SPEAKER IDENTIFICATION USING MULTIBAND LINEAR PREDICTIVE CODE

Asst. Lect. Ahmed K. Hassan

Receipted: 17 /9 /2006
Accepted: 14/11/2007

Abstract

This paper presents an effective method for improving the performance of speaker identification system based on the multiresolution properly of the wavelet transform, the input speech signal is decomposed into L subbands. To capture the characteristic of the vocal tract, the liner prediction code of each band (including the linear predictive code (LPC)for full band) are calculated.

The feature recombination schemes combines the LPC of each band and LPC for full band in single feature vector then the Euclidean distance measure is used to perform the similarity measure between the test and reference speech. Experimental results shows that the proposed method achieve better performance than speaker identification using LPC and real cepstral coefficients.

الخلاصة

في هذا البحث تم تمثيل طريقة فعالة لتحسين أداء منظومة تعريف الشخص بالاعتماد على خصائص تحويل المويجة المتعددة التحليل. تم تحليل أشارة الكلام الداخلة الى L من الحزم. للحصول على خصائص الحبال الصوتية تم أستخدم مشفرة التخمين الخطي(LPC) (من ضمنها التخمين الخطي للحزمة الكاملة).

تم دمج الصفات المميزة لـــ(LPC) لكل حزمة مع LPC للحزمة الكاملة في متجه واحد وبعد ذلك تم أستخدام مقياس المسافة (Euclidean) لقياس التشابه بين الاشارة المرجعية والاشارة المختبرة. وضحت نتائج الاختبار ان أستخدام الطريقة المقترحة أعطت نتائج أفضل من منظومة التمييز بأستخدام LPC و Real Cepstral Coefficients.

1- Introduction

Speaker recognition is the process of automatically recognizing who is speaking on the basis of individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify his identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas and remote access to computers [1].

Speaker recognition can be classified into identification and verification. Speaker verification refers to the process of determining whether or not the speech samples belong to some specific speaker. On the other hand, Speaker identification is the process of determining which registered speaker provides a given utterance (word or phrase).

Speaker recognition methods can also be divided into text-independent and text-dependent methods. In a text-independent system, speaker models capture characteristics of what one is saying, while in a text-dependent system the recognition of the speaker's identity is based on his or her speaking one or more specific phrases, like passwords, card numbers, etc [2].

Many researches have been done on the feature extraction of speech. The linear predictive code (LPC) was used because of their simplicity and effectiveness in speaker recognition [3]. Other widely used feature parameters, namely, cepstral coefficients. Cepstral coefficients and their time derivatives are used as features in order to capture dynamic information and eliminate time-invariant spectral information that is generally attributed to the interposed communication channel [4].

In this paper, the multiband linear predictive code (MBLPC) is used in speaker identification system. This method is based on the multiresolution of the wavelet transform. The input speech signal is decomposed into L subband then the linear predictive code of each band (including the LPC for full band) are calculated. The feature recombination and distance measure methods are used to evaluate the task of speaker identification. This paper is organized as follows. Feature extraction is described in section 2. Distance measure is described in section 3. Section 4 presents the multiband speaker identification model. Experimental results are presented in section 5. Concluding remarks are made in section 6.

2- Feature Extraction

2-1 Linear Predictive coding (LPC): [5]

One of the most powerful speech analysis techniques is the method of linear predictive analysis. This method has become the predominant technique for estimating the basic speech parameters, e.g., pitch, formants, spectra, vocal tract area functions and for representing speech for low bit rate transmission or storage. The importance of this method lies both in its ability to provide the speed and extremely accurate estimates of the computation. The basic idea behind LPC analysis is that a speech sample can be approximated as a linear combination of past speech samples. By minimizing the sum of the squared differences (over a finite interval) between the actual speech samples and the linearly predicted ones.

