in the larynx and hence to generate a different pulse wave with different voice quality. Air flowing during breathing out from the lungs is used in generating speech sound.

Sound generated from a vibration of the larynx is obstructed by some articulators as explained in Table 1, 2 and 3. Then, sound will be transformed into consonant and vowel form which we normally hear.

Table 1. Phonetic transcription of Thai consonant and vowels

Voice Characteristic	Articulator		
	Bilabial	Dental	Palatal
Plosive	/p/ /ph/ /b/	/t/ /th/ /d/	/c/ /k/ /ch/ /kh/
Nasal	/m/	/n/	/ng/
Lateral		/l/	
Roll		/r/	
Fricative	/f/	/s/	
Semi vowel	/w/		/j/
Close vowel Semi-close- open vowel Open vowel			Front Middle Back /i/ /i:/ / ɯ / / ɯ :/ /u/ /u:/ /e/ /e:/ / ɣ / / ɣ :/ /o/ /o:/ / ɛ / / ɛ :/ /ɔ/ /ɔ:/ /a/ /a:/

Table 2. Thai consonant

Phonetic	Thai consonant
/p/	ป
/t/	ต ฏ
/c/	จ
/k/	ก
/ph/	พ ภ ผ
/th/	ท ธ ฒ ฑ ถ ฐ
/ch/	ช ฌ ฉ
/kh/	ข ค ฆ
/b/	บ
/d/	ด ฎ
/m/	ม หม
/n/	น ณ หน
/ng/	ง หง
/l/	ล ฬ หล
/r/	ร หร
/f/	ฝ ฟ
/s/	ส ศ ษ ซ
/w/	ว หว
/j/	ย ญ หย หญ

Table 3.Thaivowels

Phonetic	Thai vowel
/i/	อิ
/i:/	อี
/ ɯ /	อึ
/ ɯ :/	อื
/u/	อุ
/u:/	อู
/e/	เอะ
/e:/	เอ
/ ɣ /	เออะ
/ ɣ :/	เออ
/o/	โอะ
/o:/	โอ
/ ɛ /	แอะ
/ ɛ :/	แอ
/ɔ/	เอาะ
/ɔ:/	ออ
/a/	อะ
/a:/	อา

Articulators are the points of major closure in the vocal tract during its articulation. They can be categorized into three main placements, Bilabial, Dental and Palatal :- First, the Bilabial articulator, is a constriction between the lips and the lips closed against the upper teeth. Second, the Dental articulator, is formed between the tongue and the upper teeth. And third, the Palatal articulator, Palatal constriction is made between the tongue and the hard palate.

3. Thai Speaker Recognition

The proposed Thai speaker recognition system uses a text-dependent speaker recognition to simplify a complex speaker recognition system which can improve the accuracy of the speaker recognition task by studying the input data system.This system consists of three main phrases as shown in Figure 2 :-

1. Speech pre-processing phase,
2. Feature extraction phase, and
3. Recognition phase which consists of a training module and a testing module.

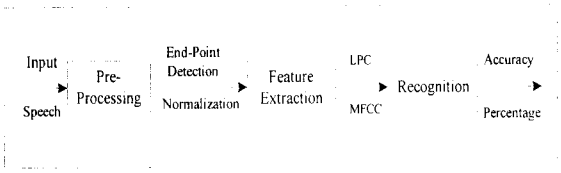

Figure 2. Speaker recognition data flow

3.1 Speech pre-processing phase

Pre-processing step is the first phase of a speaker recognition task to prepare an incoming speech waveform into a suitable format before extracting useful information in the following feature extraction phase.

Pre-processing phase consists of 3 main modules as shown in Figure 3 :- a 20-dB end-point detection, a 25-milliseconds linear time alignment normalization and a 256-byte frame windowing with 128-byte overlapping have been performed.

