International Journal of Innovation and Applied Studies
ISSN 2028-9324 Vol. 4 No. 2 Oct. 2013, pp. 353-358

http://www.issr-journals.org/ijias/

Voice identification Using a Composite Haar Wavelets and Proper Orthogonal Decomposition

Mohammed Anwer and Rezwan-Al-Islam Khan

School of Engineering and Computer Science,
Independent University, Dhaka, Bangladesh

ABSTRACT: In present day business and consumer environment, a robust voice identification system is needed to reduce false positives, and true negatives. In this work, a modified voice identification system is described using over sampled Haar wavelets followed by proper orthogonal decomposition. The audio signal is decomposed using over sampled Haar wavelets. This converts the audio signal into various non-correlating frequency bands. This allows us to calculate the linear predictive cepstral coefficient to capture the characteristics of individual speakers. Adaptive threshold was applied to reduce noise interference. This is followed by multi-layered vector quantization technique to eliminate the interference between multi-band coefficients. Finally, proper orthogonal decomposition is used to evaluate unique characteristics for capturing more details of phoneme characters. The proposed algorithm was used on KING and MAT-400 databases. These databases were chosen as previous extraction results were available for them. In the present study, the KING database were trained with three sentences, and tested with two. On the other hand, the MAT-400 database were trained with two seconds of random voice signal, and tested with other two seconds. Results were compared with vector quantization and Gaussian mixture models. The present model gave consistently better performance on speech collected through mouthpieces, but gave comparatively poor performance on audio collected on telephones. The better performance is obtained at the cost of higher computational time.

KEYWORDS: Voice identification, Haar wavelet, Proper Orthogonal Decomposition, Signal Processing, modeling.

1 INTRODUCTION

Generally, speaker recognition can be divided into two parts: speaker verification and speaker identification. Speaker verification refers to whether the speech samples belong to some specific speaker or not. However, in speaker identification, the goal is to determine which one of a group of known voices best matches the input voice samples. Certainly, how to extract and model the speaker-dependent characteristics of the speech signal is the key point. It seriously affects the performance of the system.

In past literatures for recognition models, vector quantization (VQ)[1]-[2], neural network (NN), dynamic time warping (DTW), hidden Markov model (HMM)[6], and Gaussian mixture model (GMM) were used in speaker recognition task. Some researchers combine VQ [3] and learning vector quantization (LVQ) [4] to form the group vector quantization (GVQ) that its performance is better than LVQ but it needs to be retrained after entering new samples. Also NN in speaker recognition task was used [5]. Although NN technique is robust and has high performance, it needs to be retrained after entering new samples. DTW technique [5] is not suitable for text-independent speaker recognition. GMM [7]-[10] were widely used in speaker recognition and had satisfactory performance.

In this work, an effective and robust extraction method of speech features called LPCC with multiband support based on wavelet analysis has been implemented. In order to effectively utilize all these multi-band speech features, a modified vector quantization method called eigen-codebook vector quantization (ECVQ) as the identifier. In these text-independent

Corresponding Author: Mohammed Anwer (manwer55@hotmail.com)

evaluations, the experimental results showed that this method has satisfactory computation cost and performance especially on the noisy environments.

2 MATHEMATICAL MODEL

2.1 COMPOSITE HAAR WAVELETS

The Haar wavelet transform has an established name as a simple and powerful technique for the multi-resolution decomposition of time series. Unfortunately, the standard Haar wavelet transform has several limitations. The lack of translational shift invariance is very limiting when matching is to be performed over various intervals of the input data. The composite Haar wavelet used in this work is especially useful for real-time signal analysis for our purpose.

In order to provide shift invariance, the Haar decomposition is over sampled. As for voice identification, over sampling operation will only be performed on the position axis [11]. The sampling grid for the scale axis will remain dyadic in scale. Oversampling on the position axis can be done to the highest resolution required, or any required intermediate, lower resolution. The highest available resolution that is that of the sampling rate of the input is being used. This means that we let the position translation index run over the positions *i* of the time series ψ_i. Therefore, the oversampled system will be as following [12]:

$$\psi_{m_0,n_0}(x) = 2^{-m_0}\psi(2^{-m_0}x - n_0), \qquad \left\{ \begin{array}{c} m_0 = 1,2,....,L \\ n_0 = BE,...2^L - 1 - BE \\ BE = 2^{m-1} - 1, \end{array} \right\}$$

Where, BE denotes empty boundary coefficients and is introduced here in order to ensure alignment of the indices across scales. The bands that are empty are half the width of the wavelet. They are 'empty' because position of the wavelet can be aligned with the time series at either end. An oversampled grid obtained according to the above prescription is shown in figure 1.

