Text-dependent Speaker Identification Using Neural Network On Distinctive Thai Tone Marks*

Chularat Tanprasert[1], *Chai Wutiwiwatchai*[2], *Sutat Sae-tang*[3]
Software and Language Engineering Laboratory,
22nd floor, Gypsum Metropolitan Tower, Sri-Ayudhaya Rd., Bangkok 10400, THAILAND
E-mail: [1]chulak@nectec.or.th, [2]cwutiwiw@notes.nectec.or.th, [3]sutat@notes.nectec.or.th

ABSTRACT-- This paper presents a neural network based text-dependent speaker identification system for Thai language. Linear Prediction Coefficients (LPC) are extracted from speech signal and formed feature vectors. These features are fed into multilayer perceptron (MLP) neural network with backpropagation learning algorithm for training and identification processes. Five Thai tone marks are considered very closely in choosing the sentences in order to achieve the best speaker identification accuracy. Five speaking texts with each Thai tone and a mixed tone text are comparatively experimented. Average identification rate on 9 speakers achieves above 95% when using mixed tone text, and poor results occur with middle and low tone texts, which usually cause vagueness or unclear voices.

KEY WORDS -- Text-dependent speaker identification, Linear prediction coefficients, Neural networks, Thai tone.

บทคัดย่อ--บทความฉบับนี้นำเสนอระบบระบุผู้พูดสำหรับภาษาไทยแบบกำหนดคำพูด โดยใช้โครงข่ายประสาทเทียมเป็นระบบในการจดจำ ในขั้นต้นสัญญาณเสียงจะถูกนำมาผ่านกระบวนการสกัดค่าลักษณะสำคัญ ซึ่งคือค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น (Linear Prediction Coefficients) และใช้เป็นอินพุตสำหรับโครงข่ายประสาทเทียมประเภทเพอเซปตรอนหลายชั้น (Multilayer Perceptron) โดยใช้กระบวนการเรียนรู้แบบแพร่กระจายกลับหลัง (Backpropagation) จุดมุ่งหมายของการทดลองเพื่อพิจารณาผลกระทบของการใช้เสียงวรรณยุกต์ที่แตกต่างกันต่อการระบุผู้พูดจึงกำหนดให้ใช้ประโยคในการพูด 6 ประโยค 5 ประโยคแรกแต่ละประโยคมีเสียงวรรณยุกต์เดียวจาก 5 ระดับเสียงวรรณยุกต์ในภาษาไทย ส่วนประโยคสุดท้ายมีทั้ง 5 เสียงวรรณยุกต์ผสมกัน การทดลองเปรียบเทียบการพูดประโยคแต่ละประโยคเพื่อระบุผู้พูดจำนวน 9 คน ปรากฏว่าเมื่อใช้ประโยคที่มีเสียงวรรณยุกต์ผสมจะได้ผลการระบุผู้พูดสูงที่สุดซึ่งได้ผลมากกว่า 95 เปอร์เซ็นต์ และผลการระบุผู้พูดต่ำที่สุดเมื่อใช้ประโยคที่มีเสียงวรรณยุกต์เอกซึ่งเป็นเสียงที่ต่ำและไม่ชัดเจน

คำสำคัญ -- การระบุผู้พูดแบบขึ้นกับกำหนดคำพูด, ค่าสัมประสิทธิ์การประมาณพันธะเชิงเส้น, โครงข่ายประสาทเทียม, วรรณยุกต์ไทย

1. Introduction

Speaker recognition can be separated into two groups [1,5]; speaker verification which has main task in verifying speaking person, and speaker identification which tries to answer who is speaking. Speaking text dependency is an important limitation of these systems. On text-dependent system, the first essential procedure is to select some speaking texts that can cover speaker's characteristics as much as possible for the recognition process. This task is language dependent requiring some comprehensive researches.