It is assumed that the variations with time of the vocal tract shape can be approximated with sufficient accuracy by a secession of stationary shapes. It is possible to define an all-pole transfer function H(z) that produces the output speech s(n) given the

input excitation u(n) (either an impulse or random noise) is given by:

$$H(z)=\frac{S(z)}{U(z)}=\frac{G}{1-\sum_{k=1}^{p}a_k z^{-k}} \qquad (1)$$

Thus, the linear filter is completely specified by scale factor G (gain factor) and p predictor coefficients $a_1,\ldots,a_p$. The number of coefficients p required to represent any speech segment adequately is determined by many factors, such as the length of the vocal tract, the coupling of the nasal cavities, the place of the excitation and the nature of the glottal flow function.

A major advantage of the all-pole model of the speech production is that it allows one to determine the filter parameters in a straight-forward manner by solving a set of linear equations. In the all-pole model, the speech sample s(n) at n^{th} sampling instant is related to the excitation, u(n) by the following equation:

$$s(n)=\sum_{k=1}^{p}a_k s(n-k)+Gu(n) \qquad (2)$$

where u(n) is the n^{th} sampling of the excitation and G is the gain factor. Equation (2) represents the LPC difference equation, which shows that the value of the present output may be determined by summing the weighted present input, Gu(n),and the weighted sum of the post output samples. If the excitation u(n) is white noise, the best estimate of the n^{th} speech sample based on speech samples is given by:

$$\hat{s}(n)=\sum_{k=1}^{p}a_k s(n-k) \qquad (3)$$

where $\hat{s}(n)$ is called the predicted value of s(n) and a_k is the predictor coefficient. The prediction error between the actual speech sample and the predicted sample is defined as:

$$e(n)=s(n)-\hat{s}(n) \qquad (4)$$

$$=s(n)-\sum_{k=1}^{p}a_k s(n-k) \qquad (5)$$

which is the output of a system whose transfer function is:

$$A(z)=\frac{e(z)}{S(z)}=1-\sum_{k=1}^{p}a_k z^{-k} \qquad (6)$$

where A(z) is the transfer function of the predictor error filter or the inverse filter for the system H(z). To determine the filter coefficients, a_k, the mean squared prediction error is minimized over a short-segment of speech (N). The average square of the prediction error becomes :

$$E_m=\sum_{n=0}^{N-1}e^2(n)=\sum_{n=0}^{N-1}\left[s(n)-\sum_{k=1}^{p}a_k s(n-k)\right]^2 \qquad (7)$$

The values of the estimated predictor coefficients can be determined by minimizing the partial derivatives of E_m with respect to a_k.

$$\frac{\partial E_m}{\partial a_k}=0\ (k=1,2,\ldots,p) \qquad (8)$$

This yields p linear equations:

$$\sum_{n=0}^{N-1}s(n-i)s(n)=\sum_{k=1}^{p}a_k\sum_{n=0}^{N-1-k}s(n-i)s(n-k) \qquad (9)$$

where i=0,1,…,p and k=1,2,……,p.

Defining

$$R(i) = \sum_{n=0}^{N-1-i} s(n)s(n+i) \qquad (10)$$

Then, Equation (10) can be expressed by matrix representation as:

$$\begin{bmatrix} R(0) & R(1) & .... & R(p-1) \\ R(1) & R(0) & .... & R(p-2) \\ & & & \\ R(p-1) & R(p-2) & .... & R(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \\ a_p \end{bmatrix} = \begin{bmatrix} R(1) \\ R(2) \\ \\ R(p) \end{bmatrix} \qquad (11)$$

The $p \times p$ autocorrelation matrix of the term has the form of a Toeplitz matrix, which is symmetrical and has the same values along the lines parallel to the main diagonal. This type of equation is called a Yule-Walker equation. Since the positive definition of the autocorrelation matrix is guaranteed by the definition of the autocorrelation function, an inverse matrix exists for the autocorrelation matrix. Solving the equation permits obtaining a_k.