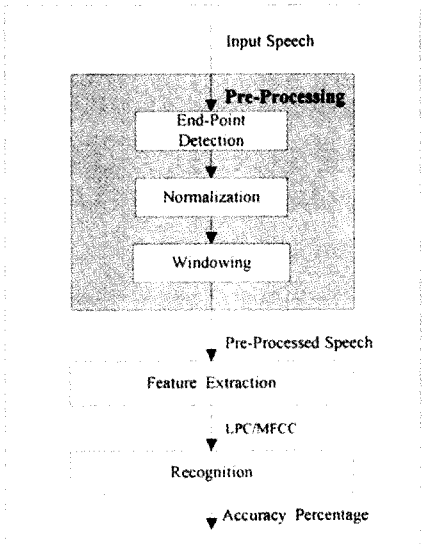

Figure 3. Speech pre-processing diagram

3.1.1 End-point detection

End-point detection is a detecting process of each spoken utterance period by determining where that particular utterance starts and ends and also decreasing additive presence of noise. Figure 4.a and Figure 4.b are examples of the utterance /ch/ before and after passing through the end-point detection process.

Figure 4.a /ch/ before processing the end-point detection phase

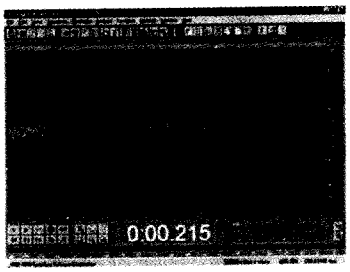

Figure 4.b /ch/ after processing the end-point detection phase

3.1.2 Normalization

After the incoming speech waveform passes through the end-point detection process, lengths of each utterance are not the same. This variation arises from difference in the recording and speakers themselves [5,10,14]. Speaker variability is a major source of performance degradation in speaker recognition. Normalization is used to adjust the length of each utterance to be the same size by modifying the spectral representation of incoming speech waveform to reduce variability between speakers. Figure 5.a and Figure 5.b are the utterance /ch/ before and after applying the normalization technique as an example.

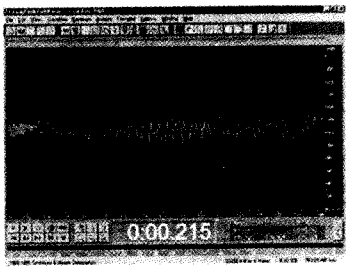

Figure 5.a /ch/ before processing the normalization phase (length = 21.5 ms.)

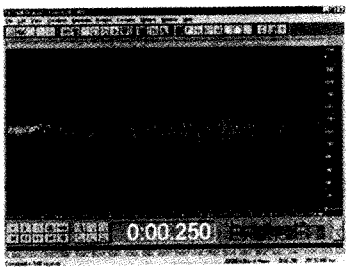

Figure 5.b /ch/ after processing the normalization phase (length = 25.0 ms.)

3.1.3 Windowing

Windowing is a process to divide a voice signal into small pieces of frames to make a nonstationary voice signal, whose attributes are variant from time to time, stable within a short period of time. Then applying frame overlapping additional with the Hamming function $W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right)$ where N is the number of data in each frame and n = 0,1, ... , N-1, to ensure that no data at the frame border is lost during frame cutting.

To find out the best alphabets for Thai speaker recognition task, the 8th coefficient order of LPC and MFCC of the three observers are used to train a neural network. Their recognition accuracy are shown in Table 4.

