An High Efficiency Feature Extraction Based on Wavelet Transform for Speaker Recognition

Ching-Han CHEN , Chia-Te CHU

Institute of Electrical Engineering,I-Shou University,1, Section 1, Hsueh-Cheng Rd.Ta-Hsu Hsiang Kaohsiung County, Taiwan, 840, R.O.C.

E-mail: pierre@isu.edu.tw , cld123@giga.net.tw

http://140.115.11.235/~chen/

Abstract*-To speaker recognition problem, firstly this paper will study and compare to various feature extraction methods include LPCC, PCA, fractal, and wavelet transform, which combined probabilistic neural network classifier. We carry out a set of experiments in speaker identification and matching .The result reveals Fractal has the best efficiency and discrete wavelet transform has the excellently high recognition rate. Besides, we will apply wavelet transform to reduce data dimension and enhance discriminative feature in speech signal, and combine LPCC, PCA, or fractal for feature extraction. The advantage of these mixed methods has the discriminative features in speaker recognition, saving system resource and speeding up recognition time. From our speech database, the average recognition of WT+LPCC in 10 times tests is 99.5% and the EER of speaker matching is 0.0. This shows the feature extraction method is combined with wavelet has excellently efficiency and performance.*

Keyword:speaker recognition; wavelet transform; probabilistic neural network; fractal

1. Introduction

Speaker recognition system determines identity of a speaker according to the information in speech signal. Its applications included information services, voice-mail, security control for confidential information areas, and remote access to computer. Nevertheless these algorithms to perform speaker recognition are high complexity and memory requirements, as not to be suitable for high efficiency applications. Thus, the simple and high reliable speaker recognition system is desired.

To address the problems related complexity and memory requirements, the aim of this paper is to provide a low complexity, best performance and high efficiency speaker recognition system. Consequently, we propose the combination of wavelet transform together with a feature extraction method such as LPCC [3], [4], PCA [4], [7] and fractal [5]. The wavelet transform [1], [2] is regarded as preprocessing of speech signal. The goal of preprocessing is to reduce the dimension and enhance the robust feature of the speech signal. By selecting a number of voiceprint feature, feature are fed to the probabilistic neural network (PNN) for a series of experiment in speaker identification and verification. The results show the proposed method has the best performance, high efficiency and very low complexity.

2. Speaker Feature Extraction

2.1 LPC-Derived Cepstral Coefficients (LPCC)

The advantage of feature extraction is for the dimension reduction and representation of original signal. The LPCC has been widely used to extract feature in speech signal. Finally, some existing methods are also evaluated for comparison.

The linear predictive coding (LPC)[3][4] is one of the most popular techniques for speech analysis. The LPC generates prediction errors e(n).If the all-pole model was good, e(n) would be very small. Thus, e(n) can be stated as the ideal excitations of the all-pole model. In speaker recognition, the LPC coefficients will be transformed into LPCC as feature vectors, because the robust and reliability of the LPCC coefficients is better that LPC ones.

Autocorrelation sequence $R(k)$ are obtained in the analysis of speech signals x.

$$R(k)=\sum_{n=-\infty}^{\infty}x(n)x[n+k]R(k)\quad k=0,\cdots,p+1$$

Where n is the sample index, and k is time shift.

The Levinson-Durbin algorithm is an iterative method of computing the LPC coefficients.

$$E^{(0)}=R(0)$$
$$for\quad i=1:p$$
$$k_i=R(i)-\sum_{j=1}^{i-1}a_j^{(i-1)}R(i-j)\quad a_i^{(i)}=k_i/E^{(i-1)}$$
$$for\quad j=1:i-1$$
$$a_j^{(i)}=a_j^{(i-1)}-k_i a_{i-j}^{(i-1)}$$
$$end$$
$$E^{(i)}=(1-k_i^2)E^{(i-1)}$$
$$LPC\quad coefficients=a_i=a_i^{(p)}\quad 1\le i\le p$$

After LPC coefficients was obtained, we will obtain LPCC coefficients through transform:

$$c_m=a_m+\sum_{k=1}^{m-1}\left(\frac{k}{m}\right)c_k a_{m-k}\quad 1\le m\le p$$

Since the LPCC coefficients can improve the robust and reliability of features extract from speech signal, they are also regarded as feature vectors.

