Mel-frequency cepstral coefficients extraction based on wavelet transform for speaker identification in reverberant environments

Noha Korany

E. E. Dept., Faculty of engineering, Alexandria University, Egypt. E-Mail: nokorany@hotmail.com

Abstract

This paper aims to improve the performance of speaker identification in reverberant environments, as reverberation affects the perceived signals, and degrades the identification performance.

Wavelet decomposition is applied to the noisy reverberant speech so that clean speech frequency band is obtained, and wavelet-based Mel-frequency cepstral coefficients are extracted. Gaussian mixture model classifier is trained on a data set of clean speech. Then, it is used to determine the speaker identity.

Wavelet parameters are investigated to improve the identification rate in reverberant environments. Experimental evaluation shows that employing wavelet-based Mel-frequency cepstral coefficients by the classifier reduces the influence of reverberation on the extracted features, and increases significantly the performance of the identification process.

Introduction

Speaker recognition process aims to automatically establish the identity of an individual based on his voice.

Mel-frequency cepstral coefficients (MFCC) are widely used for speaker identification. Employing MFCC for the identification process yields to a good performance in clean environments. Noise affects the whole MFCC feature vector even in the presence of band-limited noise. However, the performance of the identifier degrades[1]. This implies the application of noise elimination techniques [2].

Wavelet transform is widely used for signal detection and de-noising. This paper suggests the extraction of a wavelet-based MFCC feature vector. Discrete Wavelet transform is employed for the decomposition of the speaker utterance. Then, MFCC feature vector is extracted from the lowest frequency sub-band. The extracted wavelet-based MFCC feature vector is more effective than the conventional MFCC.

The paper aims to evaluate the performance of the identification process using the proposed feature vector. It is organized as follows. The next section suggests the extraction of the wavelet-based Mel-frequency cepstral coefficients. Section 3 discusses the effect of reverberation on the acoustic quality of the room. Section 4 presents the speaker identification model. Section 5 describes the database followed by a discussion of the simulation and the results. The last section summarizes the main conclusions.

Extraction of Wavelet-based Mel-frequency cepstral coefficients

The first step in an automatic speech recognition system is feature extraction. MFCC has been widely used for automatic speech and speaker recognition, but they are noise sensitive coefficients. MFCC is extracted as follows [3]. First, the spectrum is warped to the Mel-scale as defined in equation (1), where f is the frequency in Hz. Then, the log-power spectrum is transformed to the cepstral domain using discrete cosine transform (DCT). The result of the conversion is the MFCC feature vector.

$$Mel(f) = 2595\, log(1 + f/700) \qquad (1)$$

Wavelet transform is an effective tool for noise reduction. It is used to decompose a signal into shifted and scaled versions of a particular wavelets. There are families of wavelet to employ. Daubechies wavelets are implemented on speech signals [4][5]. The decomposition process is performed using a pair of filters which convolve the input signal and then decimate it into approximation coefficients (low frequency component) and detail coefficients (high frequency component). The process is repeated until a final level is reached. Figure 1 shows an example for a speech signal, whereas figure 2 show the corresponding wavelet components for decomposition level that equals one.

A robust method, that is based on wavelet transform, is proposed for speech feature extraction. Figure 3 shows the block diagram for the extraction of wavelet-based MFCC feature vectors. First, the wavelet transform is applied iteratively for the decomposition of speech signals. The decomposition level is chosen to eliminate noise from the signals. Then, MFCC feature vector of the lowest frequency sub-band is calculated.