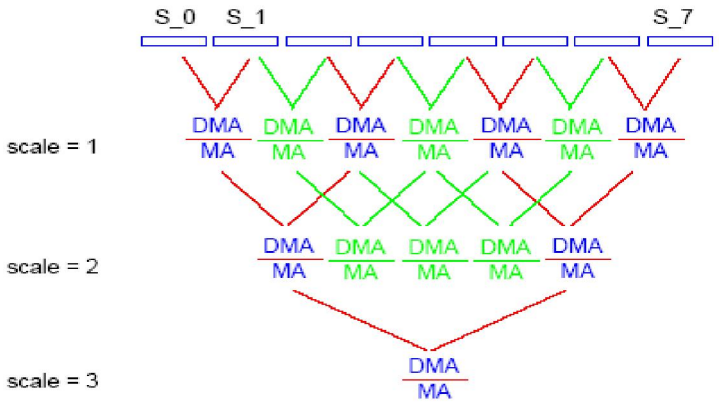

Fig. 1. Composite Haar decomposition. Highlighted (in red) is the dyadic pyramid [11]

MA=DMA is localized decomposition coefficients on the grid points. Of course it is possible to select the dyadic grid from the oversampled representation for any arbitrary starting location. One such possible (centered) selection has been highlighted in figure 1.

2.2 PRINCIPAL COMPONENT ANALYSIS

Assuming zero empirical mean (the empirical mean of the distribution has been subtracted away from the data set), the principal component w_1 of a dataset *x* can be defined as [13]:

$$w_1 = \arg\max_{\|w\|=1} E\left\{ \left(w^T x \right)^2 \right\}$$

There are two algorithms currently used to calculate PCA of a dataset that are: covariance method and the correlation method [13]. For this work covariance method is implemented.

3 SYSTEM DESCRIPTION

The target of implemented prototype system (figure 2) is to create voice signature of a user and then to isolate the target user from a group of speakers. The two distinct parts of this system one is voice signature creation and other one is speaker detection.

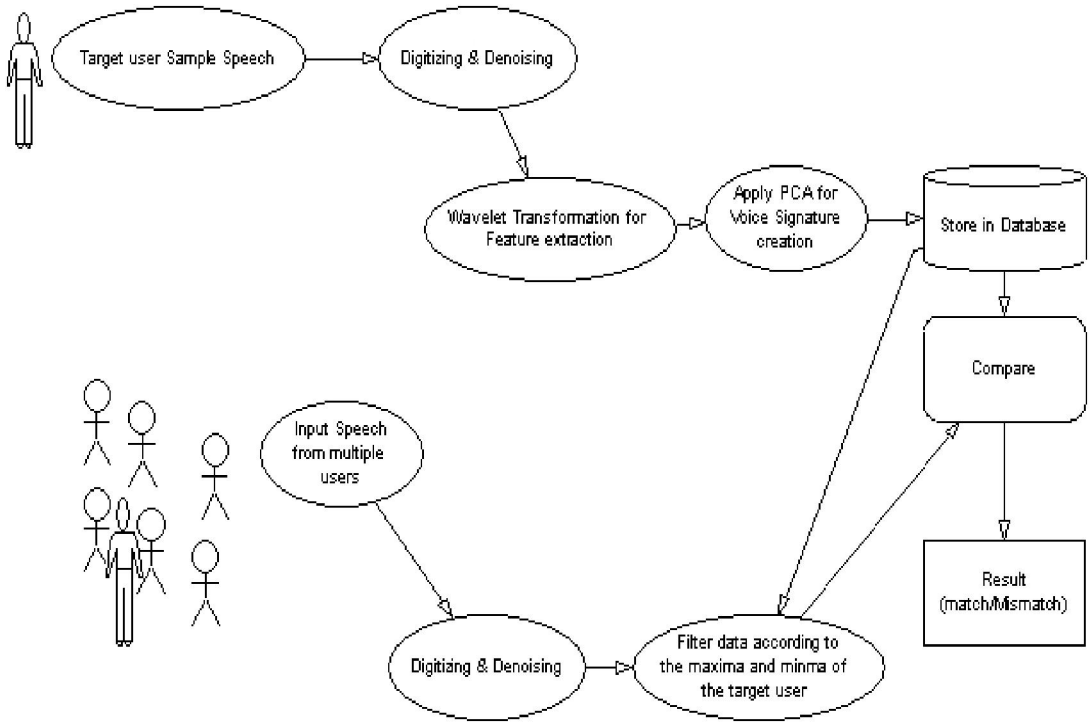

Fig. 2. Prototype System

3.1 FEATURE EXTRACTION METHOD

The wavelet transform is applied to decompose the input signal into various non-correlating frequency bands. Thus the linear predictive cepstral coefficients (LPCC) within all approximation channels are calculated to capture the characteristic of individual speaker. The main reason of using LPCC parameters is its good representation on the envelope of speech spectrum of vowel and its simplicity [9].