Thai language consists of 44 consonants, 18 vowels, 10 Thai digits, and 5 tonal levels or tone marks. Tone marks, different from intonation, can be added in Thai words and caused different meaning. This characteristic of Thai language is experimented carefully to consider its effect. Thai tone mark is an outstanding feature like tone marks in Mandarin or Korean. Five Thai tonal levels can be classified into dynamic and static tones. Dynamic

*This article is a reprint of the article appeared in the Proceedings of IEEE-INNS International Joint Conference on Neural Network, July, 1999.

tones, rising and falling tones, have a great change of fundamental frequency with respect to duration, while the static tones, low, middle, and high tones, are quite stable. Figure 1 [6] illustrates a graph of fundamental frequency of each tone plotting with average time duration.

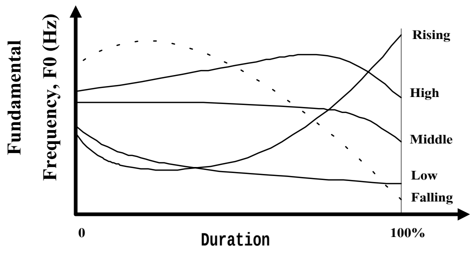

Figure 1. Fundamental frequency curve with respect to speech duration for each Thai tone.

Speaker identification system can conclusively be modeled as shown in Figure 2 [2]. Features are some quantities, which are extracted from preprocessed speech and can be used to represent the whole speech signal. Two sets of features consisted of spectral and prosodic features are proposed in many researches. Many works have concluded that the spectral based features such as Linear Prediction Coefficients (LPC), Cepstral coefficients, and their derivatives, have more efficient than prosodic-based features like fundamental frequency, formant frequency, and speech energy [3-5]. Main reasons are easiness in extraction and their consistency over several variations. Also, there are many recognition system proposed for speaker recognition such as a well-known nonlinear time-aligned techniques called Dynamic Time Warping (DTW) [1,4], Vector Quantization (VQ) [12], Artificial Neural Network (ANN) [10], and statistical Hidden Markov Model (HMM) [1,5]. What technique should be used depends on some factors in the implemented system.

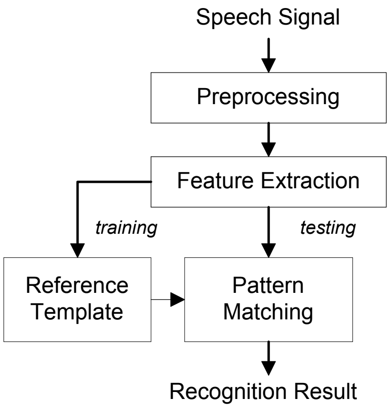

Figure 2. Speaker recognition model

In this paper, time duration of speaking voices is normalized via a time normalization technique. Hence, equal number of features, LPC is extracted from equal time-duration speech and can suitably form the input of backpropagation neural network, which has an essential limitation on number of input but has a strength of nonlinear mapping or recognizing [8].

2. Proposed Speaker Identification System

In the proposed system, digital speech signal is passed through a preprocessing procedure, which performs energy-based endpoint detection and time normalization. LPC features are then extracted from preprocessed speech and the input vectors have passed through the backpropagation learning algorithm with multilayered perceptron network for both training and evaluation processes. Figure 3 shows overall model of the proposed system.

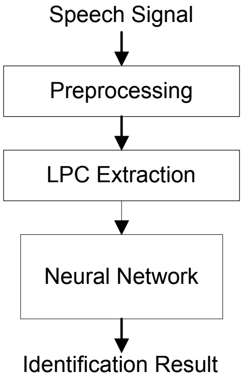

Figure 3. Overall proposed speaker identification model

2.1 Linear Prediction Coefficient (LPC)

A well-known speech coding strategy is to assume that each speech sample can be estimated from linear combination of p-previous speech samples as described in the following equation.

$$\hat{X}(n) \approx \sum_{k=1}^{p} a_k X(n-k) \qquad (1)$$

where $X(n)$ is a speech sample n, a_k's are p-order linear prediction coefficients (LPC), which can represent a speech signal. The way to find value of a_k is to minimize an error between an original speech sample and an estimated speech signal as shown in equation 2 by differentiation with each a_k.

$$E_n = \sum_m \left[\bar{X}_n(m) - \sum_{k=1}^{p} a_k X_n(m-k) \right]^2 \qquad (2)$$

Equation 2 forms an error of the whole speech frame instead an error of each sample. N denotes index of speech frame and m denotes index of speech sample within frame n. P equations derived from the differentiation of error equation are then solved to obtain the values of a_k by Levinson-Durbin algorithm [7].

