Multiband Approach to Robust Text-Independent Speaker Identification

Wan-Chen Chen[*+], Ching-Tang Hsieh[*], and Eugene Lai[*]

Abstract

This paper presents an 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. To capture the characteristics of the vocal tract, the linear predictive cepstral coefficients (LPCCs) of each band are calculated. Furthermore, the cepstral mean normalization technique is applied to all computed features in order to provide similar parameter statistics in all acoustic environments. In order to effectively utilize these multiband speech features, we use feature recombination and likelihood recombination methods to evaluate the task of 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 the independent GMM for each band. Experimental results show that both proposed methods achieve better performance than GMM using full-band LPCCs and mel-frequency cepstral coefficients (MFCCs) when the speaker identification is evaluated in the presence of clean and noisy environments.

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

1. Introduction

In general, speaker recognition can be divided into two parts: speaker verification and speaker

---

[*] Department of Electrical Engineering, Tamkang University, Taipei, Taiwan, Republic of China

[+] Department of Electronic Engineering, St. John's & St. Mary's Institute of Technology, Taipei, Taiwan, Republic of China

E-mail: steven@mail.sjsmit.edu.tw, hsieh@ee.tku.edu.tw, elai@ee.tku.edu.tw

identification. Speaker verification refers to the process of determining 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 tasks, the speech can be either text-dependent or text-independent. Text-dependent means that the text used in the test system must be the same as that used in the training system, while text-independent means that no limitation is placed on the text used in the test system. Certainly, the method used to extract and model the speaker-dependent characteristics of a speech signal seriously affects the performance of a speaker recognition system.

Many researches have been done on the feature extraction of speech. The linear predictive cepstral coefficients (LPCCs) were used because of their simplicity and effectiveness in speaker/speech recognition [Atal 1974, White and Neely 1976]. Other widely used feature parameters, namely, the mel-frequency cepstral coefficients (MFCCs) [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 method is based on the fact that human perception of the frequency contents of sounds does not follow a linear scale. The above two most commonly used feature extraction techniques do not provide invariant parameterization of speech; the representation of the speech signal tends to change under various noise conditions. The performance of these speaker identification systems may be severely degraded when a mismatch between the training and testing environments occurs. Various types of speech enhancement and noise elimination techniques have been applied to feature extraction. Typically, the nonlinear spectral subtraction algorithms [Lockwood and Boudy 1992] have provided only minor performance gains after extensive parameter optimization. Furui [1981] used the cepstral mean normalization (CMN) technique to eliminate channel bias by subtracting off the global average cepstral vector from each cepstral vector. Another way to minimize the channel filter effects is to use the time derivatives of cepstral coefficients [Soong and Rosenberg 1988]. 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.

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 partial band-limited noise corruption affects all the feature vector components. The multiband approach deals with this problem by performing acoustic feature analysis independently on a set of frequency subbands [Hermansky *et al.* 1996]. Since the resulting coefficients are computed independently, a band-limited noise signal does not spread over the entire feature space. In our previous works [Hsieh and Wang 2001, Hsieh *et al.* 2002,

2003], we proposed a multiband feature extraction method in which features from various subbands and the full band are combined to form a single feature vector. This feature extraction method was evaluated in a speaker identification system using vector quantization (VQ), group vector quantization, and the Gaussian mixture model (GMM) as identifiers. The experimental results showed that this multiband feature is more effective and robust than the full-band LPCC and MFCC features, particularly in noisy environments.

In past studies on recognition models, VQ [Soong *et al.* 1985, Buck *et al.* 1985, Furui 1991], dynamic time warping (DTW) [Furui 1981], the hidden Markov model (HMM) [Poritz 1982, Tishby 1991], and GMM [Reynolds and Rose 1995, Alamo *et al.* 1996, Pellom and Hansen 1998, Miyajima *et al.* 2001] were used to perform speaker recognition. The DTW technique is effective in text-dependent speaker recognition, but it is not suitable for text-independent speaker recognition. HMM is widely used in speech recognition, and it is also commonly used in text-dependent speaker verification. It has been shown that VQ is very effective for speaker recognition. Although the performance of VQ is not as good as that of GMM [Reynolds and Rose 1995], VQ is computationally more efficient than GMM. GMM [Reynolds 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 multiband linear predictive cepstral coefficients (MBLPCCs) proposed previously [Hsieh and Wang 2001, Hsieh *et al.* 2002, 2003] are used as the front end of the speaker identification system. Then, cepstral mean normalization is applied to these multiband speech features to provide similar parameter statistics in all acoustic environments. In order to effectively utilize these multiband speech features, we use feature recombination and likelihood recombination methods in the GMM recognition models to evaluate the task of text-independent speaker identification. The experimental results show that the proposed multiband methods outperform GMM using full-band LPCC and MFCC features.

