Expert Systems with Applications 36 (2009) 8056–8063

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Speaker identification based on the frame linear predictive coding spectrum technique

Jian-Da Wu *, Bing-Fu Lin

Graduate Institute of Vehicle Engineering, National Changhua University of Education, 1 Jin-De Rd., Changhua City, Changhua 500, Taiwan

ARTICLE INFO

Keywords:
Speaker identification
Linear predictive coding
Gaussian mixture model
General regression neural network

ABSTRACT

In this paper, a frame linear predictive coding spectrum (FLPCS) technique for speaker identification is presented. Traditionally, linear predictive coding (LPC) was applied in many speech recognition applications, nevertheless, the modification of LPC termed FLPCS is proposed in this study for speaker identification. The analysis procedure consists of feature extraction and voice classification. In the stage of feature extraction, the representative characteristics were extracted using the FLPCS technique. Through the approach, the size of the feature vector of a speaker can be reduced within an acceptable recognition rate. In the stage of classification, general regression neural network (GRNN) and Gaussian mixture model (GMM) were applied because of their rapid response and simplicity in implementation. In the experimental investigation, performances of different order FLPCS coefficients which were induced from the LPC spectrum were compared with one another. Further, the capability analysis on GRNN and GMM was also described. The experimental results showed GMM can achieve a better recognition rate with feature extraction using the FLPCS method. It is also suggested the GMM can complete training and identification in a very short time.

1. Introduction

Communications among human beings are simple and common, but the same action between human and computer is difficult beyond our expectation. Till now, many scientists are still working on how to make a machine not only to decipher what people say, but also to understand the meaning implied by the words. Fortunately, due to the great promotion of computing ability in microprocessors, applications of voice signal processing have been increasing rapidly in recent years, such as sound recognition or speaker identification. In speech recognition, the system has to recognize various commands from various speakers. Thus, many algorithms have been developed to extract the vital components hidden in speech signals, for example, spectral or cepstral characteristics. On the other hand, speaker identification originated from the needs for security monitoring in many important buildings or facilities. By applying this mechanism, the approaching people can be observed. In addition, speaker identification is utilized in suspect identification because of its properties of non-contact characteristic.

In this paper, a text-dependent speaker identification system is proposed. The advantage of text-dependent lies in the sentence used for recognition does not need to be very long; it can simply be a word or an utterance. Besides, unlike text-independent system, shorter sentences can increase classification speed. Generally speaking, implementing speaker identification can be divided into two stages: the first is feature extraction and the second is speaker classification based on the extracted features (Sarikaya, Pellom, & Hansen, 1998). At the stage of feature extraction, the extracted features should be capable of separating the speakers from each other in its space. In traditional techniques, the speech features are usually obtained by Fourier transforms and short time Fourier transforms. However, these techniques are unsuitable for speaker identification because they accept stationary signal within a given time frame and may therefore lack the ability to analyze the non-stationary signals or signals in transient state (Avci & Akpolat, 2006). Therefore, many algorithms were developed to find a better representation of a speaker, for example: linear predictive coding (LPC) technique (Adami & Barone, 2001; Haydar, Demirekler, & Yurtseven, 1998; Wutiwiwatchai, Achariyakulporn, & Tanprasert, 1999), Mel frequency cepstral coefficient (MFCC) (Mashao & Skosan, 2006; Sroka & Braida, 2005) and wavelet (Lung, 2006; Wu & Lin, 2009; Wu & Ye, 2009). In this paper, an improved method based on LPC is proposed. In fact, LPC is not a new method, it was developed in 1960s (Atal, 2006), but is popular and widely used till today because LPC coefficients representing a speaker by modeling vocal tract parameters and the data size are very suitable for speech compression through the digital channel. In the present study, the focus will be on modifying LPC coefficients and reducing the size of feature vectors.

* Corresponding author.
E-mail address: Jdwu@cc.ncue.edu.tw (J.-D. Wu).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.10.051

On the selection of classifier, the general regressive neural network (GRNN) and Gaussian mixture model (GMM) were chosen to be applied in the classification stage. Both two classifiers are popular in many pattern recognition fields because they can achieve good performance and the training time is well-satisfied. In the following section, an experimental investigation was carried to form a comparable result between these two approaches. Both their advantages and disadvantages will be compared with each other.

2. Principles of feature extraction and classification

2.1. Linear predictive coding technique

In modern signal processing, the analysis procedure extracts useful information from the structure of a signal. LPC technique is a developed algorithm used in speech analysis for many years. Its basic idea comes from a model representing the resonances of the human vocal tract. In general, speech sounds are produced by acoustic excitation of the vocal tract. During the production of voiced sounds, the vocal tract is excited by a series of nearly periodic pulses generated by the vocal cords. With unvoiced sounds, the excitation is provided by air passing turbulently through constrictions in the tract (Atal & Hanauer, 1971). Fig. 1 shows the speech signal production model in which the speech synthesizer strongly depends on the estimation of a_p. Here, a_p are autoregressive parameters obtained from the linear prediction method, and provide better results to characterize human speech. The use of these parameters assumes the speech signal can be represented as the output signal of an all pole digital filter in which the excitation is an impulse sequence with a frequency equal to the pitch of the speech signal under analysis when the segment is voiced, or with noise when the segment is unvoiced (Perez-Meana, 2007). The steps of acquiring a_p are described as follows:

(a) The input signal is segmented in 20 ms with 10 ms overlapping.
(b) Apply the window function to these segments to avoid distortion of the segmented speech because of the discontinuities introduced during the segmentation process, typically the Hamming window is used. The Hamming window is given by the following equation:

$$W(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), \quad \text{for} \quad 0 \leqslant n \leqslant N-1, \tag{1}$$

where N is the number of samples of the used segment.
(c) Estimate the prediction order and calculate the linear prediction coefficients for each segment.