2.2 Multilayer Perceptron (MLP) with Backpropagation Learning Algorithm

In this paper, we have chosen to use a backpropagation neural network [8,9,11] since it has been successfully applied to many pattern classification problems including speaker recognition [10] and our problem has been considered to be suitable with the supervised rule.

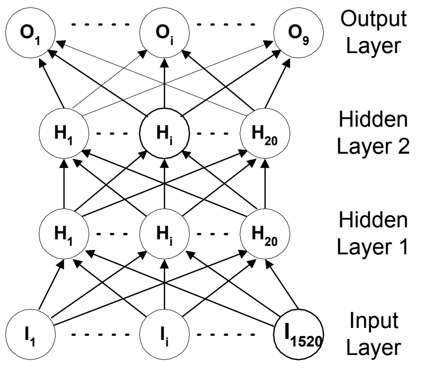

Figure 4. The structure of MLP

MLP neural network we used consists of four layers; one input layer, two hidden layers and one output layer. The structure of the backpropagation neural network is shown in Figure 4. The first layer has 1,520 input neurons (152 frames x 10 LPC-orders) which are fully connected to the first hidden layer. The two next hidden layers consist of 20 neurons per layer. The last layer is the output layer consisting of 9 neurons which one output neuron represented one speaker. All four layers are fully feed-forwarded.

3. Experimental Results

Our experiments were conducted on six sentences. Each sentence consists of five words. The first five sentences contain the same word in a tonal level of middle, low, falling, high and rising, respectively. The last sentence contains one word from each tonal level. So the last sentence has five different tonal levels. All six sentences used in our experiments are shown in the Table 1.

In the experiment, we recorded in an office environment at 11.025 kHz sampling rate, 16 bit and single channel. The average lengths of each sentence are 2.5 second. The utterance is obtained from 9 local speakers (4 females and 5 males). Each speaker was asked to pronounce each sentence ten times per day for two days. Therefore, the total utterances are 180 utterances for each sentence. False starts, pauses, and background noise are manually controlled in this database. The data were randomly divided into two sets, with 135 utterances in the training set and 45 utterances in the test set.

- Training set consists of 8 and 7 times from the first and second day, respectively. Total is 135 utterances (15 x 9).
- Test set consists of 2 and 3 times from the first and second day, respectively. Total is 45 utterances (5 x 9).

Each utterance was pre-emphasized, time normalized, and performed the 10th order autocorrelation analysis. The results were used to find the LPC feature vector. The LPC analysis was performed for every 20-ms speech frame with 5-ms overlapping.

Table 1. Six sentences in our experiments.

No.	Sentences	Phonetics	Tone mark
1	เอ-เอ-เอ-เอ-เอ	/?e:0 ?e:0 ?e:0 ?e:0 ?e:0/	Middle
2	เอ่-เอ่-เอ่-เอ่-เอ่	/?e:1 ?e:1 ?e:1 ?e:1 ?e:1/	Low
3	เอ้-เอ้-เอ้-เอ้-เอ้	/?e:2 ?e:2 ?e:2 ?e:2 ?e:2/	Falling
4	เอ๊-เอ๊-เอ๊-เอ๊-เอ๊	/?e:3 ?e:3 ?e:3 ?e:3 ?e:3/	High
5	เอ๋-เอ๋-เอ๋-เอ๋-เอ๋	/?e:4 ?e:4 ?e:4 ?e:4 ?e:4/	Rising
6	เอ-เอ่-เอ้-เอ๊-เอ๋	/?e:0 ?e:1 ?e:2 ?e:3 ?e:4/	All 5 tones

Two experiments were conducted. The first one was performed by using backpropagation network on the training and testing sets. The neural network simulator used in our experiments is SNNS (Stuttgart Neural Network Simulator) [9]. The second experiment was conducted by applying the Euclidean distance on the mean of distances of 20 utterances of each speaker on each sentence.