Table 4. Percentage of recognition accuracy based on each articulator

Articulator	Alphabet	Trainer1	Trainer2	Trainer3	Avg. per person
Bilabial (ริมฝีปาก)	/p/	91.14%	97.31%	70.99%	86.48%
	/ph/	80.79%	88.88%	88.96%	86.21%
	/b/	73.34%	93.96%	89.59%	85.63%
	/m/	93.73%	69.32%	90.66%	84.57%
	/f/	94.47%	96.29%	99.18%	96.65%
	/w/	82.67%	63.72%	92.08%	79.49%
Dental (ฟัน)	/t/	95.99%	78.49%	81.51%	85.33%
	/th/	99.09%	88.97%	87.49%	91.85%
	/d/	96.52%	67.10%	97.76%	87.12%
	/n/	95.88%	84.33%	66.02%	82.07%
	/l/	92.08%	82.67%	70.04%	81.60%
	/r/	98.59%	66.86%	93.91%	86.45%
	/s/	97.55%	63.72%	94.03%	85.10%
Palatal (เพดานอ่อน-เพดานแข็ง)	/c/	94.23%	48.99%	72.06%	71.76%
	/ch/	86.41%	92.47%	51.26%	76.71%
	/k/	94.56%	80.57%	93.74%	89.63%
	/kh/	68.25%	73.98%	16.52%	52.92%
	/ng/	92.08%	70.04%	41.30%	67.80%
	/j/	65.23%	63.08%	82.67%	70.33%

Using percentage of recognition accuracy of the Bilabial articulator, alphabet /p/, /ph/, /b/ and /f/ give 86.48%, 86.21%, 85.63% and 96.65% respectively which are greater than 84.57% from alphabet /m/ and 79.49% from alphabet /w/. Therefore, /p/, /ph/, /b/ and /f/ have been chosen to be the best representatives of the Bilabial articulator in a continuous experiment. Applying the same criteria on the Dental and Palatal articulator to choose their representatives, Table 5 is generated.

Table 5. Selected alphabets for our test

Articulator	1st Alphabet	2nd Alphabet	3rd Alphabet
Bilabial	/p/	/ph/	/f/
Dental	/t/	/d/	/r/
Palatal	/c/	/ch/	/j/
Palatal-Dental-Bilabial (mix)	/k/	/th/	/b/

Four meaningful Thai sentences are defined to cover all categories specified in Table 5 including with their phonetics representation as shown in Table 6.

Table 6. Phonetics of each utterance based on a voice articulator in a text dependent recognition system

Voice Articulator	Thai Sentences	Phonetics
Bilabial	เปิดเพลงฟัง	[*pɣ ˧˩ d*], [*phleŋ*] and [*faŋ*]
Dental	ต้นคาวเรือง	[*t ˧˥ n*], [*da : w*] and [*rɯaŋ*]
Palatal	แจกโชคใหญ่	[*cɛ ˧˩ k*], [*cho ˧˩ k*] and [*jàj*]
Palatal-Dental-Bilabial	กาแฟอร์โบ	[*ka :*], [*thɣ :*] and [*bo :*]

The most significant factor that affects a speaker recognition's performance is a variability in voice characteristics. This variation arises from the speaker him/herself, from differences in recording, media transmission and noises. Therefore, this paper proposes to have 25 repetitions of 19 persons by dividing 75 percentage of data to be a training set and the last 25 percentage of data as a testing set. Total number of samples used in this experiment is 5,700 which comes from 19 (persons) * 4(sentences) * 3(words) * 25 (repetitions).

3.2 Feature extraction phase

Feature extraction is a process to obtain an amount of necessary data to process in the recognition stage. Thereby, Linear Prediction Coefficient (LPC) and Mel Frequency Cepstral Coefficient (MFCC) which are the prominent features from spectral envelope will be used as the main features [2,4,10,13,14,16,19,21].

LPC is a parametric analysis model which assumes that parameters are representing a vibration of wave within the vocal tract and a speech at time n can be approximated as a linear combination of the past p speech samples. Whereas MFCC is a nonparametric model that describes speech as a representative of spectrum envelope in form of the formant frequencies and pitch harmonics [4].

Figure 6 shows over all tested features. Speech passed from the pre-processing phase will continue to process in the feature extraction phase. Two copies of them have been utilized, the first copy is extracted through LPC model whereas the second one is extracted via MFCC model. Either in LPC or MFCC model, four extracted-feature files are generated for the 8^{th}, 12^{th}, 16^{th} and 20^{th} coefficient order, respectively.