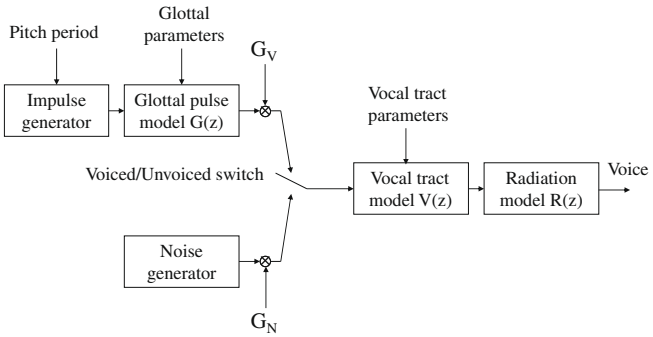

Fig. 1. Flow chart of speech synthesis.

After the speech signal is segmented, the p autocorrelation coefficients are estimated, where p is the linear predictor order. The autocorrelation function can be estimated using the biased or unbiased autocorrelation algorithms (Childers, 2000). Once autocorrelation coefficients p are evaluated from each segment, the signal at time n can be rewritten as a linear combination from the pass samples of the input signal:

$$\widehat{s}(n) = -(a_1 s(n-1) + a_2 s(n-2) + \ldots + a_p(n-p)), \tag{2}$$

or

$$\widehat{s}(n) = -\sum_{k=1}^{p} a_k s(n-k), \quad k = 1, 2, \ldots, p. \tag{3}$$

Therefore, it is affirmed a filter can be designed to estimate the data at time n only using the previous data at time $n-1$:

$$\widehat{s}(n) = -a_e s(n-1), \tag{4}$$

where a_e is the linear prediction coefficient. To evaluate a_e, the prediction error is minimized expectantly between $s(n)$ and $s(n-1)$:

$$e(n) = s(n) - \widehat{s}(n) = s(n) + a_e s(n-1). \tag{5}$$

After a series of calculations, $s(n)$ is evaluated and rewritten in the Z domain:

$$S(z) = \frac{E(z)}{1 + \left[\sum_{k=1}^{p} a_k z^{-k}\right]} = \frac{E(z)}{A(z)}, \tag{6}$$

where

$$A(z) = 1 + \sum_{k=1}^{p} a_k z^{-k}.$$

Eq. (6) denotes the transfer function of an all pole filter shown in Fig. 2. The poles of transfer function are the zeros of the polynomial in the denominator on the right side of Eq. (6). The linear filter thus has a total of p poles which are either real or occur in conjugate pairs. Moreover, for the linear filter to be stable, the poles must be inside the unit circle (Atal & Hanauer, 1971). Once the prediction coefficients a_k are obtained, a more specific algorithm to extract the features for representing a speaker is applied. Therefore, the frame based linear predictive coding spectrum (FLPCS) coefficients is introduced in this paper.

2.2. Frame based linear predictive coding spectrum

LPC coefficients provide good reproduction of human vocal tract in speech synthesis. In this paper, these coefficients were not used as features for speaker identification directly due to the complexity of the data dimension. Moreover, during the procedures of acquiring LPC coefficients, some signal-freed segments were also involved in the LPC computation. This may lead to difficulties in the classification stages because speech signals always have different location on the time axis. Therefore, a new algorithm called FLPCS coefficients is proposed in this paper to solve this dilemma. Basically, FLPCS coefficients are based on LPC but have different ways of extracting features. The mathematical expression is defined as

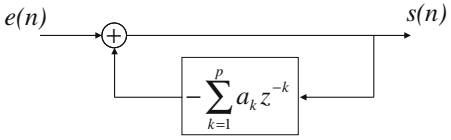

Fig. 2. Transfer function of the all pole filter.

$$X_{FLPCS}(\hat{a}_{k_j}) = \frac{1}{N}\sum_{i=1}^{N} a_{k_{ij}}, \quad j = 1, 2, \ldots, p, \tag{7}$$

where a_k denotes the prediction coefficients and N, p denote the time frames and predictive orders, respectively. In fact, Eq. (7) transfers two-dimensional LPC coefficients into one-dimensional FLPCS coefficients. Thus, the data complexity is reduced on the one hand, but on the other hand, the demur on time location can be ignored.

FLPCS reflects the tendency of LPC coefficients in another way which is similar to the energy accumulation used in the wavelet technique (Wu & Lin, 2009; Zheng, Li, & Chen, 2002). However, FLPCS can provide a faster and easier approach to obtain features for representing a speaker. Fig. 3 shows the LPC spectrum in which the amplitude of the coefficients changes with the speech content and time frames. Fig. 4 shows a two-dimensional feature vector extracted by FLPCS. Before sending FLPCS coefficients into the classifier, a post-process procedure is needed to normalize these coefficients.