This paper is organized as follows. The proposed algorithm for extracting speech features is described in section 2. Section 3 presents the multiband speaker recognition models. Experimental results and comparisons with the conventional full-band GMM are presented in section 4. Concluding remarks are made in section 5.

2. Multiband Features Based on Wavelet Transform

The recent interest in the multiband feature extraction approach has mainly been attributed to Allen's paper [Allen 1994], where it is argued that the human auditory system processes features from different subbands independently, and that the merging is done at some higher point of processing to produce a final decision. The advantages of using multiband processing

are multifold and have been described in earlier publications [Bourlard and Dupont 1996, Tibrewala and Hermansky 1997, Mirghafori and Morgan 1998]. The major drawback of a pure subband-based approach may be that information about the correlation among various subbands is lost. Therefore, we suggest that full-band features should not be ignored, but should be combined with subband features to maximize recognition accuracy. A similar approach that combines information from the full band and subbands at the recognition stage was found to improve recognition performance [Mirghafori and Morgan 1998]. It is not a trivial matter to decide at which temporal level the subband features should be combined. In the multiband approach [Bourlard and Dupont 1996, Tibrewala and Hermansky 1997], different classifiers for each band are used, and likelihood recombination is done at the HMM state, phone or word level. In another approach [Okawa *et al.* 1998, Hariharan *et al.* 2001], the individual features of each subband are combined into a single feature vector prior to decoding. In our approach, the full band and subband features are also used in the recognition model.

Based on time-frequency multiresolution analysis, the effective and robust MBLPCC features are used as the front end of the speaker identification system. First, the LPCCs are 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 subband and a higher frequency subband. To capture the characteristics of an individual speaker, the LPCCs of the lower frequency subband are calculated. There are two main reasons for using the LPCC parameters: their good representation of the envelope of the speech spectrum of vowels, and their simplicity. Based on this mechanism, we can easily extract the multiresolution features from all lower frequency subband signals simply by iteratively applying the wavelet transform to decompose the lower frequency subband signals, as depicted in Figure 1. As shown in Figure 1, the wavelet transform can be realized by using a pair of finite impulse response (FIR) filters, h and g, which are low-pass and high-pass filters, respectively, and by performing the down-sampling operation ($\downarrow 2$). The down-sampling operation is used to discard the odd-numbered samples in a sample sequence after filtering is performed.
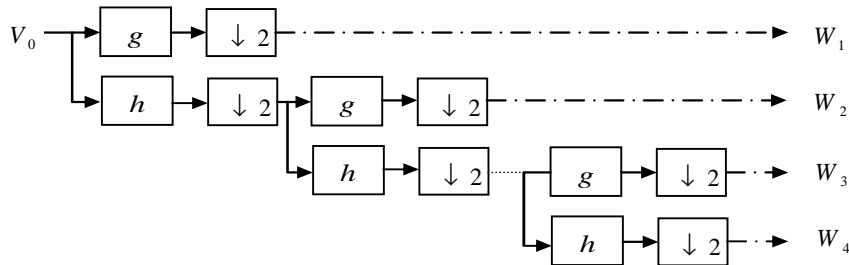

