Proceedings of the 5th WSEAS International Conference on Telecommunications and Informatics, Istanbul, Turkey, May 27-29, 2006 (pp145-149)

Tone Analysis in Harmonic-Frequency Domain and Feature Reduction using KLT+LVQ for Thai Isolated Word Recognition

SARITCHAI PREDAWAN[1]
PRASIT JIYAPANICHKUL[2] and CHOM KIMPAN[3]
Faculty of Information Technology
Rangsit University
Muang-Ake, Paholyotin Road, Patumtani, 12000
THAILAND

CHAI WUTIWIWATCHAI[4]
National Electronics and Computer Technology Center
Ministry of Science and Technology
112 Thailand Science Park, Phahon Yothin Rd.,
Klong 1, Klong Luang, Pathumthani 12120
THAILAND

Abstract: - This work uses intonation analysis in Harmonic Frequency Domain [1] to group Thai isolate words into five groups according to tones and applies Karhunen-Loeve transformation (KLT) to capture only principal components of original set of features. Principal components are projections of eigenvectors that have the higher eigenvalues given original data. The eigenvectors are used as input of Learning Vector Quantization (LVQ) which is a word recognizer. LVQ is a kind of supervised neural network that employs the concept of "winner-take-all" property for classification tasks. By applying the KLT+LVQ method for recognizing ten Thai numeric patterns 0-9 average recognition accuracy achieves 90%.

Key-Words:-Karhunen-Loève Transform, Learning Vector Quantization, Harmonic-Frequency Transform, Feature Reduction.

1 Introduction

Unlike English language that intonations are often used for emphasizing syllables. Intonations in Thai are used to distinguish the word's meanings. Every Thai single-syllable word has its own intonation categorized into five groups, which are low("ake"), medium("saman"), falling("toe"), high("dtee") and rising("jattawa"). Research in Thai tone extraction presented by Ramalingam [4] is compared with our work. Besides, several Thai speech recognition approaches that have been elaborated for years such as in [2] and [5] are studied. The correctness values of these approaches are 74.9% [2] and 89.4% [5]. Although, the results of these experiments can be accepted, the vast amount of time spent for learning all elements of words is still be the most important problem. The main objective for this approach is to reduce the large volume of speech data into low dimensional independent patterns by finding the eigenvectors that can be used later as word representatives. The compression method for used in this paper is Karhunen-Loeve transformation (KLT)

In the part of recognition model improving, there are two main approaches parametric and non-parametric. Hidden Markov Model and Neural Networks are good examples in each approaches. Several kinds of neural networks are applied for speech processing task for example in [6],[7],[2] and [8]. In these networks, Learning Vector Quantization is distinguished for discriminative tasks.In [9], the authors show the 8 power of combining LVQ with HMM for English isolated word recognition. By this powerful classifier, they got 83% accuracy.

2.1 Phonetics of Thai Language

A Thai sentence consists of various words composed from one or more syllables. The smallest part of each syllable is called a phoneme. Every Thai syllable is organized from two up to six phonemes. Luksaneeyanawin [3] has classified Thai phonemes into two groups i.e. segmental phonemes, consisting

of consonants and vowels, and supra segmental phonemes, for example tone and accent phonemes. In our study we divide a syllable into phonemes based on their time occurrences and characteristics. These phonemes are 1) the initial consonant, 2) the vowel, 3) the secondary vowel, 4) the syllable ending, and 5) the tonal. It can be concluded that the tonal is the most important phoneme embedding in every Thai syllable [3] and the tones always locate on the vowels and the resonant syllable ending phonemes [1]. In Fig. 1, the illustration of consonants and vowels of the a Thai numeral syllable "khu:a:w" (nine in English) uttered by six speakers are presented.

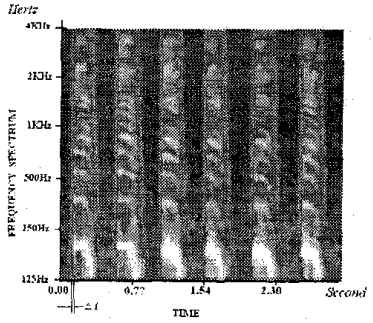

Fig.1.The spectrograms of a Thai numeral syllable uttered by six speakers

In Fig. 1, The consonant and vowel area can be clearly distinguished. The consonant part of each spectrogram is located on the bottom of spectrogram and in the range of frequencies starting from 250 Hz are vowel areas that are represented in the form frequency resonance. The possible direction of vowel contouring [1] is shown in Table 1.