2.3. Post-process of signals

FLPCS coefficients stand for the speech characteristics of a speaker in an amplitude-like way. But the distribution of these coefficients is closely related to the volume of the speakers. Any dissimilar volume will change the position on the y-axis and leads to misclassification. Therefore, a post-process is needed to eliminate this phenomenon. The post-process is given by the following equation (Lou & Loparo, 2004):

$$X_p(a_k) = \frac{X(a_k) - \mu}{\sigma}, \tag{8}$$

where $X(a_k)$ and $Xp(a_k)$ denote the before and after post-process coefficients; μ and σ are the mean and standard deviation of the vector $X(a_k)$, respectively. The post-process makes the coefficients comparable regardless of differences in magnitude and remains their origin form. In Fig. 5a, it can be seen the FLPCS coefficients have similar shape but are dispersedly distributed. Fig. 5b shows these coefficients distributed very well after post-process.

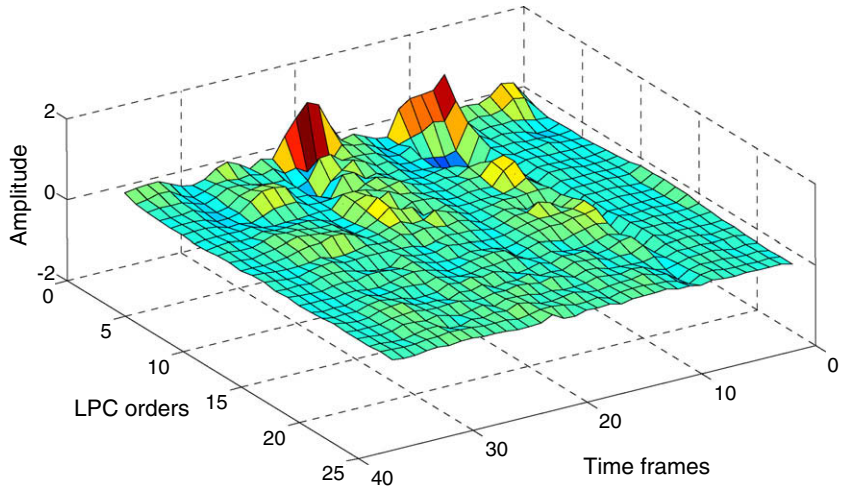

Fig. 3. LPC spectrum of a speech signal.

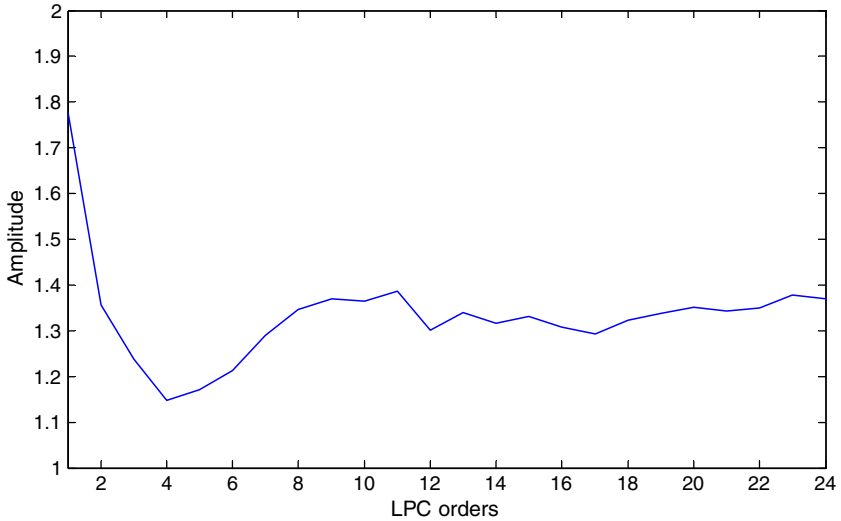

Fig. 4. FLPCS coefficients of a speaker extracted from LPC.