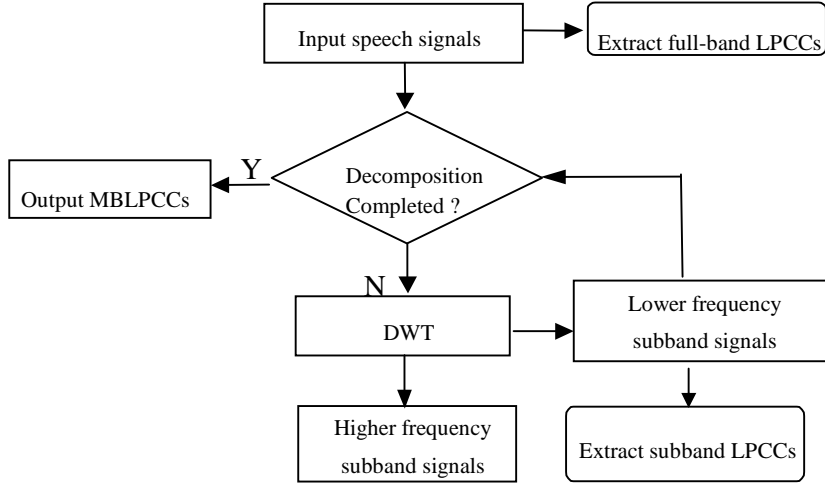**Figure 1. Two-band analysis tree for a discrete wavelet transform**

Figure 2. Features extraction algorithm of MBLPCCs

The schematic flow of the proposed feature extraction method is shown in Figure 2. After the full-band LPCCs are 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 LPCCs are extracted from this lower frequency subband. The recursive decomposition process enables us to easily acquire the multiband features of the speech signal. Based on the concept of the proposed method, the number of MBLPCCs depends on the level of the decomposition process. If speech signals bandlimited from 0 to 4000 Hz are decomposed into two subbands, then three bands signals, (0-4000), (0-2000), and (0-1000) Hz, will be generated. Since the spectra of the three bands will overlap in the lower frequency region, the proposed multiband feature extraction method focuses on the spectrum of the speech signal in the low frequency region similar to extracting MFCC features.

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

$$\hat{X}_k(t) = X_k(t) - m_k , \tag{1}$$

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

In this paper, the orthonormal basis of DWT is based on the 16 coefficients of the quadrature mirror filters (QMF) introduced by Daubechies [1988] (see the Appendix).

3. Multiband Speaker Recognition Models

As explained in section 1, GMM is widely used to perform text-independent speaker recognition and achieves good performance. Here, we use GMM as the classifier. Our initial strategy for multiband speaker recognition is based on straightforward recombination of the cepstral coefficients from each subband (including the full band) to form a single feature vector, which is used to train GMM. We call this identifier model the feature combination Gaussian mixture model (FCGMM). The structure of FCGMM is shown in Figure 3. First, the input signal is decomposed into L subbands. In the "extract LPCC" block, the LPCC features extracted from each band (including the full band) are further normalized to zero mean by using the cepstral mean normalization technique. Finally, the LPCCs from each subband (including the full band) are recombined to form a single feature vector that is used to train GMM. The advantages of this approach are that: (1) it is possible to model the correlation among the feature vectors of each band; (2) acoustic modeling is simpler.
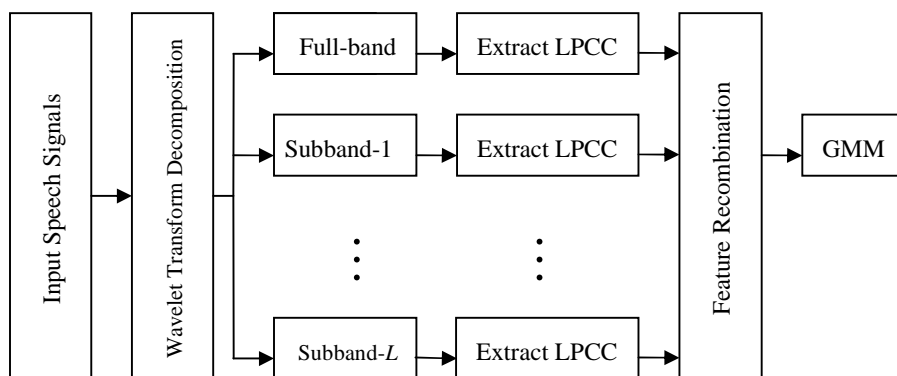