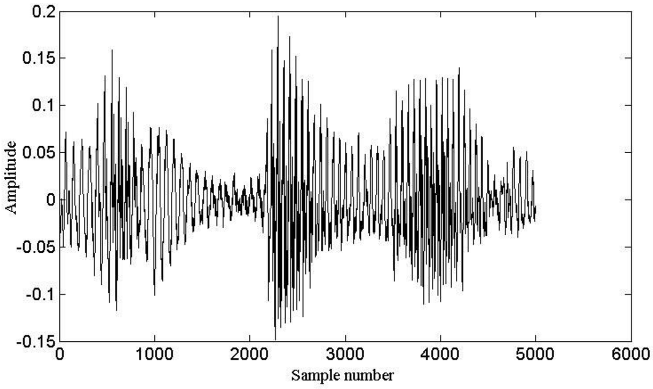

Figure 1: Speech signal, sampling frequency = 8 kHz.

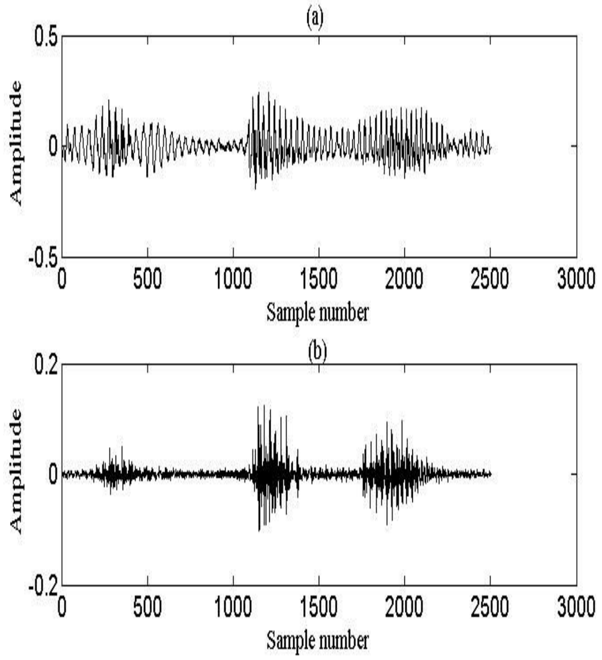

Figure 2: The wavelet decomposition for the speech signal in figure 1. Decomposition level =1. (a) Low-frequency component. (b) High-frequency component.

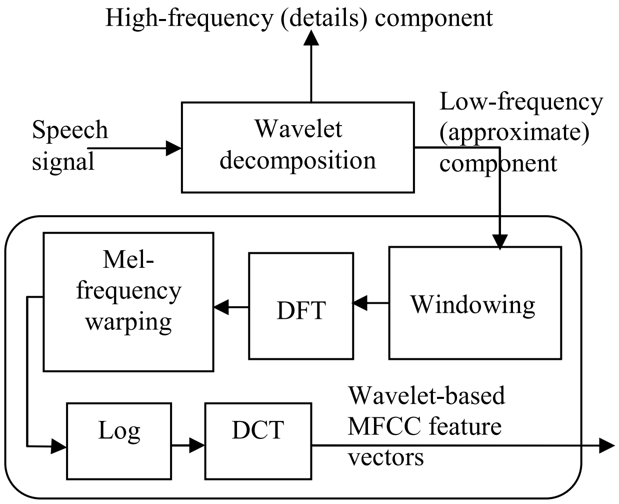
High-frequency (details) component
Speech signal
Wavelet decomposition
Low-frequency (approximate) component
Mel-frequency warping
DFT
Windowing
Log
DCT
Wavelet-based MFCC feature vectors

Figure 3: Extraction of wavelet-based MFCC feature vectors.