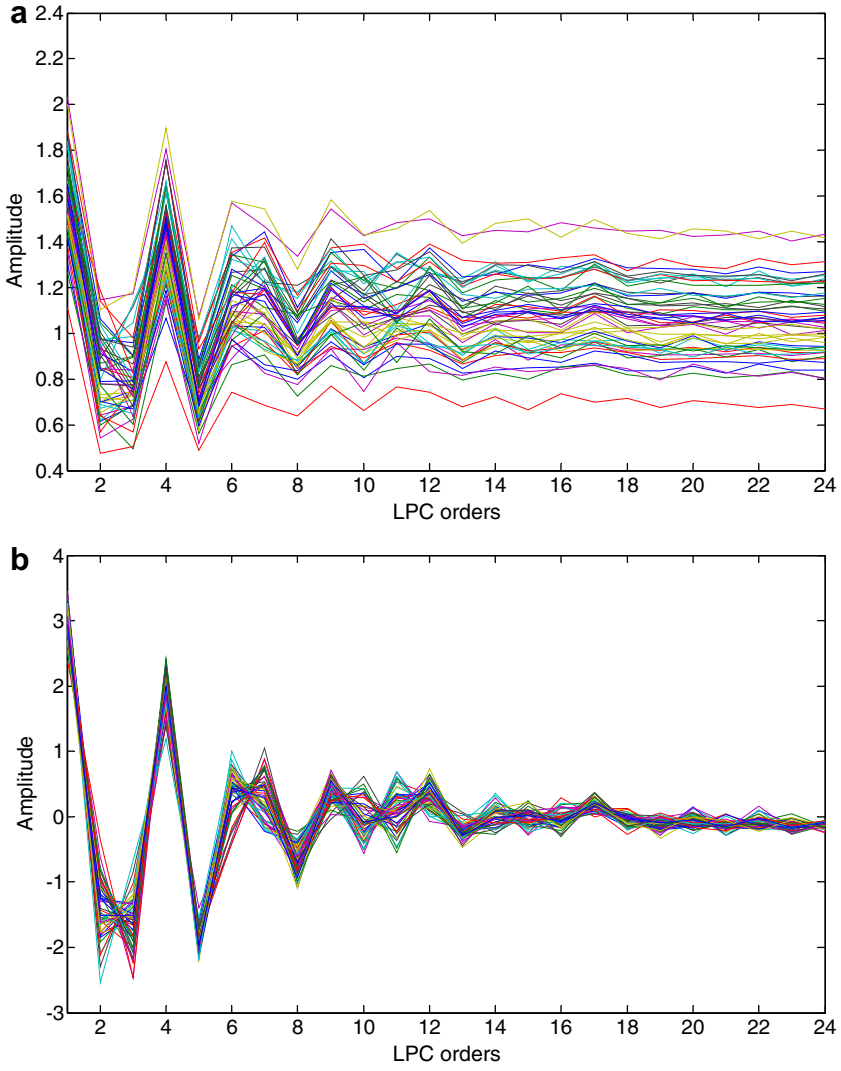

Fig. 5. 50 FLPCS vectors of a female speaker (a) before post-process and (b) after post-process.

2.4. Classifiers for speaker identification

2.4.1. General regression neural network

GRNN was first proposed by Specht (1991) and is widely used in many recognition tasks. Fig. 6 shows the block diagram of the GRNN architecture. It is a one-passing learning algorithm, which can be used for estimating continuous variables such as some transient content in speech signal. In addition, it does not require an iterative training procedure to converge to the desired solution as in the back-propagation (BP) neural network.

By definition, the regression of a dependent variable y on an independent x estimates the most probable value for y, given x and a training set. The GRNN is a method for estimating the joint probability function of x and y to produce the estimated value of y, given only a training set. Assume $f(x,y)$ represents the known joint continuous probability density function (PDF) of a vector random variable, x, and a scalar random variable. Let X be a particular measured value of the random variable x. The conditional mean of y given X is given by

$$E[y|X] = \frac{\int_{-\infty}^{\infty} yf(X,y)dy}{\int_{-\infty}^{\infty} f(X,y)dy}. \tag{9}$$

when the density $f(x,y)$ is unknown, it must usually be estimated from the sample of observations of x and y. The probability estimator $f(X,Y)$ is based upon sample values X^i and Y^i of the random variables x and y, where n is the number of sample observations and p is the dimension of the vector variable x:

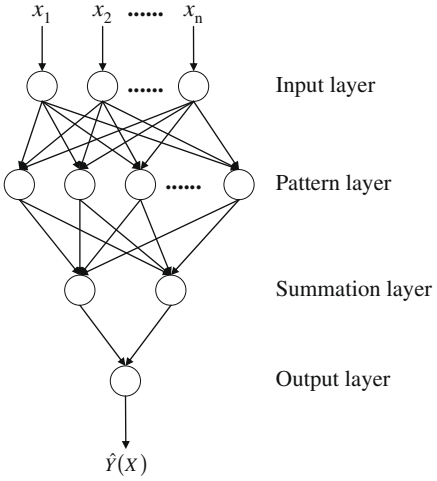

Fig. 6. Block diagram of GRNN architecture.

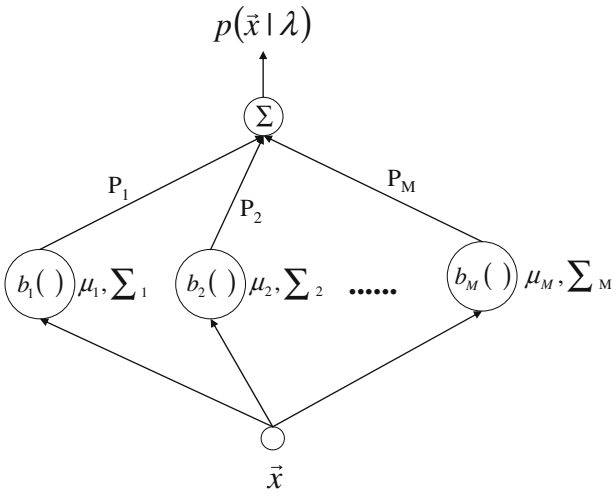

Fig. 7. Depiction of an M component Gaussian mixture density.