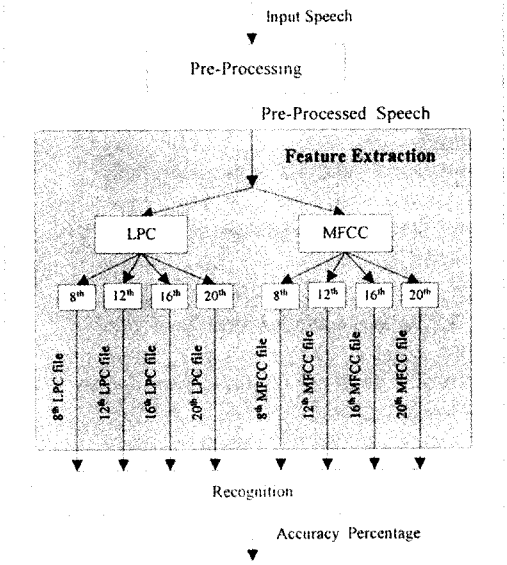

Figure 6. Feature extraction diagram

Some LPC and MFCC features have been generated in Table 7 for an example.

Table 7. LPC and MFCC features of utterance /ch/ based on each coefficient order

Feature	8^{th} coeff. order	12^{th} coeff. order	16^{th} coeff. order	20^{th} coeff. order
LPC				
MFCC				

3.3 Recognition phase

Recognition phase is a process to identify a speaker by using extracted features from the feature extraction procedure. A neural network based on backpropagation learning algorithm is a computationally efficient method and is one of the best supervised learning rule artificial neural network models which has a potential to adapt itself by using a powerful Log-Sigmoid activation function with a gradient descent weight technique and propagating error backward to suit for an unknown pattern recognition as described in Figure 7 [2,13,18].

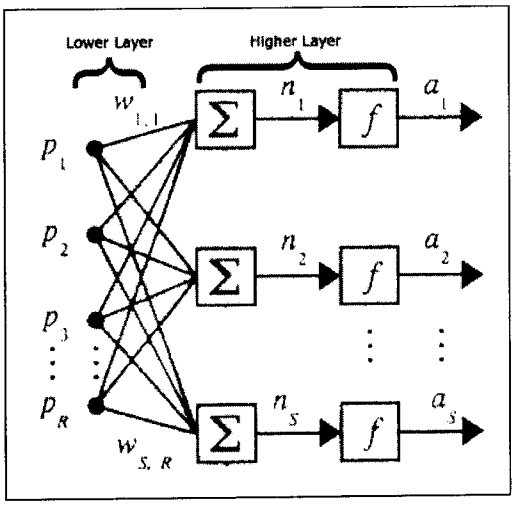

Figure 7. Backpropagation neural network

Therefore, this proposed speaker recognition system uses backpropagation network as a speaker recognition engine via SNNS (Stuttgart Neural Network Simulator) which is divided into two modules, a training module and a testing module, as depicted in Figure 8, and its architecture is also shown in Table 8.

3.3.1 Training Module

The training module is used part for training a neural network by presenting extracted features as its inputs with a corresponding target output. Since a network is learnt for a particular coefficient order of each utterance, either LPC or MFCC features, thus a training set has its input composed of seventy five percent of recorded data or eighteen of twenty-five patterns of each utterance from all speakers and will be learnt until a threshold, for example MSE < 0.001, is met.

