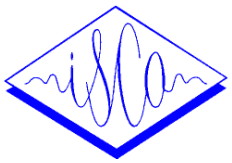

International Symposium on Chinese Spoken Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

COMPACT SPEECH FEATURES BASED ON WAVELET TRANSFORM AND PCA WITH APPLICATION TO SPEAKER IDENTIFICATION

Ching-Tang HSIEH

Department of Electrical Engineering,
Tamkang University, Taipei
hsieh@ee.tku.edu.tw

Wan-Chen CHEN

Department of Electrical Engineering,
Tamkang University, Taipei
steven@mail.sjsmit.edu.tw

Eugene LAI
Department of Electrical Engineering,
Tamkang University, Taipei
elai@ee.tku.edu.tw

You-Chuang WANG
Department of Electrical Engineering,
Tamkang University, Taipei
g6350015@tkgis.tku.edu.tw

ABSTRACT

The main goal of this paper is to find some effective methods to improve the performance of speaker identification system. In speaker identification, we use wavelet transform to decompose the speech signals into several frequency bands and then use cepstral coefficients to capture the individualities of vocal track within the interested bands based on the acoustic characteristic of human ear. In addition, an adaptive wavelet-based filtering mechanism is applied to eliminate the small variation of wavelet coefficients caused by noise. In order to effectively utilize all these multi-band speech features, we propose a modified vector quantization method called multi-layer eigen-codebook vector quantization (MLECVQ) as the identifier. This model uses the multi-layer concept to eliminate the interference between the multi-band coefficients and then uses the principal component analysis (PCA) method to evaluate the codebooks for capturing more details of phoneme character. Experimental results show that the proposed method is better than the GMM+MFCC model on computational cost and recognition performance under clean and noisy speech data evaluations.

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.

There are many researchers focusing their subjects on speaker recognition. Linear predictive cepstral coefficients (LPCC) are used because of its simplicity and effectiveness in speaker/speech recognition [1], [2]. Another widely used feature parameters, mel frequency cepstral coefficients (MFCC) [3] were calculated based on the acoustic characteristic of human ears. Cepstral coefficients and their time differential functions are used as the feature spaces [4]. In that paper, Furui uses mean normalization to improve the identification performance by minimizing intersession variability. Phan et al. [5] use the technique of wavelet transform to divide the speech signals into four octaves by using the quadrature mirror filters (QMFs). The experiments show that the performance is badly bothered by noises because of the simplified extraction method.

In past literatures for recognition models, vector quantization (VQ)[6]-[8], neural network (NN), dynamic time warping (DTW), hidden Markov model (HMM)[12], and Gaussian mixture model (GMM) were used in speaker recognition task. He et al. [9] combine VQ and learning vector quantization (LVQ) [10] to form the group vector quantization (GVQ) that its performance is better than LVQ but it needs to be retrained after entering new samples. Phan et al. [5] use NN in speaker recognition task. Although NN technique is robust and has high performance, it needs to be retrained after entering new samples. DTW technique [11] is not suitable for text-independent speaker recognition. GMM [13]-[16] were widely used in speaker recognition and had satisfactory performance.

In this paper, we propose an effective and robust extraction method of speech features called multi-band linear predictive cepstral coefficients (MBLPCC) based on time-frequency multi-resolution analysis. In order to effectively utilize all these multi-band speech features, we propose a modified vector quantization method called multi-layer eigen-codebook vector quantization (MLECVQ) as the identifier. In our text-independent evaluations, the experimental results show that the proposed method has satisfactory computation cost and performance especially on the noisy environments.

This paper is organized as follows. The proposed extraction algorithm of speech features is described in Section 2. Section 3 gives the proposed speaker recognition model. Experimental results and several comparisons with the other existing methods are presented in Section 4. Concluding remarks are given in Section 5.

2. MULTI-BAND FEATURES BASED ON WAVELET TRANSFORM

The wavelet transform is applied to decompose the input signal into various uncorrelating 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. Consequently, an effective and robust speech feature can be constructed by using the proposed multi-band linear predictive cepstral coefficients (MBLPCC).