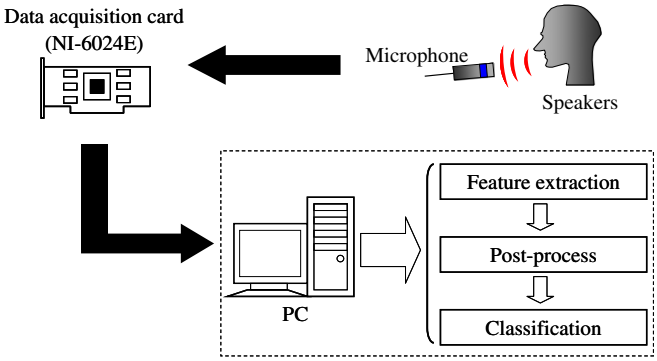

Fig. 8. Experimental setup of the proposed speaker identification system.

$$\hat{f}(X,Y)=\frac{1}{(2\pi)^{(P+1)/2}\sigma^{(P+1)}}\frac{1}{n}\times\sum_{i=1}^{n}\times\exp\left[-\frac{\left(X-X^i\right)^T(X-X^i)}{2\sigma^2}\right]\exp\left[-\frac{(Y-Y^i)^2}{2\sigma^2}\right]. \tag{10}$$

Table 1
Five cases in the speech recording experiment.

Sentence	Taiwan Tongyong Romanization	English representation
1	wǒ shìh wáng siaō míng	Speaker's name
2	cǐng kai mén	Open the door
3	san liòu sìh ba	A set of password (3648)
4	fa dòng yǐn cíng	Start the engine
5	fáng dào cǐ dòng	Enable security system

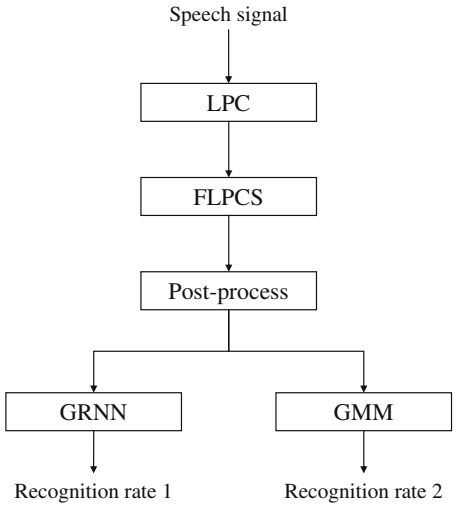

Fig. 9. Flow chart of the speech analysis.

A physical interpretation of the probability estimate $\hat{f}(X,Y)$ is it assigns the sample probability of width σ for each sample X^i and Y^i, and the probability estimate is the sum of those sample probabilities. Defining the scalar function

$$D_i^2=(X-X^i)^T(X-X^i) \tag{11}$$

and performing the indicated integration yields:

$$\hat{Y}(X)=\frac{\sum_{i=1}^{n}Y^i\exp\left[-\frac{D_i^2}{2\sigma^2}\right]}{\sum_{i=1}^{n}\exp\left[-\frac{D_i^2}{2\sigma^2}\right]}. \tag{12}$$

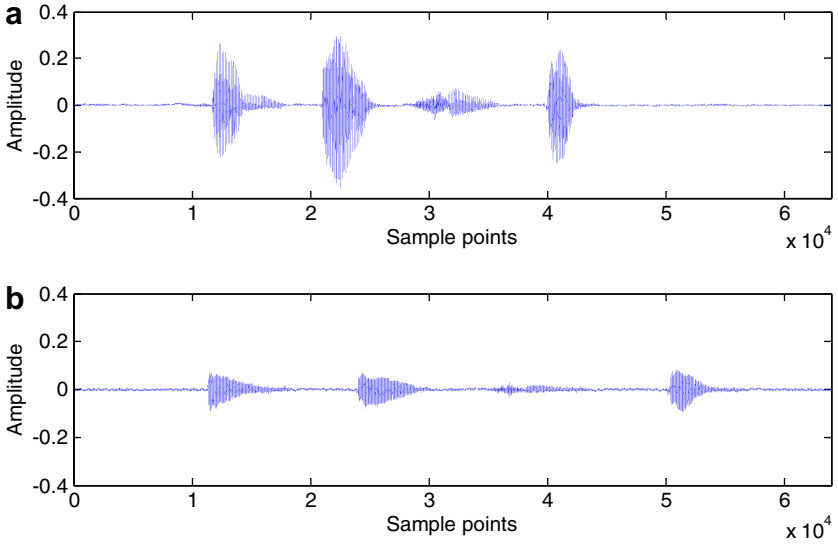

Fig. 10. Speech signal diagram in time domain: (a) male and (b) female.