The identification rates of all six training sentences by neural network are all approaching 100%. The identification rates in percentage on testing data between

*This article is a reprint of the article appeared in the Proceedings of IEEE-INNS International Joint Conference on Neural Network, July, 1999.

the backpropagation network and the Euclidean distance are shown in Table 2. It is found that the ANN produces the best accuracy with a mixed five tone sentence at above 95% and give lowest average identification rate with sentence of middle tonal level at 68.89%. Meanwhile, the Euclidean distance gives the corresponding same result with ANN. The mixed tone sentence gives the best performance at 82% identification rate, while the worst result was obtained from sentence with middle tone mark at only 65%.

Table 2. Comparison of the identification rate between backpropagation network and Euclidean distance.

Sentence No.	**Identification rate (%)**	
	Neural Network	**Euclidean Distance**
1	68.89	65.00
2	77.78	69.00
3	88.89	74.00
4	82.22	73.00
5	86.67	80.00
6	95.56	82.00

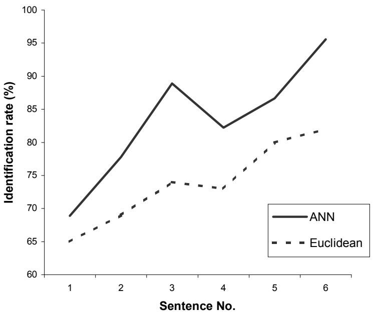

Figure 5. The comparison of the identification rates between ANN and Euclidean distance.

In Figure 5, the identification rates in percentage of ANN are better than ones of the Euclidean distance, but the best sentence of each technique are alike, while the sentence with middle tonal level gives the lowest identification rate in both techniques also. Therefore, the experimental results have shown that tonal levels of Thai tone marks are important to Thai speaker identification system. Furthermore, the sentence with several tone marks produces a better accuracy than a sentence with single tone mark with the same length of words.

Further experiment has performed with ten isolated Thai digits (0-9) on 20 speakers. Voices of each digit are recorded 10 times per week for 4 weeks, 7 times are used for training and the rest are used for testing. Table 3 illustrates phonetic of Thai digits, their tonal levels, and identification results.

Table 3. Identification results using speaking text of isolated Thai digits

Thai digit	Phonetics	Tone mark	Identification rate (%)
0	/su:n4/	Rising	80.00
1	/nvng1/	Low	80.83
2	/s@:ng4/	Rising	82.08
3	/sa:m4/	Rising	82.08
4	/si:1/	Low	79.17
5	/ha:2/	Falling	84.58
6	/hok1/	Low	60.42
7	/cet1/	Low	78.33
8	/pa:t1/	Low	70.83
9	/kao:2/	Falling	84.17

Larger number of speakers, and also shorter speaking text duration may normally pull down the performance of identification system as shown in Table 3. However, the results still confirm our conclusion on tone effect. Falling tone such as digit "5" and "9", and rising tone such as digit "2" and "3", can give quite high identification rates, while for example; digit "4", "6", and "8" conduct poor results. Another interesting notice is the worst identification obtained by digit "6", which may be caused by not only a low tone digit, but also a very short utterance digit.

Future researches of Thai text-dependent speaker identification system will be deeply investigated on the suitable length of speaking text, Thai alphabets and vowels, and other preprocessing and identification techniques.

4. Conclusion

LPC speech feature and a neural network with backpropagation training algorithm are appropriate to use for a Thai text-dependent speaker identification system on distinctive Thai tone marks. The best average identification rate is over 95% on nine Thai native speakers with ANN. In the comparison between ANN and Euclidean distance as an identification engine, ANN shows a great powerful nonlinear recognition performance. To improve system capability, mixed tone speech in appropriate length of speech duration should be selected for speaking sentence since it can cover more personal characteristics than using each tone in all utterances.

