Robust Speaker Identification System Based on Wavelet Transform and Gaussian Mixture Model

Wan-Chen Chen
Dept. of Electronic Engineering
St. John's & St. Mary's Institute of Technology
Taipei
steven@mail.sjsmit.edu.tw

Ching-Tang Hsieh
Dept. of Electrical Engineering
Tamkang University
Taipei
hsieh@ee.tku.edu.tw

Eugene Lai
Dept. of Electrical Engineering
Tamkang University
Taipei
elai@ee.tku.edu.tw

Abstract

This paper presents some effective method for improving the performance of a speaker identification system. Based on the multiresolution property of the wavelet transform, the input speech signal is decomposed into various frequency bands in order not to spread noise distortions over the entire feature space. The linear predictive cepstral coefficients (LPCC) of each band are calculated. Furthermore, the cepstral mean normalization technique is applied to all computed features. In order to effectively utilize these multiband speech features, we use feature recombination and likelihood recombination methods to evaluate the task of the text-independent speaker identification. The feature recombination scheme combines the cepstral coefficients of each band to form a single feature vector used to train the Gaussian mixture model (GMM). The likelihood recombination scheme combines the likelihood scores of independent GMM for each band. Experimental results show that the proposed both methods outperform the GMM model using full-band LPCC and MFCC features in both clean and noisy environments.

Keywords: speaker identification, wavelet transform, linear predictive cepstral coefficients (LPCC), mel-frequency cepstral coefficients (MFCC), Gaussian mixture model (GMM).

1 Introduction

In general, speaker recognition can be divided into two parts: speaker verification and speaker identification. Speaker verification refers to whether or not the speech samples belong to some specific speaker. However, in speaker identification, the goal is to determine which one of a group of known voices best matches the input voice sample. Furthermore, in both of the tasks the speech can be either text-dependent or text-independent. Text-dependent means that the text used in the training system must be the same as that used in the test system, while text-independent means that there is no limitation on the text used in the test system. Certainly, how to extract and model the speaker-dependent characteristics of the speech signal is the key point which seriously affects the performance of the speaker recognition system.

Much research has been done on the task of speech feature extraction. The linear predictive cepstral coefficients (LPCC) (Atal, 1974; White and Neely, 1976) were used because of their simplicity and effectiveness in speaker/speech recognition. Other widely used feature parameters, mel-frequency cepstral coefficients (MFCC) (Vergin et al., 1999), were calculated by using a filter-bank approach in which the set of filters had equal bandwidths with respect to the mel-scale frequencies. This was based on the fact that the human perception of frequency contents of sounds did not follow a linear scale. The above two most commonly used feature extraction techniques do not provide an invariant parameterization for speech—the representation of the speech signal tends to change due to various noise conditions. The performance of

the speaker identification systems may severely degrade in the case of a mismatch between training and test environments. Various types of speech enhancement and noise elimination techniques have been applied to feature extraction. Typically, the nonlinear spectral subtraction (Lockwood and Boudy, 1992) had provided only minor performance gains after extensive parameter optimization. Furui (1981) used the cepstral mean normalization technique to eliminate the channel bias by subtracting off the global average cepstral vector from each cepstral vector. Other way to minimize the channel filter effects is to use the cepstrum difference coefficients (Soong and Rosenberg, 1988). Cepstral coefficients and their time difference functions were used as the features in order to capture the dynamic information and eliminate the time-invariant spectral information generally attributed to the interposed communication channel.