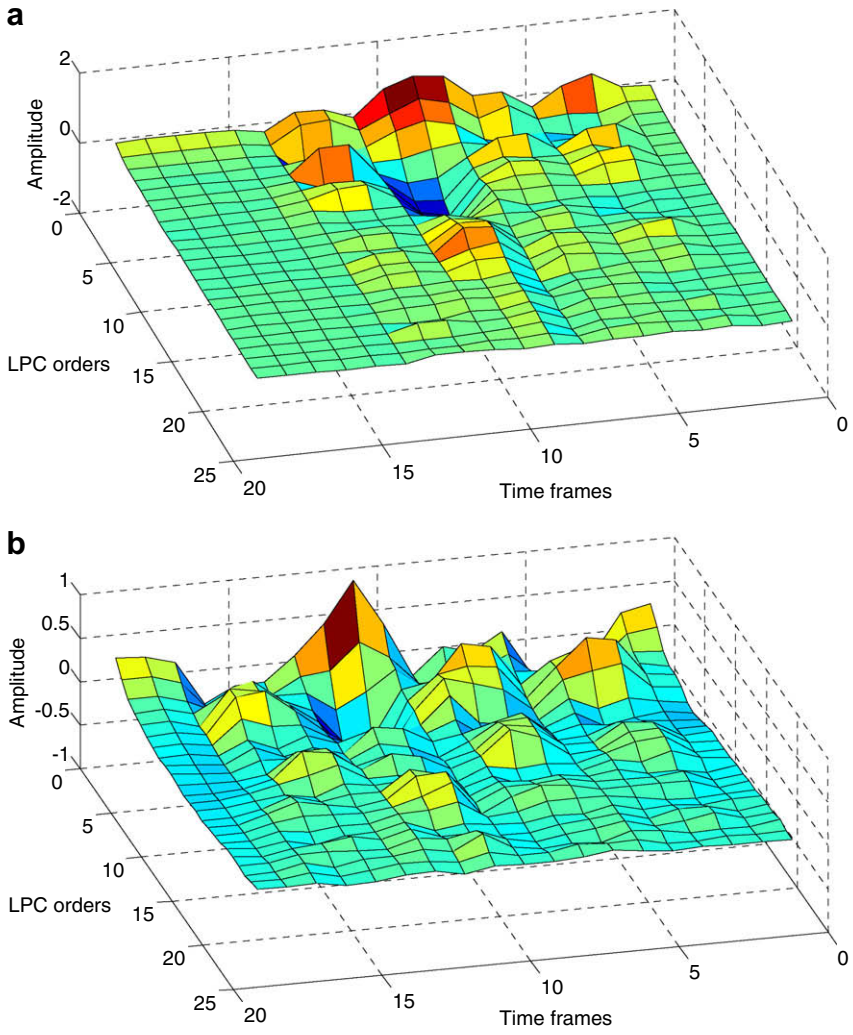

Fig. 11. LPC spectrum of the speech: (a) male and (b) female.

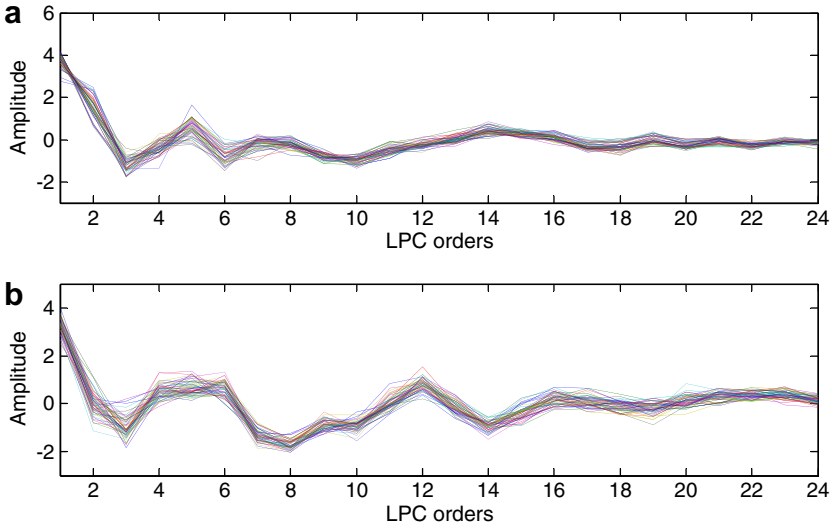

Fig. 12. FLPCS representation for the speaker: (a) male and (b) female.

When the smoothing parameter σ is large, the estimated density is forced to be smooth and the limit becomes a multivariate Gaussian with covariance $\sigma^2 I$. On the other hand, a smaller value of σ allows the estimated density to assume non-Gaussian shapes, but with the hazard that wild points may have a significant effect on the estimation.

2.4.2. Gaussian mixture model

GMM is extensively used for classification tasks especially in speaker identification (Markov & Nakagawa, 1998; Reynolds & Rose, 1995; Ruiz, Domingo, & Hernandez, 1999). Because GMM can smoothly estimate the density distribution of the data clusters, its accuracy and performance are outstanding. Fig. 7 shows the structure of a GMM model. A Gaussian mixture density can be written as a weight sum of M component densities, which is given by the following equation:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}), \tag{13}$$

where $\vec{x}$ is a D-dimensional random vector, p_i, $i = 1,\ldots,M$, are the mixture weights and $b_i(\vec{x})$, $i = 1,\ldots,M$, are the component densities. Each component density is a D-variate Gaussian function of the form:

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2}|\sum_i|^{1/2}} \exp\left[\frac{-1}{2}(\vec{x}-\vec{\mu}_i)' \sum_i^{-1}(\vec{x}-\vec{\mu}_i)\right]. \tag{14}$$

The mixture weights satisfy the constraint with mean vector $\vec{\mu}_i$ and covariance matrix $\sum_i$:

$$\sum_{i=1}^{M} p_i = 1. \tag{15}$$

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are given collectively by the notation:

$$\lambda = \left\{p_i, \vec{\mu}_i, \sum_i\right\}, \tag{16}$$

where $i = 1,\ldots, M$. For speaker identification, each speaker is represented by a GMM and is referred to by his or her model λ.