Figure 3. Structure of FCGMM

Our next approach combines the likelihood scores of the independent GMM for each band, as illustrated in Figure 4. We call this identifier model the likelihood combination Gaussian mixture model (LCGMM). First, the input signal is decomposed into L subbands. Then the LPCC features extracted from each band are further normalized to zero mean by using the cepstral mean normalization technique. Finally, different GMM classifiers are applied independently to each band, and the likelihood scores of all the GMM classifiers are combined to obtain the global likelihood scores and a global decision.
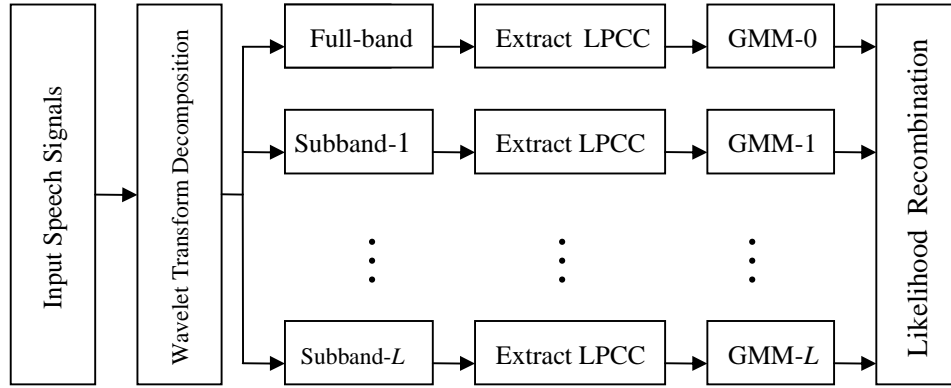
Figure 4. Structure of LCGMM

For speaker identification, a group of *S* speakers is represented by LCGMMs, $\lambda_1, \lambda_2, \ldots, \lambda_S$. 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 log-likelihood $\log P(X_i/\lambda_{ki})$ of band *i* for a specific speaker *k* is evaluated, the combined log-likelihood $\log P(X/\lambda_k)$ for the LCGMM of a specific speaker *k* is determined as the sum of the log-likelihood $\log P(X_i/\lambda_{ki})$ for all bands as follows:

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

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

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

4. Experimental Results

This section presents experiments conducted to evaluate application of FCGMM and LCGMM to text-independent speaker identification. The first experiment studied the effect of the decomposition level. The next experiment compared the performance of FCGMM and LCGMM with that of the conventional GMM using full-band LPCC and MFCC features.

4.1 Database Description and Parameter Setting

The proposed multiband approaches were evaluated using the KING speech database [Godfrey *et al.* 1994] for 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 that were recorded at different times. Each section consists of about 30 seconds of actual speech. The speech from a section was recorded locally using a microphone and was transmitted over a long distance telephone link, thus providing a high-quality (clean) version and a telephone quality version of the speech. The speech signals were recorded at 8 kHz and 16 bits per sample. In our experiments, noisy speech was generated by adding Gaussian noise to the clean version speech at the desired SNR. In order to eliminate silence segments from an utterance, simple segmentation based on the signal energy of each speech frame was performed. All the experiments were performed using five sections of speech from 20 speakers. For each speaker, 90 seconds of speech cut from three clean version sections provided the training utterances. The other two sections were divided into nonoverlapping segments 2 seconds in length and provided the testing utterances.

In both experiments conducted in this study, each frame of an analyzed utterance had 256 samples with 128 overlapping samples. Furthermore, 20 orders of LPCCs for each frequency band were calculated, and the first order coefficient was discarded. For our multiband approach, we used 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 Effect of the Decomposition Level

As explained in section 2, the number of subbands depends on the decomposition level of the wavelet transform. The first experiment evaluated the effect of the number of bands used in the FCGMM and LCGMM recognition models with 50 mixtures in both clean and noisy environments. The experimental results are shown in Table 1. One could see that the 3-band FCGMM achieved better performance under low SNR conditions (for example, 15 dB, 10 dB and 5 dB), but poorer performance under clean and 20 dB SNR conditions, compared with the 2-band FCGMM. Since the 2-band FCGMM used (0-4000) and (0-2000)Hz features, and the 3-band FCGMM used (0-4000), (0-2000) and (0-1000)Hz features, the feature derived from the lower frequency region (below 1kHz) was more robust than the feature derived from the higher frequency region under low SNR conditions.