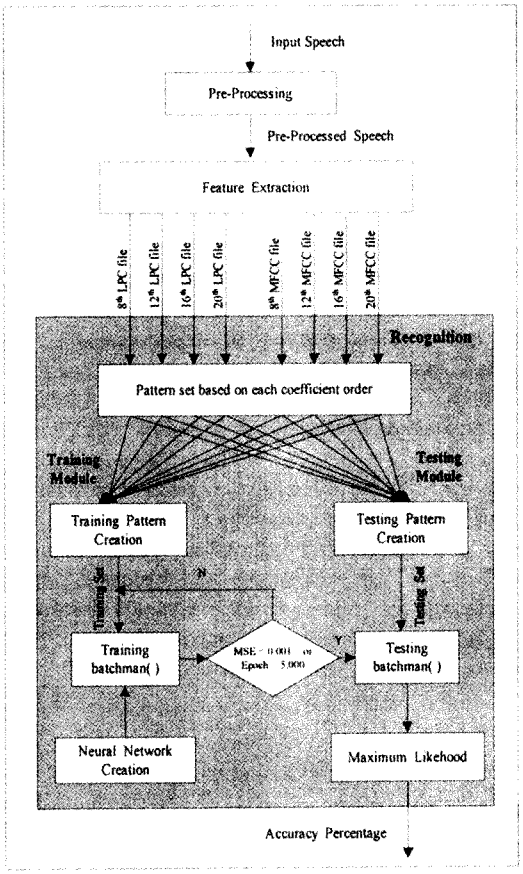

Figure 8. Proposed recognition diagram

Table 8. Number of neurons in each layer of different networks

LPC/MFCC Coefficient	Input Node	Hidden Node	Output Node
8	128	50	19
12	192	60	19
16	256	70	19
20	320	80	19

3.3.2 Testing module

When the training module is stopped,the system recognition accuracy is tested by using the remaining twenty five percentage or seven of twenty-five patterns based on each utterance.

Not only a MLP with backpropagation is applied in this recognition phase, but maximum likelihood is also performed. Maximum likelihood criteria is one of the most famous and

simple decision rules used in this field. This rule is applied as a criteria in a testing process to recognize a speaker by choosing the highest recognition accuracy among its neighbors. A speaker whose output node returns a maximum probability has potential to be the identified speaker. Additionally, the best voice articulator which is properly used in a speaker recognition should be the one which gives the highest average percentage of recognition accuracy.

4. Experimental Results

Applying all processes stated in the previous section, speech signal has been pre-processed and has extracted features in order to gain some useful characteristics for a backpropagation network to identify a speaker. An average percentage of recognition accuracy has been summarized in Figure 9 and Figure 10 by showing the relationship between coefficient order in x-axis and percentage of accuracy in y-axis and Table 9 by showing a relationship between voice articulator in columns and percentage of accuracy in rows. The highest obtained recognition accuracy, from the proposed speaker recognition system, is 97.24%.

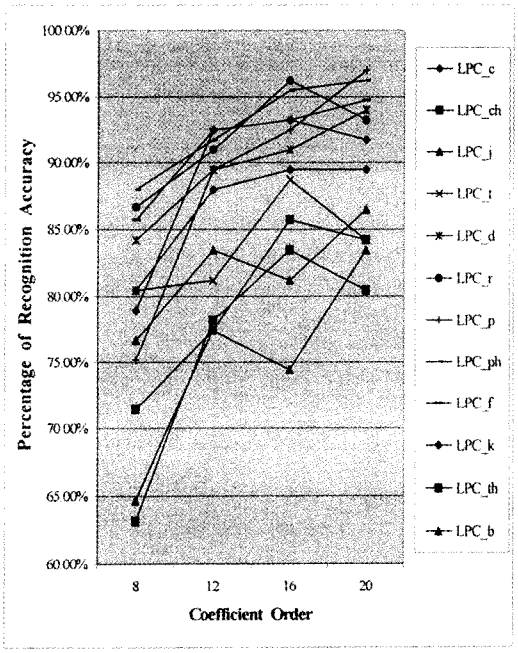

Figure 9. The 8th, 12th, 16th and 20th coefficient order of LPC for all utterances

Figure 10. The 8th, 12th, 16th and 20th coefficient order of MFCC for all utterances