Effect of reverberant environments

The room reverberation is simulated by means of comb filters [6]. Figure 4 shows the room impulse response that is simulated for reverberation time that equals 2 s. The impulse response consists of the direct sound followed by a train of equidistant reflections. The time distribution affects the room acoustic quality, as it can cause coloration [7]. Then, Reverberation affects clearly the perceived signals. However, it degrades the performance of the speaker identification process [8]. The perception of coloration is predicted using the autocorrelation analysis. The temporal diffusion index [9], Δ, is introduced in equation (2), where $\varphi(t = 0)$is the central maximum of the autocorrelation, and $\varphi(t = \tau)$is the next side maximum. Coloration is audible if this side maximum is caused by a single strong reflection or by a number of successive reflections.

$$\Delta = \varphi(t = o)/\varphi(t = \tau) \tag{2}$$

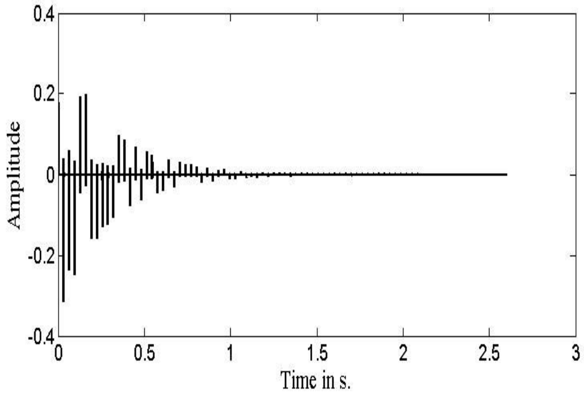

Figure 4: Room reverberation. Reverberation time = 2 s.

Speaker identification model

Once the feature sets of sound signal data have been extracted, a classifier system must be implemented. In this paper, Gaussian mixture model (GMM) is used [3]. The GMM uses density models to describe the distribution of a data set. Based on this model, the class-conditional probability of a target, that evaluates the probability of a feature vector for a given target model, can be computed.

In the training phase, each speaker is represented by a reference model λ_i that has unique, dependent, parameter values. Within the test phase, the feature vector is computed for each frame of the test signal. If x^n is a test feature vector in frame *n*, and the class-conditional probability density function is denoted as $p(x^n|\lambda_i)$, then the target identification is executed by determining the target model, i^*, that maximizes this class model-conditional probability density as shown in equation (3), where S is the number of the reference models.

$$i^* = arg \max_{1 \le i \le S} p(x^n|\lambda_i) \tag{3}$$

Database

The Database consists of 5 English sentences that are spoken by 12 speakers and are used for the classification problem. Now the data set contains (5×12= 60) audio files from 12 different speakers. Each file was sampled at 8 kHz.

Simulation & Results

Two experiments are conducted. The first experiment is conducted to find the effect of using wavelet transform on the perception of coloration. Three room impulse responses are simulated using comb filters, and the reverberation time T_{60} varies from 1s to 3s. Then, the wavelet transform is applied, and the low-frequency (approximate) component is obtained. The temporal diffusion index is calculated for each simulated room response, and for its corresponding approximate component. Table 1 shows the corresponding values for the temporal diffusion index, wavelet decomposition level that equals three is used while calculating the approximate component. Coloration is strongly perceived at T_{60} = 3s, where the smallest value for the temporal diffusion index is found (Δ_1 = 2.77). Comparing the value of the temporal diffusion index without wavelet decomposition Δ_1 to that with wavelet decomposition Δ_2, it is obvious that Δ_2 is greater than Δ_1 at T_{60} = 2 s and 3 s, which means that the perceived coloration is reduced when applying wavelet transform, whereas at T_{60} = 1 s the perceived coloration does not differ significantly.

Table 1: Temporal diffusion index, Δ_1 for the simulated room response , and Δ_2 for its approximate component using various values of reverberation time, T_{60}. Wavelet decomposition level=3.

T_{60}(s)	1	2	3
Δ_1	5.4	4.12	2.77
Δ_2	5	5.42	4.09

The second experiment aims to find the optimal wavelet decomposition level that maximize the identification rate in reverberant environments. One sentence per speaker is employed within the training phase, whereas the remaining ones are used within the test phase. Clean data is employed within the train. Room reverberation is simulated by means of comb filters using a certain value of reverberation time. The room response is convolved with each of the remaining signals to obtain the reverberant ones. The reverberant signals are used within the test phase.