Conventionally, feature extraction is carried out by computing acoustic feature vectors over the full band of the spectral representation of speech. The major drawback of this approach is that even a partial band-limited noise corruption affects all feature vector components. The multiband approach (Hermansky et al., 1996) coped with this problem by performing the acoustic feature analysis independently for a set of frequency subbands. Since the resulting coefficients were computed independently, a band-limited noise signal did not spread over the entire feature space. The major drawback of a pure subband-based approach may be that information on the correlation between various subbands is lost. Therefore, an approach of combining the information from full-band and subbands at the recognition stage produced recognition improvements (Mirghafori and Morgan, 1998). It is not trivial to decide at which temporal level the combination of subband features should be carried out. In the multiband approach (Bourlard and Dupont, 1996; Hermansky and Malayath, 1998), different classifiers for each band were used and the likelihood recombination was done at hidden Markov model (HMM) state, phone, or word level. In other approach, Okawa et al. (1998) proposed the combination of the individual features of each subband into a single feature vector prior to decoding. In our previous study (Hsieh and Wang, 2001), we proposed a multiband linear predictive cepstral coefficients (MBLPCC) method in which the LPCC features from various subbands and full-band were combined to form a single feature vector. This feature extraction method was evaluated for speaker identification system using vector quantization (VQ) as the identifier. The experimental results showed that this method was more effective and robust than the full-band LPCC and MFCC features, particularly in noisy environments.

In past studies for recognition models, dynamic time warping (DTW) (Furui, 1981), HMM (Poritz, 1982; Tishby, 1991), and Gaussian mixture model (GMM) (Miyajima et al., 2001; Alamo et al., 1996; Pellom and Hansen, 1998) were used in speaker recognition. The DTW technique is effective in the text-dependent speaker recognition, but it is not suitable for the text-independent speaker recognition. The HMM is widely used in speech recognition and it is also commonly used in the text-dependent speaker verification. The GMM (Reynold and Rose, 1995) provides a probabilistic model of the underlying sounds of a person's voice. It is computationally more efficient than HMM and has been widely used in text-independent speaker recognition.

In this study, the MBLPCC features proposed previously are used as the front end of the speaker identification system. Then the cepstral mean normalization is applied to these multiband speech features to provide similar parameter statistics in all acoustic environments. In order to effectively utilize all multiband speech features, we use the features recombination and the likelihood recombination methods in GMM speaker models to evaluate the task of the text-independent speaker identification. The experimental results show that the proposed methods outperform the GMM using full-band LPCC and MFCC features.

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 models. Experimental results and comparisons with conventional full-band GMM are presented in Section 4. Concluding remarks are given in Section 5.

2 Multiresolution Features Based on Wavelet Transform

On the basis of the time-frequency multiresolution analysis, the effective and robust MBLPCC is used as the front end of the speaker identification system. First, the LPCC is extracted from the full-

band input signal. Then the wavelet transform is applied to decompose the input signal into two frequency subbands: a lower frequency approximated subband and a higher frequency detailed subband. For capturing the characteristics of an individual speaker, the LPCC of the lower frequency subband is calculated. The main reason for using the LPCC parameters is their good representation on the envelope of speech spectrum of vowels and their simplicity. Based on this mechanism, one can easily extract the multiresolution features from all lower frequency subband signals simply iteratively applying the wavelet transform to decompose the lower frequency subband signals, as depicted in Figure 1. In Figure 1, the wavelet transform can be realized by using a pair of the finite impulse response (FIR) filters H and G, which are low-pass and high-pass filters, respectively, and the down-sampling operation (↓2). The down-sampling operation is used to discard the odd-numbered samples in a sample sequence after filtering.

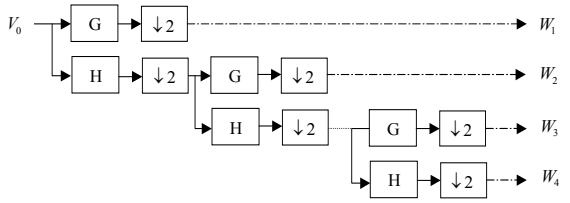

Figure 1. Two-band analysis tree for a discrete wavelet transform.

The schematic flow of the proposed feature extraction method is shown in Figure 2. In Figure 2, after the full-band LPCC is extracted from the input speech signal, the discrete wavelet transform (DWT) is applied to decompose the input signal into a lower frequency subband and the subband LPCC is extracted from this subband. The recursive decomposition process lets us easily acquire the multiband features of the speech signal. According to the concept of the proposed method, the number of MBLPCC coefficients depends on the level of decomposition process.