3. Experimental investigation and analysis

To evaluate the proposed method, a sound recording experiment was carried. In the beginning, utterances of a speaker were recorded by a microphone with a data acquisition system. Then the recorded voices were taking operation in a Pentium level PC. Fig. 8 shows the experimental setup of the proposed speaker identification system. The measured maximum frequency was 8 kHz while the sampling rate was set at 16 kHz. The recording apparatus consists of a microphone (PCB 130D20) and a data acquisition card (NI-6024E). The speech database comprises 50 speakers including 25 female and 25 male speakers. Each speaker repeated an assigned sentence 50 times, and there were five sentences in all, as shown in Table 1. In the proposed identification system, speech signals were framed by a series of Hamming windows where the frame size is 20 ms and the overlapped length is 12.5 ms.

The complete analysis flowchart is shown in Fig. 9, pointing out the speech signals were framed and pre-emphasized by Hamming windows before extracting the LPC coefficients. Then the post-process operation was taken after FLPCS coefficients were obtained from the LPC spectrum. The performance of the FLPCS method was evaluated by different classifiers including GMM and GRNN, which are rapid in training procedure and have the potential for real-time applications. Fig. 10 demonstrates a male and a female speaker's speech signal diagrams in the time domain. The LPC

Table 2
Recognition rates using GRNN in various LPC orders and sentences.

LPC orders	GRNN				
	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
13	77.2	83.8	80	78	81.2
25	84	88	85	90.8	88.6
50	87.2	90	88.2	91.6	88.8

(Recognition rate: %).

Table 3
Recognition rates using GMM in various LPC orders and sentences.

LPC orders	GMM				
	Sentence 1	Sentence 2	Sentence 3	Sentence 4	Sentence 5
13	85	93.4	92.2	89.8	92.2
25	95.8	98.2	95	97.8	97.6
50	97	98.6	96.4	99.2	99

(Recognition rate: %).

Table 4
Overall recognition rates of the classifiers under different number of training samples (LPC order = 25).

Classifier	Number of training samples		
	10	20	30
GRNN	81.36	85.60	87.28
GMM	89.14	94.31	96.88

(Recognition rate: %).

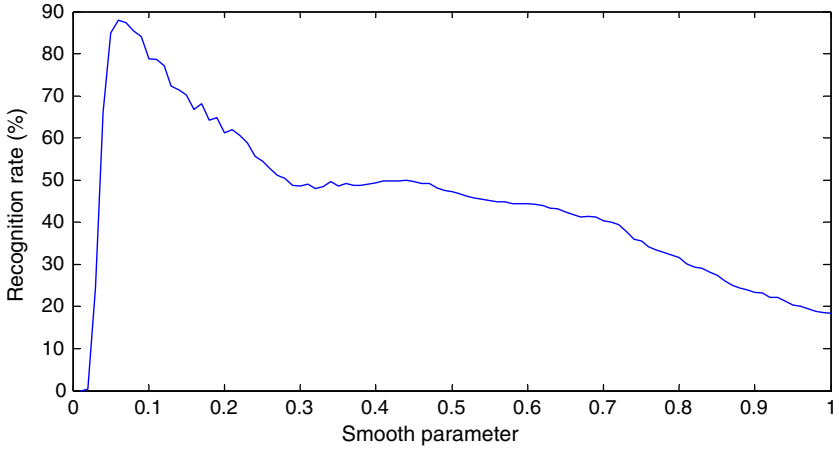

Fig. 13. Iterated computation of GRNN, in which the optima smoothing parameter is determined at where the maximum recognition rate is occurred.

spectrum diagram is indicated individually in Fig. 11, in which it was found the voiced phoneme produces positive coefficients, on the other hand, the unvoiced phoneme simultaneously produces negative coefficients. The FLPCS coefficients, shown in Fig. 12, were extracted from LPC spectrum according to Eq. (8). The overall experimental results include:

(a) The recognition rates of FLPCS coefficients using GRNN with different LPC orders are summarized in Table 2.
(b) The recognition rates of FLPCS coefficients using GMM with different LPC orders are summarized in Table 3.
(c) Effects from the various numbers of training samples on classifiers' performance were also evaluated in Table 4.

In the experiments, three different LPC orders were set to reveal the correlation between LPC order and the recognition rate. In Tables 2 and 3, it was found the recognition rate is proportional to LPC orders. With more coefficients, the higher recognition rate was acquired. Besides, the increase on LPC coefficients did not tremendously burden the system load. But the use of these parameters still has its limitation since the number of parameters slightly affects the recognition rate, especially when recognition rate reached over 95%, it did not produce significant improvement in performance even though double the FLPCS coefficients were used. Moreover, for both GRNN and GMM, the increase in parameters also affects the training time, nevertheless, it remains in an acceptable range.