The equation for the autocorrelation method can be effectively solved by the Durbin's recursive solution method.

2-2 Real Cepstral Coefficient (RCC)

If s(n) is the input sequence, the real cepstral coefficient C_{rc} (n) can be calculated by the following equations [4]

$$S(k) = \sum_{n=0}^{N-1} s(n) e^{-j\frac{2\pi}{N}kn} \quad 0 \le k \le N-1 \qquad (12)$$

$$\hat{S}(k) = \log|S(k)| \quad 0 \le k \le N-1 \qquad (13)$$

$$c_{rc}(n) = \frac{1}{N} \sum_{k=0}^{N-1} \hat{S}(k) e^{j\frac{2\pi}{N}kn}$$

$$0 \le n \le N_c - 1 \qquad (14)$$

where Equation (12) is the DFT of the input sequence, Equation (13) gives the logarithm of the absolute value of the DFT of the input, and Equation (14) gives the real cepstral coefficient of the input sequence. N_c is the cepstral coefficient order.

The real cepstrum is mainly used as feature vector as an improvement over the direct usage of LPC based cepstral features of a given speaker in the process of speaker identification.

The cepstrum, however, ignores the phase of the time-dependent Fourier representation and therefore, the time-dependent cepstrum cannot uniquely represent the speech waveform.Nevertheless, it is seen that the cepstrum is a convenient basis for estimating pitch, voicing and formant frequencies.

The real cepstrum can also be found from the spectrogram of the signal instead of the spectral component, therefore, Equations (12-14) can be rewritten as [6]

$$S_p(k) = \log(|S(k)|^2) \qquad (15)$$

$$c_{rcs}(n) = \frac{1}{N} \sum_{k=0}^{N-1} S_p(k) e^{j\frac{2\pi}{N}nk} \quad 1 \le n \le p \qquad (16)$$

where S_p (k) is the natural logarithm of the spectrogram of the signal s(n) and c_{rcs} (n) is the real cepstral based on the spectrogram of the signal. Figure (1) shows the block diagram of the real cepstrum

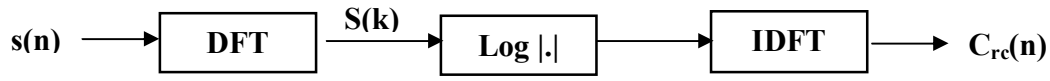

Figure (1) Block diagram of the real cepstrum.

2-3 Discrete Wavelet Transform (DWT)

The general form of an L-level DWT is written in terms of L detail sequences, $d_j(k)$ for j=1,2,….,L, and the L-th level approximation sequence, $c_L(k)$ as follows [7]:

$$f(t) = \sum_k c_L(k)\ \phi_L(t) + \sum_{j=1}^{L}\sum_k d_j(k)\,\psi_j(t) \qquad (17)$$

where $\phi_L(t)$ is the L-th level scaling function and $\psi_j(t)$ for j=1,2,…,L are wavelet function sequences for L different levels.

In order to work directly with the wavelet transform coefficients, the relationship between the detailed coefficients at a given level in terms of those at previous level is used. In general, the discrete signal is assumed the highest achievable approximation sequence, referred to as 0-th level scaling coefficients. The approximation and detail sequences at level j are given by [7]:

$$c_{j+1}(k) = \sum_m h_o(m-2k)\,c_j(m) \qquad (18)$$

and

$$d_{j+1}(k) = \sum_m h_1(m-2k)\,c_j(m) \qquad (19)$$

Equations (18) and (19) state that approximation sequence at higher scale (lower level index), with the wavelet and scaling filters, $h_o(t)$ and $h_1(t)$ respectively, can be used to calculate the detail and approximation sequences (or discrete wavelet transform coefficients) at lower scales.

The scaling coefficients are related to wavelet coefficients by:

$$h_1(n) = (-1)^n h_o(N-n) \qquad (20)$$

where N is a finite odd length of quadrature mirror filter.