2.2 Principal Component Analysis (PCA)

The purpose of PCA [3][7] is to find a set of new basis vectors, which can reduce the dimensions of feature vectors. All the speech vectors are as the rows of a matrix X. M is the mean vector of X. Thus the covariance matrix $E=(X-M)(X-M)'$ is transformed to a diagonal matrix $A=P'EP$. The columns of the matrix P are the eigenvectors of E and they are the principal components. As the principal components are very small, we can discard they. Then the dimension of feature vectors is reduced.

2.3 Iterated Function System (IFS) and Fractals

IFS is capable of effectively describing complex shapes and textures by fractals. Most natural phonemes are irregular and rough. A new and popular technique is maturing for which representation of shape and texture is effected by a class of deterministic fractals. Mandelbrot presented the idea of fractals in 1973 and Barnsley developed the IFS theory over the past few years. Barnsley emphasized on the practical application of natural fractals. IFS is a tool to generate fractals. It is a collection of contractive functions or mapping which when iteratively applied converges to the fixed point. The limit of this iterative

process or the fixed point is called the attractor of the IFS[5].

W_i are affine transforms of a special nature with following boundary conditions:

$$W_i\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} a_i & 0 \\ c_i & d_i \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} e_i \\ f_i \end{bmatrix}$$

$$W_i\begin{bmatrix} x_0 \\ F_0 \end{bmatrix} = \begin{bmatrix} x_{i-1} \\ F_{i-1} \end{bmatrix}$$

$$W_i\begin{bmatrix} x_N \\ F_N \end{bmatrix} = \begin{bmatrix} x_i \\ F_i \end{bmatrix}$$

In speech signal, x is the time domain and y is frequency domain. The interval is i.

By expanding the above matrix equations, the following equations are used to obtain IFS variables.

$$a_i x_0 + e_i = x_{i-1}$$
$$a_i x_N + e_i = x_i$$
$$c_i x_0 + d_i F_0 + f_i = F_{i-1}$$
$$c_i x_N + d_i F_N + f_i = F_i$$

a_i, c_i, d_i, e_i and f_i are IFS variables.

The IFS variables a and e are related to the time scale, they are not used here as variables. The d_i is estimated with the following equation.

$$d_i = \frac{|F_{\max}|}{(F_i \max - F_i \min)/2}$$

Let F_i max and F_i min be the maximum and minimum sample amplitude in the interval i and Fmax the maximum amplitude in the whole sample.

The purpose for IFS variables c, d, and f are to obtain eigenvalues of covariance. The superscript * is regard as raw data before standardization.

$$x_i^* = [c_i \quad d_i \quad f_i]$$

$$A^* = [x_1^* \quad x_2^* \quad \cdots \quad x_N^*]^T$$

$$m_j = \frac{1}{N}\sum_{i=1}^{N} x_{ij}^*$$

$$S_j^2 = \frac{1}{N}\sum_{i=1}^{N}(x_{ij}^* - m_j)^2$$

Let m_j be average and S_j^2 be the variance of the jth feature. Thus the standardized data x_{ij} can be obtained.

$$x_{ij} = \frac{(x_{ij}^* - m_j)}{S_j}$$

Then

$$A = [x_1 \quad x_2 \quad \cdots \quad x_N]^t$$

Finally, the covariance matrix R is defined as

$$R = \frac{1}{N} A^t A$$

The advantage of fractal is representation of signal by few feature vectors. This can save system resource and improve efficiency.

IFS coefficients are combined with covariance matrix to obtain egienvalues for speaker recognition. However, experiments reveal that the method is not proper for speaker recognition. To overcome this problem, we resort to wavelet transform to decompose speech signal into high and low frequency, and analysis on several approximations signals.

2.4 Wavelet Transform

The wavelet [1], [2], [9], [10] is constructed from two-channels filter bank as fig (2). In wavelet decomposition of speech signal, a speech signal is put through both a low-pass filter L and a high-pass filter H and the results are both low frequency components A[n] and high frequency components D[n]. The signal y [n] is reconstructed by the construction filters $\tilde{H}$ and $\tilde{L}$.

The wavelet filters are used to decompose signal s into high and low frequency by

convolution.

$$D[n]=\sum_{k=-\infty}^{\infty} s[k]\cdot H[n-k] \Leftrightarrow D=\langle s,H\rangle$$

$$A[n]=\sum_{k=-\infty}^{\infty} s[k]\cdot L[n-k] \Leftrightarrow A=\langle s,L\rangle$$

In order to construct multichannel filter, we can cascade channel filter banks. Fig (3) is a 3-level symmetric octave structure filter bank. This is an important concept from mu lti-resolution analysis (MRA).