Though GRNN is rapid in training procedure as mentioned in Section 2.4.1, such a description only aims at one situation where the smoothing parameter is determined. The smoothing parameter is an important parameter in deciding whether the GRNN can smoothly estimate the data. For an unknown data vector, there is no mathematical expression to find the ideal value. To solve this problem, the only avenue is to depend on the iterative experiments. Fig. 13 shows the repeated computation procedure in which the smoothing parameter was determined when the maximum recognition rate occurred. For GMM, iterative training is not needed, but the number of PDF used to fit the feature vector still needs to be set. In the experimental work, the uncertainty in choosing the number of PDF is becoming a dilemma. However, if the complexity of the system is considered, it is suggested 2 PDFs is suitable for approximating all data.

4. Conclusions

In this paper, a speaker identification system based FLPCS is presented. The advantage of FLPCS is it transfers the speech data into a few values, and the computing speed is also satisfied. In the beginning of feature extraction, LPC coefficients were obtained by analyzing the vocal tract parameters of a speaker. Then FLPCS coefficients were extracted from LPC and used as a feature vector. The classification stage includes GRNN and GMM. Experimental results showed both GRNN and GMM are rapid in the single training procedure. However, GRNN needs iterative computation to determine the optimal smoothing parameter. Thus, GRNN spent more time than GMM and made the system unable to react in a reasonable time. As a comparison, GMM is much faster, also, GMM performed well in the recognition rate. The experimental results revealed the proposed FLPCS technique with GMM can accomplish the speaker identification in a short time and achieve a satisfactory recognition rate.

Acknowledgement

The study was supported by the National Science Council of Taiwan, Republic of China, under Project No. NSC-96-2622-E-018-001-CC3.

References

Adami, A. G., & Barone, D. A. C. (2001). A speaker identification system using a model of artificial neural networks for an elevator application. *Information Sciences, 138*, 1–5.
Atal, B. S. (2006). The history of linear prediction. *Signal Processing Magazine, IEEE, 23*, 154–161.
Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. *The Journal of the Acoustical Society of America, 50*, 637–655.
Avci, E., & Akpolat, Z. H. (2006). Speech recognition using a wavelet packet adaptive network based fuzzy inference system. *Expert Systems with Applications, 31*, 495–503.
Childers, D. G. (2000). *Speech processing and synthesis toolboxes*. New York: John Wiley & Sons.
Haydar, A., Demirekler, M., & Yurtseven, M. K. (1998). Speaker identification through use of features selected using genetic algorithm. *Electronics Letters, 34*, 39–40.
Lou, X., & Loparo, K. A. (2004). Bearing fault diagnosis on wavelet transform and fuzzy inference. *Mechanical System and Signal Processing, 18*, 1077–1095.
Lung, S. Y. (2006). Wavelet feature selection based neural networks with application to the text independent speaker identification. *Pattern Recognition, 39*, 1518–1521.
Markov, K. P., & Nakagawa, S. (1998). Text-independent speaker recognition using non-linear frame likelihood transformation. *Speech Communication, 24*, 193–209.
Mashao, D. J., & Skosan, M. (2006). Combining classifier decisions for robust speaker identification. *Pattern Recognition, 39*, 147–155.
Perez-Meana, H. (2007). *Advances in audio and speech signal processing: Technologies and applications*. Hershey: IGI Global.
Reynolds, D. A., & Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Transactions on Speech and Audio Processing, 3*, 72–83.
Ruiz, B., Domingo, P. & Hernandez, L. (1999). A dual speech/speaker recognition using GMM in speaker identification and a HMM in keyword speech recognition. In *Proceedings of the IEEE 33rd annual 1999 international Carnahan conference on security technology* (pp. 251–254).
Sarikaya, R., Pellom, B. L. & Hansen, J. H. L. (1998). Wavelet packet transform features with application to speaker identification. In *Proceedings of the IEEE nordic signal processing symposium* (pp. 81–84).
Specht, D. F. (1991). A general regression neural network. *IEEE Transactions on Neural Networks, 2*(6), 568–576.
Sroka, J. J., & Braida, L. D. (2005). Human and machine consonant recognition. *Speech Communication, 45*, 401–423.
Wu, J. D., & Lin, B. Fu. (2009). Speaker identification using discrete wavelet packet transform technique with irregular decomposition. *Expert Systems with Applications, 36*, 3136–3143.
Wu, J. D., & Ye, S. H. (2009). Driver identification based on voice signal using continuous wavelet transform and artificial neural network techniques. *Expert Systems with Applications, 36*, 1061–1069.
Wutiwiwatchai, C., Achariyakulporn, V. & Tanprasert, C. (1999). Text-dependent speaker identification using LPC and DTW for Thai language. In *Proceedings of the IEEE region 10 conference, TENCON 99* (pp. 674–677).
Zheng, H., Li, Z., & Chen, X. (2002). Gear fault diagnosis based on continuous wavelet transform. *Mechanical System and Signal Processing, 16*, 447–457.