PROCEEDINGS of the
23rd International Congress on Acoustics

9 to 13 September 2019 in Aachen, Germany

Performance evaluation of autocorrelation technique for automatic speaker identification in various environments

Noha KORANY[1]

[1] Alexandria University, Egypt

ABSTRACT

This paper aims to improve the performance of automatic speaker identification in various environments. Feature extraction is the first stage for the identification process. It is a challenge to extract those features that express the speakers' individualities. The speaker variation is highly related to the spectrum of its speech signals. Hence, the autocorrelation technique and the discrete cosine transform are applied for the feature extraction of speech signals. Then, those features are employed by a classifier engine to determine the speaker identity. In this paper, Gaussian mixture model is used as the recognition engine. The performance of the speaker identification system is investigated for various settings of the applied technique. Moreover, the effect of additive and convolutional noise on the performance of the identification process is investigated.

Keywords: Identification, Speech, Autocorrelation

1. INTRODUCTION

The speaker identification process aimed to identify the speaker based on his speech signals. Automatic speaker identification consists of two stages: feature extraction and feature classification by means of a recognition engine. Many features have been extracted and successfully employed by the classifier for the recognition purpose. Mel-frequency cepstral coefficients (MFCC), perceptual linear predictive cepstral coefficients (PLPCC) have proved good results for identifying speakers in clean environments (1). Another techniques are suggested to improve the identification process in noisy environments (2, 3).

The speech spectrum reflects the speaker characteristics as well as it carries the message information. These characteristics are related to the physical system of speech production for individuals. Hence, the feature extracted from the spectral characteristics of the speech signals are the most useful ones for automatic determination of the speaker identity. The paper applies the autocorrelation technique and the discrete cosine transform (DCT) for the feature extraction stage. This autocorrelation-based DCT feature has been specified for ships' classification by means of the noise produced by their platform (4).

The paper aims to determine the spectral components that fit the best for the identification process. The performance of the identification process is investigated in reverberant environments and in the presence of additive white Gaussian noise. The paper is organized as follows. The next section discusses the speech production mechanism and its relation to speaker individuality. Section 3 describes the technique used for feature extraction. Section 4 presents the database used, whereas section 5 describes the experiments and discuss the results. Finally, section 6 presents the paper summary and the final conclusions.

2. SPEECH PRODUCTION AND SPEAKER INDIVIDUALITY

Speech production takes place in two separate processes. The excitation process which consists of vibration of vocal folds for voiced sounds or noise turbulence for unvoiced sounds. In the second process, the sound is filtered through vocal and nasal tracts and it is radiated from the lips and nostrils.

Speaker variation is related to difference in vocal tract length, and vocal folds variation in both size

[1] Noha.korany@alexu.edu.eg

and mass. These variations between speakers affect the pitch of voiced sounds and their resonance (formant) frequencies (5).

The voiced pitch corresponds to the fundamental frequency (F0). Large variation in F0 between speakers enables normal listeners to identify them. Filter bank configurations are chosen for spectral coding of speech signals. The low frequency region (100 – 500 Hz) provides better spectral coding of F0, whereas the middle frequency region (500 – 1500 Hz) represents the vowel formants. The high frequency region (1500 – 5000 Hz) represents the temporal speech aspects (6).

3. FEATURE EXTRACTION TECHNIQUE

The speech spectrum carries the message information and the individual characteristics of the speaker. Figure 1 shows the spectrum of a speech segment for two different speakers and same output sound. It is obvious that the variation in speech spectrum is related to speaker discrimination.

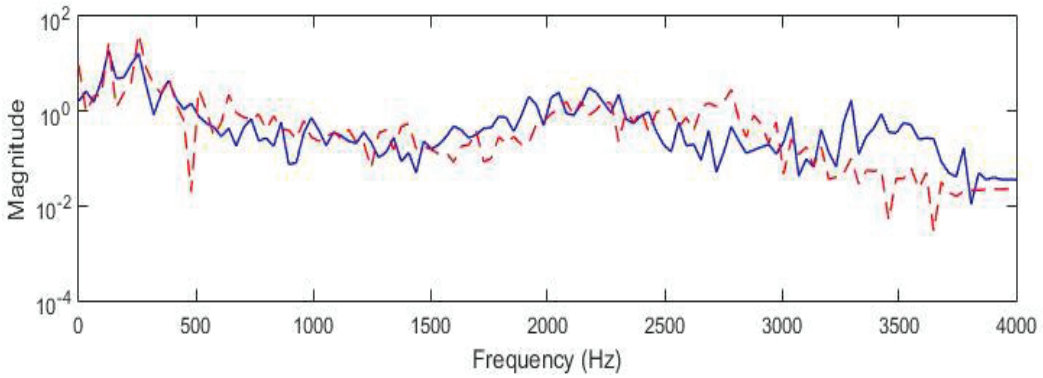

Figure 1 – Spectrum of speech segment for two different speakers.

The power spectral density describes the distribution of the power of the speech signal with frequency. For a speech segment x(t), the power spectral density $S_x(f)$ is defined in the following equation, where $R_x(\tau)$ is the autocorrelation function for the speech segment, and $\mathcal{F}$ denotes Fourier transform.

$$S_x(f) = \mathcal{F}\big(R_x(\tau)\big) \tag{1}$$

It is aimed to extract features that represent the strength of the variations with frequency, and hence characterize the speakers. This is done by applying the autocorrelation function and the discrete cosine transform (DCT) to the speech signals (4). Figure 2 shows the feature extraction method. First, the autocorrelation function of the speech signal is computed, and then windowed using hamming window of 30 ms long. Finally, DCT is applied to the windowed signal, and the autocorrelation-based DCT feature vector is obtained for each segment.