References

[1] J. P. Campbell, Jr., "Prolog to Speaker Recognition: A Tutorial", *Proceedings of IEEE*, Vol. 85, No. 9, p. 1436-1462, September 1997.

[2] W. Sintupinyo, P. Dubey, S. Sae-tang, V. Acahriyakulporn, C. Wutiwiwatchai, and C. Tanprasert, "LPC-based Thai Speaker Identification using DTW", Proceedings of 1999 *NSTDA Annual Conference*, Thailand, p. 238-246, March-April 1999. (in Thai)

[3] R. J. Mammone, X. Zhang, R. P. Ramachandran, "Robust Speaker Recognition, A Feature-based Approach", *IEEE Signal Processing Magazine*, p. 58-71, September 1996.

[4] G. R. Doddington, "Speaker Recognition-Identifying People by their Voices", *Proceedings of IEEE*, Vol. 73, No. 11, p. 1651-1663, November 1985.

[5] S. Furui, "Digital Speech Processing, Synthesis, and Recognition", New York and Basel: Marcel Dekker, Inc, 1989.

[6] Thubthong N., “A Thai Speech Recognition System Based on Phonemic Distinctive Features”, *Thesis of the Master Degree of Science Graduate School, Chulalonkorn University*, Thailand, 1995.

[7] C. Wutiwiwatchai, "Thai Polysyllabic-word Recognition using Neural Network and Fuzzy technique", *Thesis of the Master Degree of Electrical Engineering, Chulalongkorn University*, Thailand, 1997.

[8] L Fausette, “Fundamentals of Neural Networks–Architecture, Algorithm, and Applications”, Prentice-Hall, 1994.

[9] 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.

[10] Y. Yan, M. Fanty, and R. Cole, "Speech Recognition Using Neural Networks with Forward-backward Probability Generated Targets", *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, Munich, April 1997.

[11] E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation" *Parallel Diatributed Processing: Explorations in the Microstructure of Cognition*, Vol. 1, Foundations, A Bradford Book, MIT Press, 1986.

[12] F. K. Soong, A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A Vector Quantization Approach to Speaker Recognition", *Proceedings of International Conference on Acoustics, Speech, and Signal Processing*, p. 387-390, 1985.

Biographies

Chularat Tanprasert received her BS (Mathematics 1st honor, 1989) from Chulalongkorn University, and her MS (1991) and PhD (1994) degrees in computer science from University of Southwestern Louisiana, USA. After her gradation, she has joined the Software and Language Engineering Laboratory (SLL) of NECTEC and worked in Thai OCR project since then. She has also been involved in a number of practical projects such as Thai word processor, web 13, and speaker identification. Her mission as a computer scientist is to advance the theory and application of information processing and computer technologies. Her areas of interest include neural networks, pattern recognition, database systems, artificial intelligence, data mining, and knowledge base.

Chai Wutiwiwatchai received his B.Eng (Electrical Engineering with the 1st honor) from Thammasat University in 1994 and his M.Eng (Digital Signal Processing, Electrical Engineering) from Chulalongkorn University in 1998. He joined the Software and Language Engineering Laboratory (SLL), National Electronics and Computer Technology Center (NECTEC) as an Assistant Researcher since July 1998. His expertise is on speech processing especially speech and speaker recognition, which are currently researched at NECTEC. His interest is on pattern recognition and speech technology.

Sutat Sae-Tang received his BA in Management Information Systems (MIS) with second honours from Rajamangala Institute of Technology in 1993 and MS in Computer Science from Mahidol University in 1998. He joined the Software and Language Engineering Laboratory (SLL), NECTEC since November 1997 as a Research Assistant. His expertise are Internet/Intranet Programming, Database System, System Analysis and Design and PC Programming. He is involved in Thai OCR and Speaker Identification project. His research interests are Optical Character Recognition, Pattern/Speech Recognition, Image Processing and Internet/Intranet Technology.

*This article is a reprint of the article appeared in the Proceedings of IEEE-INNS International Joint Conference on Neural Network, July, 1999.