Tone	Low	Mid	Fall	High	Rise
Symbol	\	-	^	/	v

Table 1. Type of Intonation Flow in Thai Language

2.2 Intonation Flow Analysis

Tone analysis of Thai syllables in computer system can be validated by considering the direction of frequency shifts of each signal peak by verifying the differences of two adjacent frequencies in time-axis. Consider Fig. 2

$$E(x,T) = \sum_{i=4}^{123} \left[S(i-x,T) - S(i,T-1) \right]^2 \Big| \; x = -4,-3,..,3,4$$

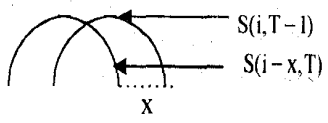

Fig. 2 Frequency Shifting Model to find E(x,T)

If all frequencies in time T and T-1 are similar and all frequency peaks are not changed, the minimum x that causes *E(x,q* equal 0 should be zero. However, if there is transformation of frequency peak in up direction the $x > 0$, and $x < 0$ in otherwise. Since a spectrogram includes not only voice phonemes, but also consists some noises that are not significant in recognition. We separate useful phonemes from noise by considering the energy levels in the spectrogram. The one with lower energy means noise and the others that have energy levels higher than specified thresholds are defined as voice phonemes. In our work, the noises are filtered by using (1):

$$P_{signal} = \begin{cases} 1, & P_{av(min)}(T) - P_{av(min)} \geq \varepsilon \, (P_{av(max)} - P_{av(min)}) \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

P_{signal} is logical value determining whether is used phoneme. $P_{av(max)}$ and $P_{av(min)}$ are maximum and minimum average energies of observed area. ε is coefficient value [0-1]. Then, searching for the vowel area by (2):

$$V_{signal} = \begin{cases} 1, & \dfrac{E(X_{min},T)}{|X_{max} - X_{min}| - E(X_{min},T)} \geq V_{threshold} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

V_{signal} is logical value determining whether is vowel phoneme. X_{max} and X_{min} are values that cause E(x,T) is maximum and minimum average energies of observed area. $V_{threshold}$ is coefficient value for vowel determining [0-1]. The estimated value of tone contouring in discrete time X(T), are calculated form (3)

$$X(T) = X_{min} + \frac{E(x_{min-1,}T) - E(x_{min+1,}T)}{2[E(x_{min-1,}T) - 2E(x_{min,}T)\,E(x_{min+1,}T)]} \quad (3)$$

3.1 Karhunen-Loeve Transformation

An orthogonal transform called the Karhunen-Loeve Transform (KLT) has been well known in principal components analysis or hotelling transformation. It has been applied in various applications in which input data are enormous such as image processing, speech processing etc. The main purpose is to decompose the signals into completely decorrelated components in the form of empirical basis functions that contain the majority of the variations in the original data. The decorrelated components are often called eigenvectors and the scaling constants used to reconstruct the spectra are generally known as coefficient vectors, as shown in Fig. 3.

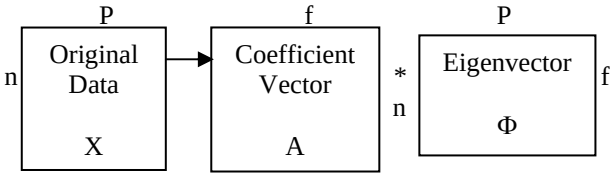

Fig.3 The KLT breaks apart the spectral data into the most common spectral variations (eigenvectors) and the corresponding coefficient vectors

The Karhunen-Loeve decomposition process is started by adapting the original matrix to be a square matrix. The covariance matrix ($\sum_x$) calculation, shown in (4), is selected to achieve in this approach. Given original data $X = [x(t_1)\ldots x(t_n)]^T$, the covariance matrix of X is ($\sum_x$) where

$$(\sum_x)_{jk} = E[\{x(j) - \bar{x}(j)\}\{x^*(k) - \bar{x}^*(k)\}] \quad (4)$$

where P is the total number of input data vector. Let Φ_i, i = 1,…,n, be the eigenvectors of ($\sum_x$), calculated form (5) for which their associated eigenvalues λ_i are arranged in descending order such that $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$.

$$S\Phi_i = \lambda_i \Phi I \quad (5)$$

Those terms associated with smaller eigenvalues can be discarded. In this approach, we choose the eigenvectors projected from eigenvalues that have the total equal 95% of the whole summation of eigenvalues as word representative. The eigenvector Φ_i has a coefficient vector A_i associates with it, computed using

$$A_i = X\Phi_i^T \quad (6)$$

The coefficient vectors, of dimension P, are the projection of the original data matrix X used to monitor the changes over time for each dominant eigenvector. The transformation matrix used to reconstruct the original image is formed the *f* eigenvectors corresponding to the largest *k* eigenvalues. The X´ vectors will be *f*–dimensional and the reconstruction is given by (7)

$$X' = A_K\Phi + \bar{m} \quad (7)$$

It can be shown in the following section that the mean square error MSE between X´ and X for a specified *k* the error is minimized by selecting the *k* eigenvectors associated with the largest eigenvalues.