Let the function f(t) be a discretely sampled function. The decomposition of f(t) in the wavelet basis is done by recursive filtering with H_o and H_1 with down-sampling of factor of two in each set. A lower resolution signal is delivered by low pass filtering with half-band low pass filter H_o followed by down-sampled by two. The higher resolution (or detail) is computed by a high pass filter H_1 followed by down-sampling by two [7].

The coefficients $h_o(n)$ and $h_1(n)$, used to construct the set of scaling and wavelet basis, are low pass (H_o) and high pass (H_1)FIR filter coefficients respectively. $H_o=\{h_o(n)\}$ and $H_1=\{h_1(n)\}$. According to the Equation (20), H_1 is the reverse of H_o. [11]

Figure (2) shows filter bank of discrete wavelet transform. The symbol ↓2 is down-sampler (decimator) that it takes a signal x(n) as input and produces an output of y(n)=x(2n), which means half of the data is discarded.

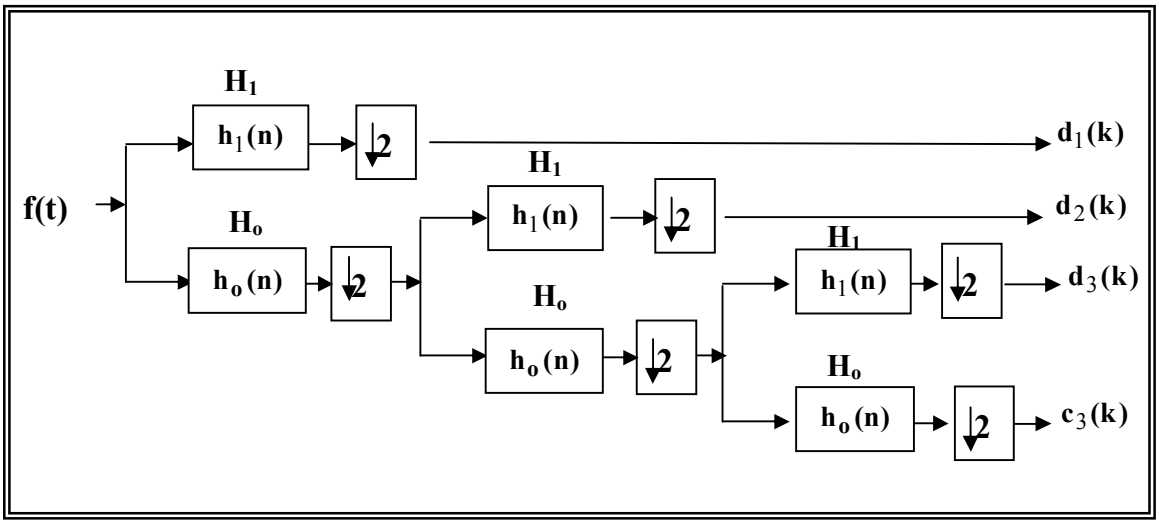

Figure (2) Filter bank of discrete wavelet transform

3- Distance Measure

For the speaker identification task, the unknown the speech is compared with all reference speech. This can be done through a distance measure. A simple geometric distance measure can be used. That is the Euclidean distance measure. The Euclidean distance can be defined as [8]:

$$D(x-y)=(a_x-a_y)^T(a_x-a_y) \quad (21)$$

where a_x and a_y are prediction coefficients for reference and tested speech respectively.

The decision rule is to select the Pattern that best matches the unknown. In this approach, the minimum distance classifier is used. This classifier assigns the unknown speech pattern to the nearest reference speech pattern.

4- Multiband Linear Predictive Code (MBLPC) Speaker Identification Model

Figure (3) shows speaker identification using multiband combination feature model.