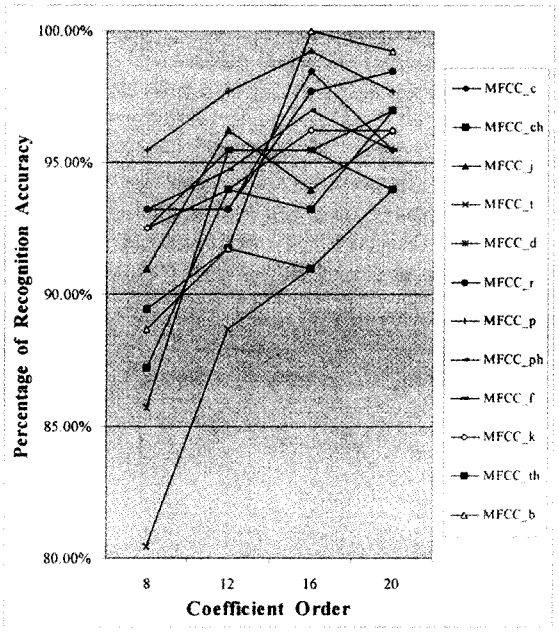

Table 9. Recognition accuracy on the 8th, 12th, 16th and 20th coefficient order of LPC and MFCC of each articulator

Articulator		Bilabial	Dental	Palatal	Mix	
Alphabet		/p/ /ph/ /f/	/t/ /d/ /r/	/c/ /ch/ /j/	/k/ /th/ /b/	
LPC	8	82.96%	82.89%	68.92%	76.19%	
	12	91.23%	87.22%	82.71%	82.96%	
	16	93.73%	91.17%	83.71%	85.46%	
	20	95.99%	90.48%	85.21%	86.72%	
Avg. Accuracy on LPC		90.98%	87.94%	82.83%	80.14%	85.47%
MFCC	8	93.73%	86.22%	89.47%	91.23%	
	12	95.99%	92.48%	93.23%	93.73%	
	16	97.24%	94.74%	96.49%	94.49%	
	20	95.74%	96.49%	97.49%	95.24%	
Avg. Accuracy on MFCC		95.68%	92.48%	93.67%	94.17%	94.00%

Additionally, these experimental results demonstrate that :-

1. Different voice articulators give different performances in the proposed speaker recognition system by contributing 94% and 85.47% averagely for MFCC and LPC model respectively.
2. Among three categories of voice articulators, the Bilabial articulator gives the best recognition accuracy average for either MFCC model, 95.68% or LPC model, 90.98%.
3. The higher coefficient order increases either on LPC or MFCC, a higher percentage of accuracy is generated correspondingly.
4. MFCC is the more powerful extracted feature from speech envelope comparing with another extracted feature, LPC.
5. Even though the highest percentage of recognition accuracy is obtained from the 20^{th} coefficient order of MFCC produced from three articulators, 97.49%, its computational time is unacceptable. It takes around twice as much training time as the 16^{th} MFCC of utterances only generated from the Bilabial articulator which returns 97.24% accuracy. This is not a significant difference. Therefore, the 16^{th} MFCC is more suitable in the practical implementation and is the best utterance to identify speakers in the proposed speaker recognition system.

5. Conclusion

Speaker recognition is correlated with the physiological and behavioral characteristics of the speech production system of each speaker. These characteristics exist both in the spectral envelope, vocal tract characteristics, and in the articulator characteristics of speech.

This paper proposes some new knowledge on voice articulator consequently increasing recognition accuracy in a Thai speaker recognition system by implementing the 19-speakers of 3-articulator-speech experiments from a pre-processing, a feature extraction and a recognition phase. Applying end-point detection, normalization and windowing in the pre-processing phase prepares better speech frame by eliminating unuseful data from an entire utterance. Then, the 8^{th}, 12^{th}, 16^{th} and 20^{th} coefficient order of LPC and MFCC which are the most prominent features have been extracted in the feature extraction phase. Finally, a neural network with backpropagation learning algorithm is generated for training and testing a target speaker by using the maximum likelihood criteria of recognition accuracy.

The result of this experiment is evident for enlarging more knowledge in the speaker recognition researched area by using voice articulator. The useful knowledge unveiled in this paper are stated as follows:-

1. Voice articulator plays a principle role in the speaker recognition system.
2. Bilabial is the best articulator of Thai speaker recognition system which provides the highest percentage of recognition accuracy.
3. Comparing between LPC and MFCC model with four coefficient orders applied, the 16^{th} coefficient order of MFCC is the best extracted feature with an efficient computational time.