The best identification rate of LCGMM could be achieved in both clean and noisy

environments when the number of bands was set to be three. Since the features were extracted from (0-4000), (0-2000) and (0-1000) Hz subbands and the spectra of the subbands overlapped in the lower frequency region (below 1kHz), the success achieved using the MBLPCC features could be attributed to the emphasis on the spectrum of the signal in the low-frequency region.

It was found that increasing the number of bands to more than three for both models not only increased the computation time but also decreased the identification rate. In this case, the signals of the lowest frequency subband were located in the very low frequency region, which put too much emphasis on the lower frequency spectrum of speech. In addition, the number of samples within the lowest frequency subband was so small that the spectral characteristics of speech could not be estimated accurately. Consequently, the poor result in the lowest frequency subband degraded the system performance.

Table 1. Effect of number of bands on the identification rates for FCGMM and LCGMM recognition models in both clean and noisy environments.

SNR / Model		clean	20 dB	15 dB	10 dB	5 dB
FCGMM	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%
LCGMM	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 LCGMM recognition models was compared with that of the conventional GMM using full-band LPCC and MFCC features under Gaussian noise corruption. For all three models, the number of mixtures was set to be 50.

Here, the parameters of FCGMM and LCGMM were the same as those discussed in section 4.2 except that the number of bands was set to be three. The experimental results

shown in Table 2 indicate that the performance of both GMM recognition models using full-band LPCC and MFCC features was seriously degraded by Gaussian noise corruption. On the other hand, LCGMM achieved the best performance among all the models in both clean and noisy environments, and maintained robustness under low SNR conditions. GMM using full-band MFCC features achieved better performance under clean and 20 dB SNR conditions, but poorer performance under lower SNR conditions, compared with the 3-band FCGMM. GMM using full-band LPCC features achieved the poorest performance among all the models. Based on these results, it can be concluded that LCGMM is effective in representing the characteristics of individual speakers and is robust under additive Gaussian noise conditions.

Table 2. Identification rates for GMM using full-band LPCC and MFCC features, FCGMM, and LCGMM with white noise corruption.

SNR Model	Clean	20 dB	15 dB	10 dB	5 dB
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 LCGMM	94.96%	92.10%	86.89%	68.07%	43.53%

5. Conclusions

In this study, the effective and robust MBLPCC features were used as the front end of a speaker identification system. In order to effectively utilize these multiband speech features, we examined two different approaches. FCGMM combines the cepstral coefficients from each band to form a single feature vector that is used to train GMM. LCGMM recombines the likelihood scores of the independent GMM for each band. The proposed multiband approaches were evaluated using the KING speech database for text-independent speaker identification. Experimental results show that both multiband schemes are more effective and robust than the conventional GMM using full-band LPCC and MFCC features. In addition, LCGMM is more effective than FCGMM.

Acknowledgements

References

Alamo, C. M., F. J. C. Gil, C. T. Munilla, and L. H. Gomez, "Discriminative training of GMM for speaker identification," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 1 1996, pp. 89-92.

Allen, J. B., "How do humans process and recognize speech?," *IEEE Transactions on Speech and Audio Processing*, 2(4) 1994, pp. 567–577.

Atal, B., "Effectiveness of linear prediction characteristics of the speech wave for automatic speaker identification and verification," *Journal of Acoustical Society America*, 55 1974, pp. 1304-1312.

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

Buck, J. T., D. K. Burton, and J. E. Shore, "Text-dependent speaker recognition using vector quantization," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 10 1985, pp. 391-394.

Daubechies, I., "Orthonormal bases of compactly supported wavelets," *Communications on Pure and Applied Mathematics,* 41 1988, pp. 909-996.

Furui, S., "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 29(2) 1981, pp. 254-272.

Furui, S., "Comparison of speaker recognition methods using statistical features and dynamic features," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 29(3) 1981, pp. 342-350.

Furui, S., "Vector-quantization-based speech recognition and speaker recognition techniques," *Proceedings of Conference Record of the Twenty-Fifth Asilomar Conference on Signals, Systems and Computers,* 4-6 Nov., 2 1991, pp.954-958.