As the LPCC parameters are bothered by noise interference, an adaptive threshold technique is applied in each approximation channel before the next decomposition process. The most significant coefficients at each scale, with amplitude above some thresholds, are defined as follows:

$$\theta_j = \sigma_j - R_j$$

Where σ_j is the standard deviation of the wavelet coefficients within approximation channel at scale *j*, and *R* is an adjusting multiplicative factor used to restrict the threshold at a certain extent. The recursive decomposition process lets the system easily acquire the multi-band features of the speech signal. In the final stage, a combination of these LPCC values is implemented by the following equation:

$$LPCC = \underset{j=1}{\overset{L}{Append}}(LPCC_j)$$

Where *L* is the levels of decomposition process and $LPCC_j$ is the set of cepstral coefficients of the wavelet coefficients at scale *j.*

3.2 PCA FOR VOICE SIGNATURE CREATION

The PCA method [14]-[16] uses the statistical distribution of input samples to find the best projection bases. The advantages of PCA method are that the principal eigenvectors are orthogonal, and represent the directions where the signals have maximum energy. This property will speed up the convergence of model training and improve the recognition performance. While all training samples have been processed by PCA method, the evaluated mean vector $\overline{X}$ and eigenvectors $\phi_1, \phi_2, \ldots \phi_m$ may effectively describe the characteristic of samples belonging to specific word. If there is a new sample x_{new} for classification, it will be adjusted by the mean vector and project to the *m* orthonormal eigenvectors as follows:

$$p_i = (x_{new} - \bar{x})^T . \vec{\phi}_i \quad i=1,2,.....m$$

Where p_i is the projection of x_{new} on the *i*th projection base $\vec{\phi}$, and *m* is the number of the projection bases. Finally, the distance between ε and x_{new} the specific word is evaluated as follows:

$$\varepsilon = x_{new} - \bar{x} - \sum_{i=1}^{m} p_i \vec{\phi}_i$$

Where ε is a distance vector. The norm represents distance value between and the word. The new sample is belonging to the code word that is most closing to x_{new}. This training procedure known eigen-codebook vector quantization (ECVQ) [17] is described as follows:

Step 1: set up the number of the projection bases m, number of code words C_{no} maximum training times N.

Step 2: use VQ training method to calculate the centroids of all code words and classify all training samples.

Step 3: use PCA method to evaluate the mean vector $\bar{x}$ and eigenvectors $\phi_1, \phi_2, ... \phi_m$ for each code word. Figure 14 shows PCA output.

Step 4: use eigen distance formula to evaluate the distance vector ε and classify the training samples by ||ε|| .

Step 5: If the training times are less N and samples classification is not convergent, go to step 3.

Step 6: store the mean vector $\bar{x}$ and eigenvectors $\phi_1, \phi_2, \ldots \phi_m$ for each code word.

Because the LPCC is extracted from the wavelet coefficients of multi-band speech signals, in order to eliminate interference between the multi-band coefficients, the individual codebook for each band is evaluated. The total distance *total* ε between *new x* and the specific model is sum of ε in each band.

4 RESULT & DISCUSSION

The algorithm described here is applied on KING [17] and MAT-400 [18] database. KING is a database of 51 male speakers collected through microphones and telephone networks. Ten sections were recorded at different time for every speaker. MAT-400 is a Mandarin speech database of 400 speakers collected through telephone networks in Taiwan. Speakers include 216 males and 184 females. The speech signal is recorded at 8 kHz and 16 bits per sample. Furthermore, 16 orders of LPCC for each decomposition process were used. While computing LPCC, the mean normalization is used to compensate the channel effect.

In this experiment, the performance of ECVQ model is compared to conventional VQ and GMM model with Gaussian white noise corruption. In KING database, three arbitrary sentence utterances for each speaker are used as the training patterns and two seconds of speech waveform cut from the other two sentence utterances are used as the testing patterns. In MAT- 400 database, two balance sentences used as the training patterns and two seconds of other 8 balance sentences us as the testing patterns. In proposed ECVQ model, the decomposition level is 4, *j R* =0, each layer has 64 code words that are represented by a mean vector and a projection base. In VQ and GMM model, 20 orders of MFCC were used. In VQ, 100 code words were used. In GMM, 50 code words were used. The results in Table 1 show how the performance of the other models begins to degrade as the SNR of testing environment departs from that of the training environment. However, the performance of proposed ECVQ model is better than other models, and maintains its robustness at lower SNR.