Finally, the cepstral mean normalization is applied to normalize the feature vectors, so that their short-term means are normalized to zero as follows:

$$\hat{x}_k(t) = x_k(t) - \mu_k, \quad (1)$$

where $x_k(t)$ is the kth component of feature vector at time (frame) t, and μ_k is the mean of the kth component of feature vectors of a specific speaker's utterance.

In this paper, the orthonormal basis of the DWT is based on the quadrature mirror filters (QMF) introduced by Daubechies (1988).

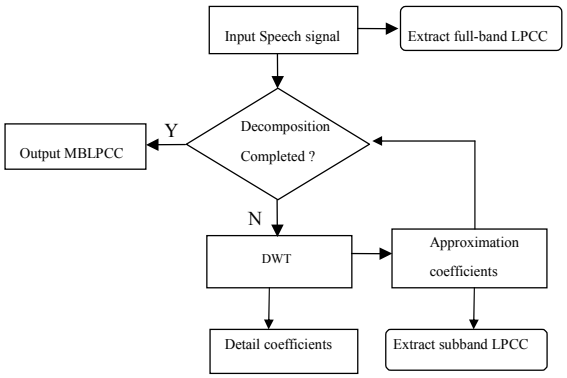

Figure 2. Features extraction algorithm of MBLPCC

3 Multiband Speaker Recognition Models

As described in Section 1, the GMM is widely used in the text-independent speaker recognition and shows good performances. Here, we use GMM as the classifier. Our initial strategy of multiband speaker recognition is based on straightforward recombination of the cepstral coefficients from each subband (including full-band) to form a single feature vector, which is used to train GMM. We name this identifier model as the feature combination Gaussian mixture models (FCGMM). The structure of the FCGMM is shown in Figure 3. The advantages of this approach are that: (1) it is possible to model the correlation between feature vectors of each band; (2) acoustic modeling becomes simpler. Other approach combines the likelihood scores of independent GMM for each band, as illustrated in Figure 4. We name this identifier model as the multi-layer Gaussian mixture models (MLGMM).

For speaker identification, a group of S speakers is represented by MLGMM's λ_1, λ_2,…, λ_S. A given speech utterance X is decomposed into L subbands. Let X_i and λ_{ki} be the feature vector and the associated GMM for band i, respectively. After the logarithmic likelihood $\log P(X_i|\lambda_{ki})$ of band i for a specific speaker k is evaluated, the combined logarithmic likelihood $\log P(X|\lambda_k)$ for the MLGMM of a specific speaker k is determined as the sum of

logarithmic likelihood $\log P(X_i|\lambda_{ki})$ for all bands as follows:

$$\log P(X \mid \lambda_k) = \sum_{i=0}^{L} \log P(X_i \mid \lambda_{ki}), \qquad (2)$$

where L is the number of subbands. When $i = 0$, the functions of the MLGMM and the conventional full-band GMM are identical. For a given speech utterance X, X is classified to the speaker $\hat{S}$ who has the maximum logarithmic probability $\log P(X \mid \lambda_{\hat{S}})$:

$$\hat{S} = \arg \max_{1 \le k \le S} \log P(X \mid \lambda_k). \qquad (3)$$

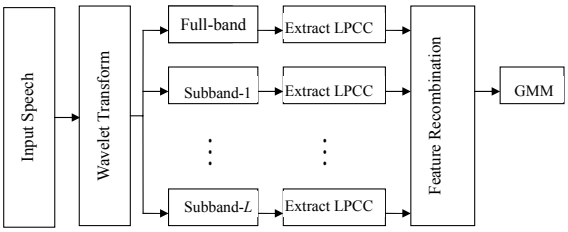

Figure 3. Structure of FCGMM

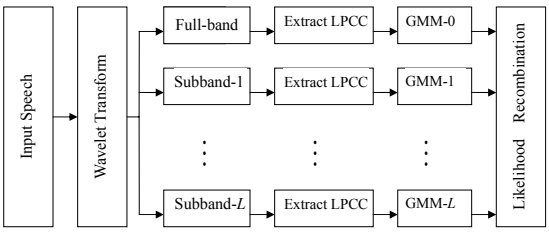

Figure 4. Structure of MLGMM

4 Experimental Results

This section presents the evaluations of the FCGMM and MLGMM for the text-independent speaker identification. The first experiment studies the effect of decomposition level. The next experiment compares the performance of the FCGMM and MLGMM with that of the conventional GMM model using the full-band LPCC and MFCC features.

4.1 Database Description and Parameters Setting

The proposed method is evaluated using the KING speech database (Godfrey et al., 1994) for the text-independent speaker identification. The KING database is a collection of conversational speech from 51 male speakers. For each speaker there are 10 sections of conversational speech recorded at different time. The waveform file of each section consists of about 30 seconds of actual speech. The speech from a section is recorded from a microphone locally and was transmitted over a long distance telephone link, providing a high-quality (clean) version and a telephone quality version of the speech. The speech signals are recorded at 8 kHz and 16 bits per sample. In our experiments, the noisy speech is generated by adding Gaussian noise to the clean version speech at the desired SNR. In order to eliminate the silence segments from an utterance, a simple segmentation based on signal energy of each speech frame is used. The experiments are evaluated using five sections of 20 speakers. For each speaker, 90 seconds of speech cut from three clean version sections are used as the training utterances. The other two sections are divided into nonoverlapping segments of 2 seconds and are used as the test utterances.

For all experiments in this paper, each frame of the analyzed utterance has 256 samples with 128 samples overlapping. Furthermore, 20 orders of LPCC for each frequency band are calculated and the first order coefficient is discarded. For our multiband approach, we use 2, 3 and 4 bands as follows:

2 bands: (0-4000), (0-2000) Hz.
3 bands: (0-4000), (0-2000), (0-1000) Hz.
4 bands: (0-4000), (0-2000), (0-1000), (0-500) Hz.

4.2 Effects of Decomposition Levels

As described in Section 2, the number of subbands depends on the decomposition levels of wavelet transform. This experiment is to evaluate the effect of number of bands used for the FCGMM and MLGMM models with 50 mixtures in both clean and noisy environments. The experimental results are shown in Table 1.

We can see that 3-band FCGMM model has better performance in low SNR conditions (for example, 15 dB, 10 dB or 5 dB), but has poorer performance in clean and 20 dB SNR conditions in

comparing with 2-band FCGMM model. The best identification rate of the MLGMM model can be achieved in both clean and noisy environments when the number of bands is set to be three. It is shown that when the number of bands greater than three for both models, it not only increases the computation time but also decreases the identification rate. In this case, the highest layer signals locate at very low frequency subband and the number of samples within highest layer subband is so small that it cannot accurately estimate the spectral characteristics of speech. Consequently, the poor result in highest layer subband will downgrade the system performance.

Table 1. Effects of number of bands on the identification rate (%) for FCGMM and MLGMM models in both clean and noisy environments.

Model \ SNR(dB)		clean	20	15	10	5
FC-GMM	2 bands	93.45	85.55	72.10	50.25	30.76
	3 bands	91.09	83.87	76.64	60.50	46.22
	4 bands	88.07	81.18	74.29	63.03	43.36
ML-GMM	2 bands	93.28	86.39	76.47	53.78	28.24
	3 bands	94.96	92.10	86.89	68.07	43.53
	4 bands	94.12	89.41	84.87	71.76	43.19

4.3 Comparison with Conventional GMM Models

In this experiment, the performance of the FCGMM and MLGMM models are compared with that of the conventional GMM model using full-band 20 orders LPCC and MFCC features under Gaussian noise corruption. For all models, the number of mixtures is set to be 50.