To analysis principal components is to examines the variance structure in data. One purpose is to consider whether the first few principal components account for most of the variation in the data. There are two ways to investigate the properly amount of principal components to represent the whole set of data with lossless information. The first is plot graph between calculates eigenvalue and their amplitudes as shown in Fig.4.

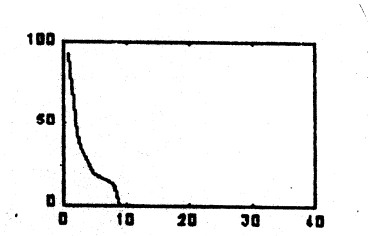

Fig. 4 eigenvalue rank vector for word '0'

Form Fig.4, the higher magnitude of eigenvalue are falled between 0 – 10, therefore we can discard all components with zero-values magnitudes without loss information. The another way to selected the number of principal components is calculated by summing all eigenvalue and approximately 90-95% of it is quite good for representation. In our approach, we select 95% of total sum.

3.2 Learning Vector Quantization (LVQ)

The Learning Vector Quantization (LVQ) is a self-organization network, originated in the neural network community by Kohonen introduction [6] as shown in Fig.5. The LVQ is a method for non-parametric classification that uses a "winner-take-all" strategy. Learning vector Quantization classifies its input data into classes that it determines by using various learning algorithms i.e. LVQ1 and LVQ2.1. The codebook vectors are created from a selected group of input vectors and then used as initial vectors. The trained and tested vectors are created from the other groups of input vectors and then used in training and classifying processes, respectively. The following subsections are mentioned of learning algorithm of LVQ.

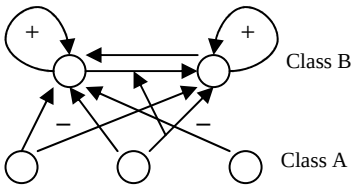

Fig. 5 The layout of the Learning Vector Quantization [10].

3.2.1 The LVQ1

Assume that a number of codebook vectors m_i, are placed into the space to approximate various domains of the input vector x by their quantized values. Usually several codebook vectors are assigned to each class of x values, and x is then decided to belong to the same class to which the nearest m_i belongs. Let

$m_c = \arg\min (\| x - m_i \|)$, (8)
where m_c is the nearest m_i to x.

Let x(t) be a sample of input and let the $m_i(t)$ represent sequences of the m_i in the discrete-time domain. The following equations define the basic LVQ1 process:

$m_c(t+1) = m_c(t) + \alpha(t)[x(t) - m_c(t)]$ (9)
;if x and m_c are in the same class

$m_c(t+1) = m_c(t) - \alpha(t)[x(t) - m_c(t)]$ (10)
;if x and m_c are in the different class,

$m_c(t+1) = m_c(t)$ (11)
;for input nodes which are not closet.

where α value is between 0 and 1 and must be constant or decrease linearly with time. The α value is 0.1 and linear decreasing in time are used in our work.

3.2.2 The LVQ2.1

With the process of the LVQ1, it is possible that there are more than one codebook vectors (m_i and m_j) are nearest to the input x. Therefore, those two vectors should be updated simultaneously. One of them has to belong to the right class and the another to a wrong class, respectively. Moreover, x must fall into a zone of values called "window", which is defined around the center of m_i and m_j. Let di and sj are distances of x from mi and mj, then x is defined to fall in a window of relative width w if

$\min (d_i / d_j , d_j / d_i) > s$, (12)
where $s = (1-w)/ (1+w)$

Then the training algorithm for the LVQ2.1 are:

$m_i(t+1) = m_i(t) - \alpha(t)[x(t) – m_i(t)]$ (13)
$m_j(t+1) = m_j(t) + \alpha(t)[x(t) – m_j(t)]$ (14)

where m_i and m_j are the two closet codebook vectors to x, whereby x and m_j belong to the same class, while x and m_i belong to different class, respectively.

3.3 Proposed model

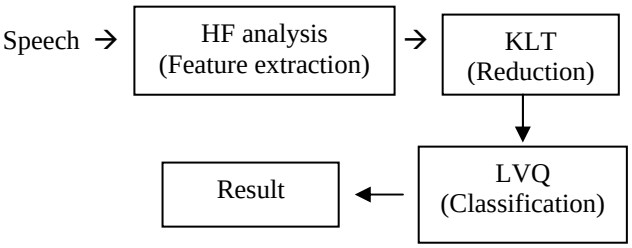

Fig. 6 Overview of Harmonic-Frequency Domain and Feature Reduction using KLT+LVQ for Thai Isolated Word Recognition

4 Experiment and Result of Recognition

The experiment for our approach is to apply the Karhunen-Loeve transformation and Learning Vector Quantization. Ten Thai numeral syllables are experiment in our work. All speech data are separated into three groups as show in each column in table 2 and 3. The 1st column, all data are absolutely separated. The 2nd column, we include all speech files in initialized step for codebook training. And in the last column, speech files for testing is also known by LVQ both in initialized and training step.