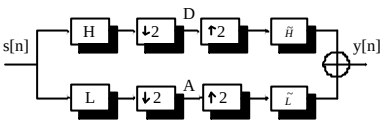

Fig 2. Two-channels filter bank

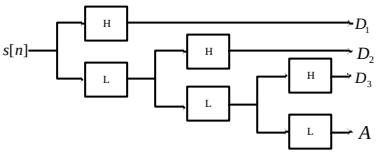

Fig 3. 3-level octave band filter bank

- **Wavelet feature extraction (WTFT)**

The detail coefficients can be used to estimate the average energy contained in the different resolution. The total average energy is the sum of the average energy contained in the detail and the lower level approximation coefficients. As a result, the ith element of a feature vector is given by

$$v_i=\frac{1}{n_i}\sum_{j=1}^{n_i} w_{i,j}^2 \quad i=1,2,\cdots,N+1$$

Where n_i is the sample number of each resolution and $w_{i,j}^2$ is the jth coefficient and ith subband. As a result, a feature vector is formed as given by

$$V=\{v_1,v_2,\cdots,v_i\}^t$$

3. Probabilistic Neural Network (PNN)

In 1988, D.F.specht have designed a very efficiency probabilistic neural network (PNN)[8] that is well adapted to manipulate classification problem. The purpose of this paper is for speaker recognition. The experiment reveals it is excellent in efficiency and performance.

The basic concept cited Bayesian classifier to PNN model as fig 4. To probability density function, it has three assumptions:

1. The classification of probability density function is the same.
2. Probability density function is Gaussian distribute.
3. The variance matrix of Gaussian distribute probability density function is diagonal matrix.

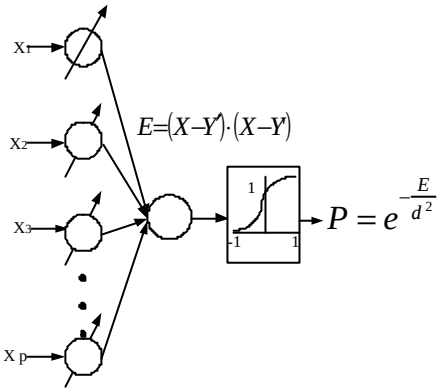

Fig 4. The simplified structure of PNN model

When the external factor is change, the PNN only change the weight of new data. The other neural network needs not to change all network weights.

The PNN model has been used for classification, because of its simplicity, performance and efficiency. Hence, this paper adopt PNN model as classifier.

4. Speaker Recognition Experiment

4.1 Experiment Prepare

In the set of experiments, this paper compares the performance and efficiency of the speech signals processing by applying wavelet and without wavelet. Thus, the speech feature is extracted and divided into two databases. The combination of feature extraction methods and PNN is to verify or identify the speaker. We performed the Chinese digit words are speech data. The speech data consists of 4 male speakers. Each of these speakers uttered "xu zi heng" about 12-20 times in 4 weeks. The speech samples were recorded using microphone in office environment and sampled at 11k Hz with an 8bit digitizer. According to individual speaker, we randomly select 5 speech data as train data; the others are as test data.

4.2 Speaker Identification in Various Feature Extraction Methods

To extract speech feature by LPCC, PCA, WTFT or fractal is combined with PNN for speaker identification. The samples are randomly selected as train data and repeated 10 times. The result is as table 1.

Table 1. Speaker identification with different feature extraction methods

	Fractal	PCA	LPCC	WTFT
1	0.65	0.6	**1**	0.95
2	0.75	0.8	0.95	**1**
3	**0.85**	0.6	0.95	0.95
4	0.8	**0.75**	0.95	0.95
5	0.65	0.7	1	0.95
6	0.7	0.6	1	1
7	0.65	0.6	0.95	0.95
8	0.65	0.65	0.8	1
9	0.75	0.75	0.95	0.95
10	0.65	0.7	0.95	1
Average rate	0.71	0.67	**0.95**	**0.97**
dimension	36	10	132	48
Feature extraction time (ms/sample)	100	**22**	91	154
Recognition time (ms/sample)	170	70	561	221

4.3 The Combination of Wavelet and Feature Extraction Methods in Speaker Identification

Here, wavelet transform is used as preprocessing of speech signal. This aim for this experiment is to reduce dimension and enhance feature of speech signal. There are three methods for feature extraction.