Godfrey, J., D. Graff, and A. Martin, "Public databases for speaker recognition and verification," *Proceedings of ESCA Workshop Automatic Speaker Recognition, Identification, Verification,* 1994, pp. 39-42.

Hariharan, R., I. Kiss, I. Viikki, "Noise robust speech parameterization using multiresolution feature extraction," *IEEE Transactions on Speech and Audio Processing*, 9(8) 2001, pp. 856-865.

Hermansky, H., S. Tibrewala, and M. Pavel, "Toward ASR on partially corrupted speech," *Proceedings of 4th International Conference on Spoken Language Processing*,1 1996, pp. 462–465.

Hsieh, C. T., and Y. C. Wang, "A robust speaker identification system based on wavelet transform," *IEICE Transactions on Information and Systems*, E84-D(7) 2001, pp.839-846.

Hsieh, C. T., E. Lai, and Y. C. Wang, "Robust Speech Features based on Wavelet Transform with application to speaker identification", *IEE Proceedings – Vision, Image and Signal Processing*, 149(2) 2002, pp.108-114.

Hsieh, C. T., E. Lai, and Y. C. Wang, "Robust speaker identification system based on wavelet transform and Gaussian mixture model," *Journal of Information Science and Engineering,* 19 2003, pp. 267-282.

Lockwood, P., and J. Boudy, "Experiments with a nonlinear spectral subtractor (NSS), hidden Markov models and the projection, for robust speech recognition in cars," *Speech Communication*, 11(2-3) 1992, pp. 215–228.

Mirghafori, N., and N. Morgan, "Combining connectionist multiband and full-band probability streams for speech recognition of natural numbers," *Proceedings of International Conference on Spoken Language Processing*, 3 1998, pp. 743–747.

Miyajima, C., Y. Hattori, K. Tokuda, T. Masuko, T. Kobayashi, and T. Kitamura, "Text-independent speaker identification using Gaussian mixture models based on multi-space probability distribution," *IEICE Transactions on Information and Systems*, E84-D(7) 2001, pp. 847-855.

Okawa, S., E. Bocchieri, and A. Potamianos, "Multi-band speech recognition in noisy environments," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2 1998, pp. 641–644.

Pellom, B. L., and J. H. L. Hansen, "An effective scoring algorithm for Gaussian mixture model based speaker identification," *IEEE Signal Processing Letters,* 5(11) 1998, pp. 281-284.

Poritz, A., "Linear predictive hidden Markov models and the speech signal," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 7 1982, pp. 1291-1294.

Reynolds D. A., and R. C. Rose, "Robust test-independent speaker identification using Gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, 3(1) 1995, pp. 72-83.

Soong, F. K., A. E. Rosenberg, L. R. Rabiner, and B. H. Juang, "A vector quantization approach to speaker recognition," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 10 1985, pp. 387-390.

Soong, F. K., and A. E. Rosenberg, "On the use of instantaneous and transitional spectral information in speaker recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* 36(6) 1988, pp. 871-879.

Tibrewala, S., and H. Hermansky, "Sub-band based recognition of noisy speech," *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing,* 2 1997, pp. 1255–11258.

Tishby, N. Z., "On the application of mixture AR hidden Markov models to text independent speaker recognition," *IEEE Transactions on Signal Processing*, 39(3) 1991, pp. 563-570.

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

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

Appendix

The low-pass QMF coefficients h_k used in this study are listed in Table 3. The coefficients of the high-pass filter g_k are calculated from h_k coefficients as follows:

$$g_k = (-1)^k h_{n-1-k} \quad k = 0,1,...,n ,$$
(4)

where *n* is the number of QMF coefficients.

Table 3. QMF coefficients h_k

h_0	0.766130	h_8	0.008685
h_1	0.433923	h_9	0.008201
h_2	-0.050202	h_{10}	-0.004354
h_3	-0.110037	h_{11}	-0.003882
h_4	0.032081	h_{12}	0.002187
h_5	0.042068	h_{13}	0.001882
h_6	-0.017176	h_{14}	-0.001104
h_7	-0.017982	h_{15}	-0.000927