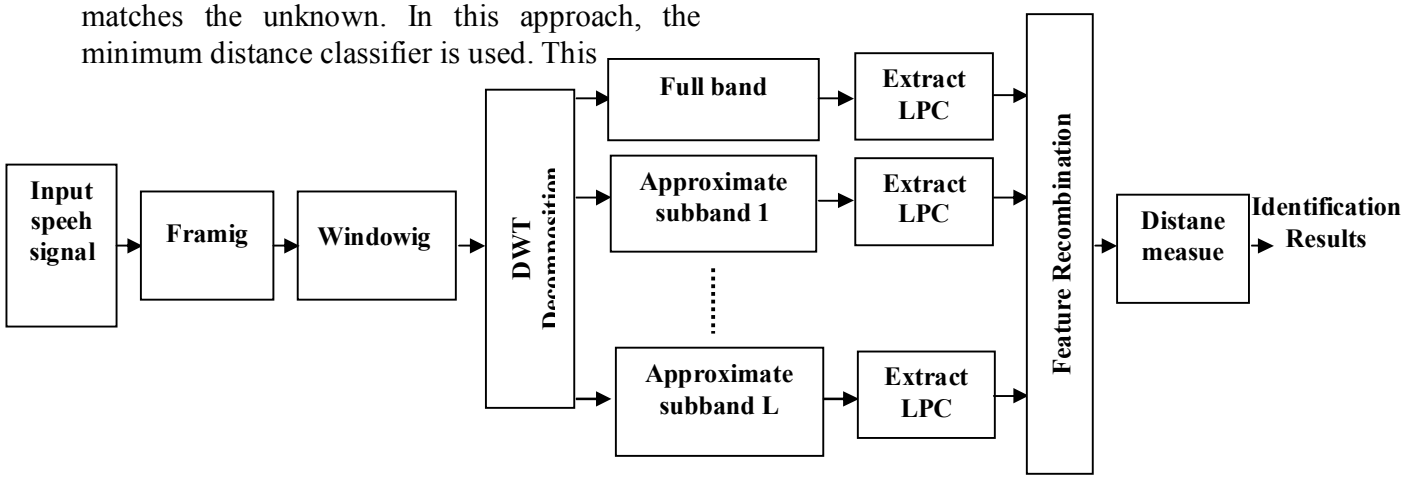

Figure (3) Block diagram of speaker identification using multiband combination feature model

Procedure

1. Framing the input speech signal.
2. Windowing the input speech signal by hamming window.
3. Obtaining wavelet transform decomposition of the input speech signal using different types of wavelet family.
4. Obtaining the approximate coefficients from the wavelet transform.
5. Extracting the LPC features from each band (including full band).
6. Recombining the LPC from each band and full band in a single feature vector.
7. Feature matching performs the similarity measure between the test and reference templates using the Euclidean distance measure.

5- Experimental Results

Simulations of speaker ident-ification using Multiband Linear Predictive Code (MBLPC) is carried out. The speech signal is sampled at 16 KHz using a computer sound blaster (in normal room conditions). The speech samples are quantized into 16 bit. The continuous speech signal is sectioned into frame of N with adjacent frames overlapping of M samples. Typically chosen values of N and M are 320 samples (about 20 ms) and 128 samples (about 8 ms) respectively. All the experiments were performed using section of speech from 15 speakers. Table (1) shows identification rate using LPC and RCC as feature extraction.

Table (1) Identification rate results using LPC and RCC as features extraction

Description	Identification rate %
LPC	73.333
RCC	80

Table (2) shows identification rate using MBLPC model with two bands and different types of wavelet family (db2, db4, db6, db8 and db10).

Table (2) Identification rate results using MBLPC model

Wavelet family	Identification rate %
db2	93.333
db4	80
db6	86.667
db8	80
db10	86.667

Table (3) shows identification rate using MBLPC model with three bands and different types of wavelet family.