6. Reference

[1] Kanjana Nakasakun, Thai Sound System. Litcrature Project, Faculty of Arts, Chulalongkorn University, 1998.

[2] Chularat Tanprasert , Vasin Sintupinyo, Premnath Du-Bae, Sutat Sae-Tang, Varin Achariyakulporn, and Chai Wudiwatchal, Thai Speaker Identification by LPC and DTW, NECTEC Journal (March-June), pp.24-35, 2000.

[3] Varin Achariyakulporn, Chai Wuttiwatchal, and Chularat Tanprasert Thai Speaker Identification System by Dynamics Time Wraping, NECTEC Journal (Junc-October), pp.108-118, 2000.

[4] Oppenheim, Alan V., Application of Digital Signal Processing, Prentice-Hall, 1978.

[5] Becchetti, Claudio, and Ricotti, Lucio Prina. Speech Recognition, Rome, John Wiley Publisher, 1999.

[6] Campbell, Joseph P., Alyea, Lisa A. and Dunn, Jeffrey S., Biometrics security: Government Application and Operations, Biometrics Consortium, CTST, 1996.

[7] Chularat Tanprasert, and Varin Achariyakulporn, Comparative Study of GMM, DTW, and ANN on Thai Speaker Identification System, Proceedings of 6^{th}

International Conference on Spoken Language Processing, pp. 234-237, 2000.

[8] Fry, D. B., The Physics of Speech. Cambridge, Cambridge University Press, 1979.

[9] Fu, Li Min., Neural Networks in Computer Intelligence, Singapore : McGraw-Hill, 1994.

[10]Furui, Sadaoki, Recent Advances in Speaker Recognition, Audio and Video Based Biometrics Person Authentication, 1997.

[11]Jain, Anil, Hong, Lin, and Pankanti, Sharath, Biometrics: Promising Frontiers for Emerging Identification Market, Communication of the ACM (February), pp. 91-98, 2000.

[12]Li, Qi, Juang, Biing Hwang, Lee, Chin Hui, Zhou, Qiru, and Soong, Frank K., Recent Advancements in Automatic Speaker Authentication, IEEE Robotics and Automation, 1999.

[13]Paoloni, A., Ragazzini, S. and Ravaioli, G., Text Independent Speaker Verification Using Multiple State Predictive Neural Networks, Audio and Video Based Biometrics Person Authentication, 1997.

[14]Rabiner, Lawrence, and Juang, Biing Hwang, Fundamentals of Speech Recognition, New Jersey : Prentice-Hall International, 1993.

[15]Rowden, Chris, Speech Processing, Master's Thesis, Department of Electronic Systems Engineering, Faculty of Engineering, University of Essex, 1992.

[16]Sambur, Marvin R., Speaker Recognition Using Orthogonal Linear Prediction, IEEE Transaction Acoustic Speech and Signal Process, ASSP-24, 1976.

[17]SNNS (Stuttgart Neural Network Simulator) User Manual Version 4.1, Institute for Parallel and Distributed High Performance Systems (IPVR), University of Stuttgart, 1995.

[18]Sutat Sae-Tang, and Chularat Tanprasert, Feature Windowing-Based for Thai Text-Dependent Speaker Identification Using MLP with Backpropagation Algorithm, Proceedings of 2000 IEEE International Symposium on Circuits and Systems (May), pp. 28-31, 2000.

[19]Thomas P. Barnwell III, Kambiz Nayebi and Craig H., Richardson., Speech Coding, Georgia : Digital Signal Processing Laboratory, 1996.

[20]Waters, Gill, Speech Processing, England : McGraw-Hill, 1991.

[21]Wassner, Hubert, and Chollet, Gerard, New Time-frequency Derived Cepstral Coefficients for Automatic Speech Recognition, IDIAP, EUSIPCO96, 1996.