As we know, the LPCC parameters are bothered by noise interference. Hence, an adaptive threshold technique is applied in each approximation channel before the next decomposition process. According to the fact that the conspicuous peaks in the time domain have large components over many wavelet scales, superfluous variations die out swiftly with increasing scale. This allows characterization of the wavelet coefficients with respect to their amplitudes. The most significant coefficients at each scale, with amplitude above some thresholds, are defined as follows:

$$\theta_j = \sigma_j \cdot R_j \tag{1}$$

Where σ_j is the standard deviation of the wavelet coefficients within approximation channel at scale *j*, and R_j is an adjusting multiplicative factor used to restrict the threshold at a certain extent. Experimental results show that using this mechanism not only reduces the influence of the noise interference but also improves the recognition rate. Finally, the schematic flow of the proposed feature extraction method is shown in Fig. 1.

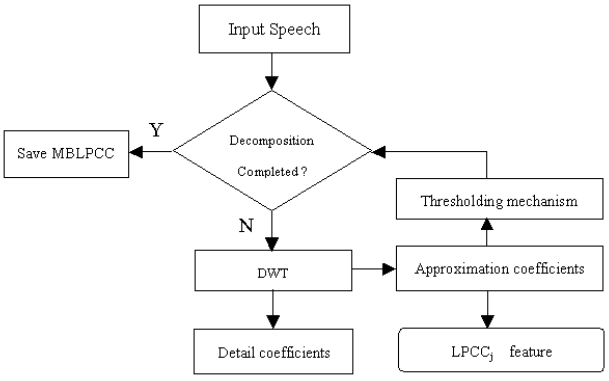

Fig. 1. Flow chart of the proposed MBLPCC feature extraction algorithm.

The recursive decomposition process lets us 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:

$$MBLPCC = \underset{j=1}{\overset{L}{Append}} (LPCC_j) \tag{2}$$

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. SPEAKER RECOGNITION MODELS

As described in Section 1, the GVQ and neural network (NN) need to be retrained after entering new training sample and DTW technique is not suitable for text-independent speaker recognition. The HMM has been widely used in speech recognition, but it is not suitable in speaker recognition for its complexity. On the contrary, GMM and VQ is widely used in speaker recognition application. Although the performance of GMM is better than VQ [13][9], VQ is suitable for our MBLPCC features because GMM is too complex to evaluate. Conventional VQ technique can't effectively describe the samples characteristic. Hence, we propose a modified vector quantization method called multi-layer eigen-codebook vector quantization (MLECVQ). This model uses the multi-layer concept to eliminate the interference between the multi-band coefficients and then uses the principal component analysis (PCA) method to evaluate the codebooks for capturing more details of phoneme character.

The PCA method [17]-[19] uses the statistical distribution of input samples to find the best projection bases. The advantages of PCA method are that the principal eigenvectors is 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 $\bar{x}$ and eigenvectors $\vec{\phi}_1, \vec{\phi}_2, ..., \vec{\phi}_m$ may effectively describe the characteristic of samples belonging to specific code 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:

$$p_i = (x_{new} - \bar{x})^T \cdot \vec{\phi}_i \qquad i = 1,2,...,m \tag{3}$$

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

$$\varepsilon = x_{new} - \bar{x} - \sum_{i=1}^{m} p_i \vec{\phi}_i \tag{4}$$

Where ε is distance vector. The norm $\|\varepsilon\|$ represents distance value between x_{new} and the specific code word. The new sample x_{new} is belong to the code word that is most closing to x_{new}. The training procedure of the proposed eigen-codebook vector quantization (ECVQ) is described as follows:

Step 1: set up the number of the projection bases *m*, number of code words C_{no} and 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 $\vec{\phi}_1, \vec{\phi}_2, ..., \vec{\phi}_m$ for each code word.

Step 4: use (4) to evaluate the distance vector ε and classify the training samples by $\|\varepsilon\|$.

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 $\vec{\phi}_1, \vec{\phi}_2, ..., \vec{\phi}_m$ for each code word.