LVQ1 learning			
Step	No. Files	No. Files	No. Files
Initialize	300	300	300
Training	500	800	900
Recognition	100	100	100*
Accuracy	83.50	85.70	93.31

*test files are take from training files

Table 2 Recognition results using LVQ1 learning method

LVQ2.1 learning			
Step	No. Files	No. Files	No. Files
Initialize	300	300	300
Training	500	800	900
Recognition	100	100	100*
Accuracy	78.67	87.96	90.97

*test files are take from training files

Table 3 Recognition results using LVQ2.1 learning method

Comparing to the method of Harmonic Frequency analysis proposed in [4], errors reduce by approximately 5%. Error distribution is shown in Table 4.

Syllables	Tone	Syllables	Tone
"1"	3	"6"	1
"2"	0	"7"	1
"3"	1	"8"	0
"4"	0	"9"	2
"5"	1	"10"	0

Table 4 Number of errors for each Thai numerals

Comparing the Thai intonation analysis approach with [4], the number of correctly identified events, 95% , though is not considerably better than 100% of [4] work. It can be acceptable we experiment with more speakers, variety words with and without the syllables ending, and in various situation i.e. noise-free, less signal-noise ratio, and pretending. Moreover, the voice identify finding method can decreased some problems in speaker-independent works. Since it uses statistical model as correlative coefficient, therefore different amplitudes, frequency and uttered period are not affected the results. However, the methods that used in our research is emphasized on statistical and mathematics model, which expenses more analyzing time.

5 Conclusion

The paper describes also how to use Karhunen-Loeve transformation for feature of speech data representation. The KL transformation has been able to compress multi-dimension waveform data to low-dimensional uncorrelated data. These are eigenvectors and coefficient vector (principal component). The principal components are the projection of the eigenvalues on the original data. We choose two principal components that have the largest two eigenvalues as speech representative. Our results in the sense of data reduction are about 80% decrement. With these principal components, we can test their efficiency by reconstructing speech signal from them using KL-expansion (inverse transform) the mean square error compare to the original are $3x10^{-4}$. Therefore KLT can yield the well-enough feature representation for the following learning and classification task.

The LVQ algorithm consists of two phases that are learning and classification phases. In learning phase, the codebook vectors are started with equal number of vectors and then adjusted by using the past observations. Two algorithms of LVQ are used i.e. LVQ1 and LVQ2.1. In the classification phase a new observation is classified using the trained vectors. As shown in Table2 and 3, the power of LVQ for recognizing Thai numeral words are quiet good. The accuracy is between 80% and 93% in the whole experiments.

References

[1] R.Kongkachandra, S.Pansong, T.Sripramong and C.Kimpan. : "Thai Intonation Analysis in Harmonic-Frequency Domain," The 1998 IEEE APCCAS. Proceeding, (1998) pp. 165-168. 1998.

[2] E.Maneenoi, et al.: "Modification of BP Algorithm for Thai Speech Recognition," NLPRS'97 (Incorporating SNLP'97) Proceeding (1997), pp. 287-291

[3] Luksaneeyanawin, S.: "Intonation in Thai," Ph.D. Thesis, University of Edinburgh, 1983

[4] Ramalingam, H.: "Extraction of Tones of Speech: An application to the Thai Language," Master Thesis (TC-95-5), Asian Institute of Technology, Thailand, 1995

[5] W.Pornsukjantra, et al: "Speaker Independent Thai Numeral Speech Recognition Using LPC and the Back Propagation Neural Network," NLPRS'97 (Incorporating SNLP'97) Proceeding, pp. 585-588. 1997.

[6] A.Biem, S.Katagiri and B.-H.Juang, "Discriminative feature extraction for speech recognition," International workshop on Neural Networks for signal proceesing, Baltimore September 1993.

[7] A.Biem, and M.Sugiyama, "A hybrid stochastic and connectionist approach to speech recognition," 1st African conference in computer science, Yaounde, Cameroon, 1992.

[8] J.Hennebert, M. Hasler, H. Dedieu, "Neural Networks in speech Recognition," Micro Computer School 94, Prague, Tchec Republic.

[9] S.Katagiri and C.H.Lee, " A New Hybrid Algorithm for Speech Recognition Base on HMM Segmentation and Learning Vector Quantization," IEEE Transactions on Speech and Audio Processing, vol. 1, no. 4, pp. 421-430, October 1993.

[10] M.Purat, T.Liebchen and P.Noll. "Lossless Transform Coding of Audio Signals," 102nd AES Convention, Muchen, Preprint No.4414, 1997.