Here, the parameters of the FCGMM and MLGMM are the same as for Section 4.2 except that the number of bands is set to be three. Experimental results in Table 2 show that the performance of both GMM models using full-band LPCC and MFCC features is seriously degraded by Gaussian noise corruption. However, the MLGMM model gives the best performance among all models in both clean and noisy environments, and maintains its robustness in low SNR conditions. The GMM model using full-band MFCC features has better performance in clean and 20 dB SNR conditions, but has poorer performance in lower SNR conditions in comparing with 3-band FCGMM. The GMM using full-band LPCC features has poorest performance among all models. Finally, it can be concluded that the MLGMM model is effective to represent the characteristics of individual speaker and is robust to the additive Gaussian noise conditions.

Table 2. Identification rates (%) for GMM using full-band LPCC and MFCC features, FCGMM, and MLGMM under white noise corruption.

Model \ SNR(dB)	Clean	20	15	10	5
GMM using full-band LPCC	88.40	77.65	61.68	35.63	19.50
GMM using full-band MFCC	92.61	85.88	73.11	51.60	32.77
3-band FCGMM	91.09	83.87	76.64	60.50	46.22
3-band MLGMM	94.96	92.10	86.89	68.07	43.53

5 Conclusion

In this study, on the basis of the time-frequency analysis of the wavelet transform, the effective and robust MBLPCC features proposed in the previous works are used as the front end of the speaker identification system. In order to effectively utilize these multiband speech features, we examine two different approaches. The FCGMM combines the cepstral coefficients from each band to form a single feature vector used to train the GMM. The MLGMM recombines the likelihood scores of independent GMM for each band. Finally, the proposed methods are evaluated using the KING speech database for the text-independent speaker identification. Experimental results show that both multiband schemes are more effective and robust than the GMM model using full-band LPCC and MFCC features. In addition, the identification rate of the MLGMM is more effective than that of the FCGMM.

Acknowledgements

This research was sponsored by the National Science Council (Taipei), under contract number NSC 92-2213-E032-026.

References

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

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.

Bourlard, H. and Dupont, S., "A new ASR approach based on independent processing and recombination of partial frequency bands," *Proc. Int. Conf. Spoken Language Processing*, pp. 426–429, 1996.

Daubechies, I., "Orthonormal bases of compactly supported wavelets," *Commun. Pure Appl. Math.,* 41, pp. 909-996, 1988.

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

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.

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.

Hermansky, H. and Malayath, N., "Spectral basis functions from discriminant analysis," *Proc. Int. Conf. Spoken Language Processing*, pp. 1379–1382, 1998.

Hermansky, H., Tibrewala, S., and Pavel, M., "Toward ASR on partially corrupted speech," *Proc. Int. Conf. Spoken Language Processing*, pp. 462–465, 1996.

Hsieh, C. T. and Wang, Y. C., "A robust speaker identification system based on wavelet transform," *IEICE Trans. Inf. & Syst.*, vol. E84-D, no. 7, pp.839-846, July 2001.

Lockwood, P. and Boudy, J., "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," *Speech Commun.*, vol. 11, no. 2-3, pp. 215–228, 1992.

Mirghafori, N. and Morgan, N., "Combining connectionist multi-band and full-band probability streams for speech recognition of natural numbers," *Proc. Int. Conf. Spoken Language Processing*, vol. 3, pp. 743–747, 1998.

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.

Okawa, S., Bocchieri, E., and Potamianos, A., "Multi-band speech recognition in noisy environments," Proc. Int. Conf. Acoustics, Speech, Signal Processing, pp. 641–644, 1998.

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.

Poritz, A., "Linear predictive hidden markov models and the speech signal," *Proc. ICASSP-82*, 2:1291-1294, 1982.

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.

Soong, F. K. and Rosenberg, A. E., "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Trans. Acoust., Speech, Signal Processing,* 36, pp. 871-879, 1988.

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.

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.

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.