Table (3) Identification rate results using MBLPC model

Wavelet family	Identification rate %
db2	93.333
db4	80
db6	86.667
db8	80
db10	93.333

Table (4) shows identification rate using MBLPC model with four bands and different types of wavelet family.

Table (4) Identification rate results using MBLPC model

Wavelet family	Identification rate %
db2	86.667
db4	80
db6	86.667
db8	80
db10	86.667

Figure (4) shows comparison between level 3 and level 4 speaker identification rate for different types of wavelet family.

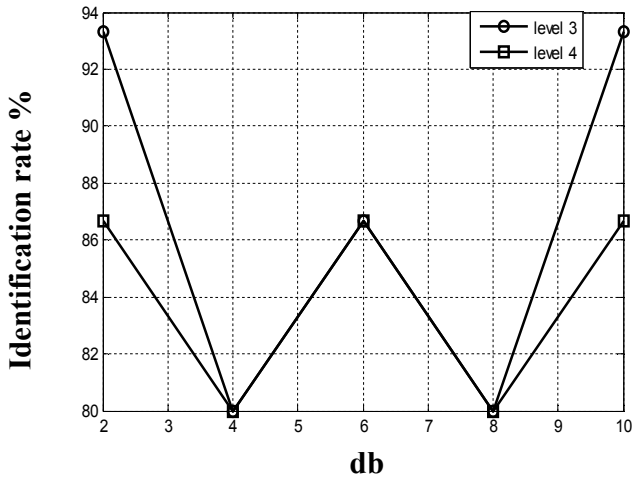

Figure (4) comparison between level 3 and level 4 speaker identification rate for different types of wavelet family

From these tables and figure (4), it is found that increasing the number of bands too more than three bands not only increased the computation time but also decreased the identification rate. In this case, the signals of the lowest frequency subband where located in the very low frequency region, which put too much emphasis on the lower frequency spectrum of speech.

6- Conclusions

The following points are concluded from the simulation results:

1. The real cepstral coefficient (RCC) gives good results for speaker identification rate compared with linear predictive code (LPC).
2. The MBLPC model gives higher identification rate compared with LPC and RCC.
3. The MBLPC model gives higher identification rate with different bands in db2.
4. The MBLPC model gives bad identification rate (80%) with different bands in db4 and db8.
5. Speaker identification using MBLPC with four bands gives bad results for identification rate compared with three bands.

7- References

[1] Richard J. Mammone et al., "Robust Speaker Recognition", IEEE Signal Processing Magazine, September 1996.

[2] Wan-Chen Chen, Ching-Tang Hsieh, and Eugene Lai, "Multiband Approach to Robust Text-Independent Speaker Identification", Computational Linguistic and Chinese Language Processing, Vol. 9, No.2, PP. 63-76, August 2004.

[3] Chi-Shi Liu, "A General Framework of Feature Extraction: Application to Speaker Recognition", IEEE Transactions on Acoustics and signal processing, Vol. 2, PP. 669-672, 1996

[4] Fadel S. Hassen, "Cepstral based speaker Recognition System", MSc. Thesis, Mustansiria University, 2003.

[5] L. R. Rabiner, and Ronald W. Schafer, " Digital Processing of Speech Signals ", Prentice-Hall, New Jercy, 1978.

[6] Yarive Epharim, and Mazin Rahim, "On Second Order Statistics and Linear Estimation of Cepstral Coefficients", IEEE Transaction Speech and Audio Processing, Vol.7, No.2, March 1999.

[7] Raghuveer M. Rao and Ajit S. Bopardikar,"Wavelet Transform: Introduction to Theory and Application", Addison Wesley Longman, Inc 1998.

[8] Hema A. Murthy, Francoise beaufays, Larry P. Heck, and Mitchell Weintraub, "Robust Text-Independent Speaker Identification over Telephone Channel", IEEE Transaction on Speech and Audio Processing, Vol.7, N0.5, September 1999.