Fractal: By wavelet transform, the speech signal is decomposed into low frequency extracted feature by fractal.

LPCC: By wavelet transform, the speech signal is decomposed into low frequency extracted feature by LPCC.

PCA: By wavelet transform, the speech signal is decomposed into low frequency extracted feature by PCA.

The result is as table 2.

Table 1. The combination wavelet transform and different feature extraction methods in speaker identification

	WT+Fractal	WT+LPCC	WT+PCA
1	0.95	1	0.65
2	0.95	1	0.6
3	0.85	1	0.75
4	0.90	1	0.65
5	0.95	0.95	0.5
6	0.85	1	0.7
7	0.85	1	0.65
8	0.9	1	0.6
9	0.85	1	0.4
10	0.90	1	0.6
Average recognition rate	0.895	**0.995**	0.61
Dimension	36	84	10
Feature extraction time (ms/sample)	77	**61**	31

Recognition time (ms/sample)	**170**	**360**	**70**

4.4 Speaker Verification Experiments

The performance of the speaker verification [12]is estimated with the EER. Table 3 is traditional feature extraction method in speaker verification. Table 4 is the combinations of wavelet transform and feature extraction method in speaker verification.

Table 3. Various feature extraction methods in speaker matching.

	Fractal	LPCC	PCA	WTFT
EER	0.1917	0.1167	0.4	**0.05**

Table 4. The combinations of wavelet transform and feature extraction method in speaker verification.

	WT+Fractal	WT+LPCC	WT+PCA
EER	0.1	**0**	0.38

4.5 Experimental Results

The results obtained are shown in Table 1. As can be seen, the best performance is WTFT in 0.97. The best efficiency and shortest dimension is fractal in 270 ms and 36.Thus, the advantage of fractal is suitable for small system resource and high efficiency.

5. Conclusion

In this paper, we explored the combination of wavelet, feature extraction method, and PNN for speaker recognition. From these experiments, it shows that the best efficiency is WT+fractal and the best performance is WT+LPCC. They provide much better results in efficiency and performance.

Owing to the low computational load of the proposed methods makes them become commercial speaker recognition system in future, because the advantage of them is in simple algorithm and fast computation.

References

[1] Jaideva C.Goswami and Andrew K. Chan, "*Fundamentals of Wavelets*" 1999.

[2] Olivier and Martin Vetteru "Wavelets and Signal Processing,"*IEEE SP MAGZINE* pp.14-38, 1991 October.

[3] Torres Humberto M., Rufiner Hugo L. "Automatic Speaker Identification by means of Mel Cepstrum, Wavelets and Wavelets Packets." *Proceeding of the 22nd Annual EMBS international Conference IEEE*, pp.978-981,July 23-28,2000.

[4] John Holmes and Wendy Holmes "*Speech Synthesis and Recognition*," 20001.

[5] Erik L.J Bohez*, T.R. Senevirathne "Speech recognition using fractals"*Pattern Recognition Society* pp.2227-2243, 2001.

[6] Specht, D.F,"Probabilistic neural network for classification mapping, or associative memory",*IEEE international conference on neural networks*, vol.1,pp.300-303, 1995.

[7] Daniel Beyerbach, "*Principal Component Analysis of Time-Frequency Representations*," Ph.D. Thesis, ECS Dept.,Boston University, Boston, MA, MAY 1991.

[8] Francis Phan, Evangelia Micheli-Tzanakou, and Samuel Sideman."Speaker Identification Using Neural Networks and Wavelets " I*EEE Engineering in Medicine and Biology*, pp92-101, January/February 2000.

[9] M. Vetterli and J. Kovacevic. *Wavelets and Subband Coding*. Prentice Hall, 1995.

[10] Yu Yue*, Zhou Jian, Wang Yiliang, Li Fengting and Ge Chenghui "On the computation of wavelet series transform," *Proceeding of ICSP* 1998.

[11] Walter G G.,"A sampling theorem for wavelet subspaces," *IEEE Trans on Information. Theory*, Vol. 38, No.3, 1992,pp.881-884.

[12] Maio, D.; Maltoni, D.; Cappelli, R.; Wayman, J.L.; Jain, A.K ., "FVC2000: fingerprint verification competition", *IEEE Transactions on Pattern Analysis and Machine Intelligence*,, Volume: 24 Issue: 3 , Mar 2002 Page(s): 402 –412.