Table 1. Identification of rates of the VQ, GMM, and MLECVQ for different SNR

Method	KING					MAT 400
	Clean (%)	20db (%)	15db (%)	Tel 30.8db	Tel 18.2db	Tel (%)
VQ	83.62	67.73	53.45	71.87	61.5	95.62
GMM	92.61	85.88	73.28	75.38	58.14	98.18
ECVQ	96.47	89.05	69.66	70.32	69.38	99.04

In other experiment with two sentence utterances of 20 speakers, the performance and computational cost of ECVQ is compared to conventional VQ and GMM model. The results are depicted in Table 2. In Table 2, *Y/ X* represents the results of 1 second and 2 second testing utterances. Although the computational time of VQ is much less than other methods, its performance also was lowest. When the number of code works was less than 64, the performance and computational time of the proposed ECVQ is better than GMM+VQ model.

Table 2. Identification rates and computations time

Code words	VQ		GMM		ECVQ	
	Recog. Rate (%)	Time (Sec)	Recog. Rate (%)	Time (Sec)	Recog. Rate (%)	Time (Sec)
16	57.26/74.12	0.15/0.25	80/90.92	1.61/3.19	84.14/92.1	1.2/2.45
32	68.78/80.84	0.25/0.5	83.83/93.28	3.05/5.95	86.81/94.79	2.65/4.85
64	69.45/84.03	0.45/0.95	84.67/93.45	7.38/11.35	88.48/96.47	8.3/16.6

5 CONCLUSION

An improved model of voice identification method using composite Haar wavelet is proposed in this work with addition of Principal Component Analysis method for detecting the pattern of speech characteristics of human voice. This model requires very little training with increased accuracy and language independence. This model may be improved further for artificial voice regeneration of a particular human voice.

REFERENCES

[1] Soong F. K., Rosenberg A. E., Rabiner L. R., and Juang B. H., "A vector quantization approach to speaker recognition," *Proceedings of ICASSP-85*, pp. 387-390, March 1985.

[2] Furui S., "Vector-quantization-based speech recognition and speaker recognition techniques," in *Proc. IEEE ICASSP*, pp. 954-958, 1991.

[3] He J., Liu L., and Palm G., "A discriminative training algorithm for VQ-Based speaker identification," *IEEE Trans. Speech Audio Processing*, vol. 7, no. 3, pp. 353-356, May 1999.

[4] Linde Y., Buzo A., and Gray R. M., "An algorithm for vector quantizer design," *IEEE Trans. Commun*., vol. 20, pp. 84-95, 1980.

[5] Furui S., "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-29, pp. 342- 350, June 1981.

[6] Tishby N. Z., "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Trans. Signal Process.,* 39, pp. 563-570, 1991.

[7] Reynolds D. A., and Rose R. C., "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Trans. Speech Audio Processing*, vol. 3, no. 1, pp. 72-83, 1995.

[8] Miyajima C., Hattori Y., Tokuda K., Masuko T., Kobayashi T., and Kitamura T., "Text-independent speaker identification using Gaussian mixture models based on multi-space probability distribution," *IEICE Trans. Inf. & Syst*., vol. E84-D, no. 7, pp. 847-855, July 2001.

[9] Alamo C. M., Gil F. J. C., Munilla C. T., and Gomez L. H., "Discriminative training of GMM for speaker identification," in *Proc. IEEE ICASSP*, pp. 89-92, 1996.

[10] Pellom B. L., and Hansen J. H. L., "An effective scoring algorithm for Gaussian mixture model based speaker identification," *IEEE Signal Processing Letters*, vol. 5, no. 11, pp. 281-284, Nov. 1998.

[11] Z.R. Struzik, "The Wavelet Transform in the Solution to the Inverse Fractal Problem," *Fractals* 3, No. 2, pp. 329-350, 1995.
[12] Z.R. Struzik, A.P.J.M. Siebes, *The Haar Wavelet Transform in the Time Series Similarity Paradigm*, in *Principles of Data Mining and Knowledge Discovery*, Eds: J.M. Zytkow, J. Rauch, Springer-Verlag, pp. 12-22, 1999.
[13] Pearson, K., "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosophical Magazine* 2, (6):559–572, 1901.
[14] Hurase H., and Nayar S., "Visual learning and recognition of 3D objects from appearance," *Int'l J. Computer Vision,* vol. 14, pp. 5-24, 1995.
[15] Belhumeur P. N., Hespanha J. P., and Kriegman D. J., "Eigenfaces vs. Fisherfaces: recognition using class specific linear projection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no. 7, pp. 711-720, July 1997.
[16] Martinez A. M., and Kak A. C., "PCA versus LDA," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228-233, Feb. 2001.
[17] Godfrey J., Graff D., and Martin A., "Public databases for speaker recognition and verification," in *Proc. ESCA Workshop Automat. Speaker Recognition, Identification, Verification*, pp. 39-42, Apr. 1994.
[18] Wang H. C., "MAT – A project to collect Mandarin speech data through telephone networks in Taiwan", *Computational Linguistics and Chinese language Processing*, pp. 73-90, vol. 2, no. 1, 1997.