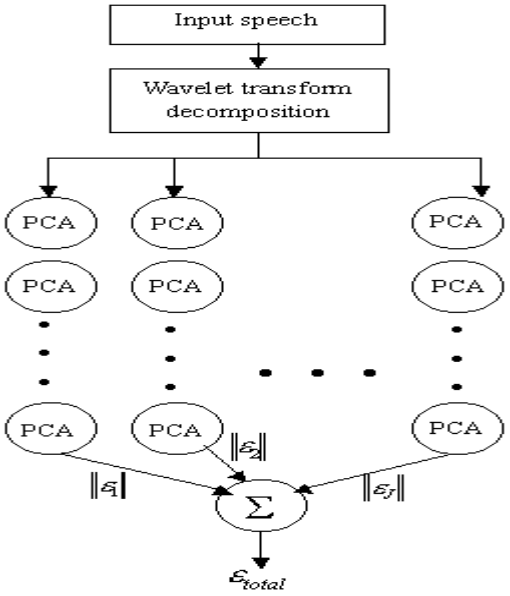

Fig. 2. The structure of the MLECVQ model

Because the proposed MBLPCC is extracted from the wavelet coefficients of multi-band speech signals, in order to eliminate interference between the multi-band coefficients, we evaluate the individual codebook for each band . The total distance ε_{total} between x_{new} and the specific model is sum of ε in each band, as shown in Fig. 2.

4. EXPERIMENTAL RESULTS

This section presents the application of MLECVQ to text-independent speaker identification. The experiment compares the performance of proposed MLECVQ with conventional VQ and GMM models.

4.1 Database Description and Parameters Setting

The proposed method is evaluated on KING [20] and MAT-400 [21] 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. In our experiments, in order to eliminate the silence section from an utterance, a simple segmentation based on signal energy of each speech frame is used.

With all experiments in this paper, the frame size of the analysis utterance is 256 samples with 128 samples overlapping. Furthermore, 16 orders of LPCC for each decomposition process were used. While computing LPCC, the mean normalization is used to compensate the channel effect.

4.2 Comparison with Other Existing Models

In this experiment, the performance of MLECVQ 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 MLECVQ model, the decomposition level is 4, R_j=0, each layer has 64 code words that is 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 MLECVQ model is better than other models, and maintains its robustness at lower SNR.

In other experiment with two sentence utterances of 20 speaker, the performance and computational cost of MLECVQ is compared to conventional VQ and GMM model. The results are depicted in Table 2. In Table 2, X / Y 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 MLECVQ is better than GMM+MFCC model.

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

Method	KING							MAT-400
	Clean	20 dB	15 dB	10 dB	5 dB	Tel-30.8dB	Tel-18.2dB	Tel
VQ	82.62%	67.73%	53.45%	35.46%	18.99%	71.87%	61.5%	95.62%
GMM	92.61%	85.88%	73.28%	52.61%	34.96%	75.38%	58.14%	98.18%
MLECVQ	96.47%	90.05%	82.65%	68.7%	47.86%	81.32%	69.38%	99.04%

Table 2. Identification rates and computational time of the VQ, GMM, and MLECVQ for clean speech data and different number of code words

Number of Code words	VQ		GMM		MLECVQ	
	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. CONCLUSIONS

In this paper, based on the time-frequency analysis of wavelet transform, we propose an effective and robust method for extracting the speech features (called MBLPCC). In order to effectively utilize all multi-band speech features, we propose a modified vector quantization method called multi-layer eigen-codebook vector quantization (MLECVQ) as the identifier. This model uses the multi-layer concept to eliminate the interference between the multi-band coefficients and then uses the principal component analysis (PCA) technique to evaluate the codebooks. Experimental results show that the proposed method is more effective and robust than GMM+MFCC method.

6. REFERENCES

[1] Atal B.: "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Acoust. Soc. Amer. J.,* 55, pp. 1304-1312, 1974.

[2] White G. M., and Neely R. B.: "Speech recognition experiments with linear prediction, bandpass filtering, and dynamic Programming," *IEEE Trans. Acoustics, Speech, Signal Proc.,* 24, (2), pp.183-188, 1976.

[3] Vergin R., O'Shaughnessy D., and Farhat A.: "Generalized mel frequency cepstral coefficients for large-vocabulary speaker-independent continuous-speech recognition," *IEEE Trans. Speech and Audio Processing,* 7, (5), pp. 525-532, 1999.

[4] Furui S.: "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. Acoust., Speech, Signal Processing,* 29, pp. 254-272, 1981.

[5] Phan F., Evangella M. T., and Sideman S.: "Speaker identification using neural networks and wavelets," *IEEE Engineering in Medicine and Biology Magazine,* 191, pp. 92-101, 2000.

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

[7] Buck J. T., Burton D. K., and Shore J. E., "Text-dependent speaker recognition using vector quantization," *Proceedings of ICASSP*-85, 1:391-394, 1985.

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

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.