Wavelet transform is applied to the speech signals to obtain the low frequency component. Next the data are segmented into approximately 128 samples per frame, overlapped by 50% of this frame. A Hamming window is then applied to each frame. 26 Mel-filter bank is constructed, then 12 MFCC coefficients are extracted. GMM is used for speaker identification, two Gaussian components are used within the model. The identification rate is defined as the ratio of the number of correctly identified targets to the total number of targets. However, the identification process is repeated using various values for the wavelet decomposition level and the identification rate is calculated for each case.

It was found that without using wavelet decomposition, the identification rate in environment of 1s reverberation time reaches 100%, whereas the identification rate is affected at reverberation time 2 s, and 3 s. However, wavelet decomposition is applied to improve the identification rate at reverberation time 2 s, and 3 s. Tables 2, and 3 show the effect of wavelet decomposition on the identification rate, IR (%), for the reverberation time 2s, and 3s respectively. The results show that the identification rate increases when applying wavelet transform to the speech signals. Table 3 shows that for a room of reverberation time of 3 s, the identification rate is 75% without wavelet transform, whereas it reaches 85.42% after applying wavelet transform. Similarly, table 2 shows that applying wavelet decomposition may yield to 100% identification rate. From table 2 and table 3, it is obvious that the highest identification rate is reached at wavelet decomposition level = 3.

Table 2: Identification Rate, IR (%) in a room of reverberation time (T60 = 2 s).

	Without wavelet decomposition	With Wavelet decomposition level			
		1	2	3	4
IR(%)	89.58	97.92	95.85	100	93.75

Table 3: Identification Rate, IR (%) in a room of reverberation time (T60= 3 s).

	Without wavelet decomposition	With Wavelet decomposition level			
		1	2	3	4
IR(%)	75	83.33	81.25	85.42	70.83

Conclusions

This paper employs the temporal diffusion index to predict the perception of coloration in reverberant environments. It is found that wavelet transform reduces the perceived coloration. The paper suggests the wavelet-based MFCC to describe the speech signal. Gaussian mixture model uses the feature vector to investigate the performance of the recognition process in reverberant environments. Employing the wavelet-based MFCC feature vector improves the identification process, and the paper specifies wavelet decomposition level that equals three, as it provides the highest identification rate.

References

[1] Antonio M. Peinado and Jose C. Segura, Speech recognition over digital channels: Robustness and standards. John Wiley & Sons Ltd, 2006.

[2] Wan-Chen Chen, Ching-Tang Hsieh and Chih-Tsu Hsu, "Robust speaker identification system based on two-stage vector quantization", Tamkang Journal of Science and Engineering 11 - 4 (2008), 357 -366.

[3] Tomi Kinnunen, and Haishou Li, “An overview of Text-independent speaker recognition: From features to supervectors”, Speech communication 52 (2010), 12 - 40.

[4] Ingrid Daubechies, "The wavelet transform, Time-frequency localization and signal analysis", IEEE transactions on information theory 36 - 5 (1990), 961 - 1005.

[5] V.S.R. Kumari and Dileep Kumar Devarakonda, " A Wavelet Based Denoising of Speech Signal ", International Journal of Engineering Trends and Technology (IJETT) 5 - 2 (2013), 107 - 115.

[6] Schroeder M.R., "Natural sounding artificial reverberators", AES 10 - 62 (1962), 219 - 223.

[7] Noha Korany, "Measuring sound coloration due to synthesized room reverberation", Fortschritte der Akustik – Deutsche Gesellschaft fuer Akustik, DAGA 2008, Dresden, Germany, 2008, 607 - 608.

[8] N. Korany, "Speaker identification in reverberant environments", Proceedings of Meetings on Acoustics, © 2013 Acoustical Society of America 19 - 060010 (2013), 1-8.

[9] Kuttruff, H., Room acoustics. Elsevier Science publishers, 2000.