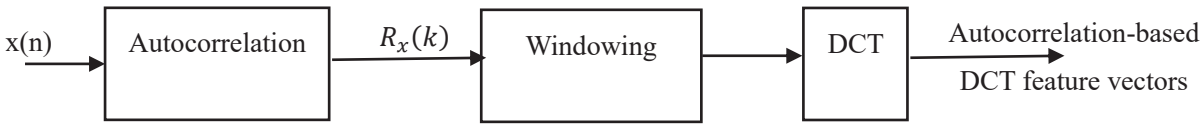

Figure 2 – Block diagram of the feature extraction technique.

4. DATABASE

The database consists of sound produced by twelve speakers, eleven males and one female. Each speech signal is sampled at 8000 Hz; 16-bit quantization level is used. Clean signals are used for the training phase, whereas reverberant signals and reverberant signals plus additive white Gaussian noise are employed for the test phase. Room reverberation is simulated using comb filters (7). Four room impulse responses are simulated for reverberation time that varies from 1 to 3 s. Voxforge database is used. A set of twelve spoken sentences is directly used within the training phase. Another set of twelve spoken sentences is chosen, and convoluted with each of the simulated room impulse response to obtain four sets of reverberant speech signals. Finally, additive white Gaussian noise is added to each speech signal of a certain reverberant set, and hence a fifth set of noisy speech signals is obtained. Those five sets of noisy speech signals are used for the test phase.

5. EXPERIMENTS AND RESULTS

The autocorrelation-based DCT feature vectors for clean and noisy speech signals are extracted, then they are employed by the Gaussian mixture model for the identification problem (8, 9). Two Gaussian components are used (10). Clean signals are used for the training phase, whereas the noisy ones are employed for the testing phase. The five sets of various type of noisy signals are used, and the identification rate is calculated using each set of these noisy signals. Three experiments are conducted to investigate the performance of the identification process in noisy environments. The experiments aim to specify the relevant frequency band that fits the best for automatic speaker identification in various noisy environments. The following equation relates the coefficient number, k, to its frequency, f. f_s is the sampling frequency, and N is the number of samples per frame.

$$k = 2(N-1)f/f_s \tag{2}$$

The first experiment aims to determine the effect of the number of the autocorrelation-based DCT coefficients on the identification rate. The feature vectors are extracted from reverberant speech signals, and they are employed by the classifier within the test phase. The number of coefficients varies from 12 to 256 according to the number of octave bands used. The DC coefficient is discarded. Table 1 shows the cutoff frequencies, the corresponding number of coefficients used, and the resulted identification rate for the various values of reverberation time.

Table 1 – The identification rate versus number of coefficients for four set of reverberant testing signals

Cutoff frequency, Hz	Number of coefficients used	Identification rate, %, for reverberation time of			
		1 s	1.5 s	2 s	3 s
177	11	83.33	66.67	83.33	66.67
354	23	83.33	75.00	91.67	75.00
707	45	91.67	75.00	83.33	83.33
1414	90	75.00	75.00	83.33	66.67
2828	180	66.67	75.00	83.33	66.67
4000	255	66.67	66.67	66.67	66.67

Table 1 concludes that the coefficients that correspond to the low frequency region (88 – 707 Hz) yields to the highest identification rate for various set of reverberant testing speech signals. For reverberation time of 1 s, maximum identification rate of 91.67% is reached when 45 autocorrelation-based DCT coefficients are used. Those 45 coefficients correspond to a cutoff frequency of 707 Hz. Similar conclusion is showed for reverberation time of 3s, and a maximum identification rate of 83.33% is obtained. In the case of 2s reverberation time, maximum identification rate of 91.67% is reached for 23 coefficients that correspond to a cutoff frequency of 354 Hz.

The second experiment is conducted to determine the number of coefficients within the low frequency region (100 – 500 Hz), which improve the identification rate in reverberant environments. Table 2 shows the cutoff frequencies for filter-bank within the low frequency region (6), the corresponding number of coefficients used, and the resulted identification rate for the various values of reverberation time. It concludes that maximum identification rate is reached for 19 autocorrelation-based DCT coefficients that correspond to a cutoff frequency of 300 Hz. Maximum identification rate of 91.67% and 100% is reached for reverberant testing signals at reverberation time of 1s and 2s respectively.

The third experiment is conducted to evaluate the effect of the presence of both additive and reverberant noise on the identification rate. In this experiment, additive white Gaussian noise is added to the reverberant speech signals and the fifth set of noisy signals is obtained. Then, 20 coefficients are extracted from those noisy signals and are employed by the classifier within the test phase. Figure 3 shows the identification rate versus various values of signal-to-noise ratio (SNR). It is found that high identification rate is obtained for SNR that equals to or greater than 5 dB. At 0 dB SNR, the identification rate is decreased to 58.33%.

Table 2 – The identification rate within the low frequency region for four set of reverberant testing signals

Cutoff frequency, Hz	Number of coefficients used	Identification rate, %, for reverberation time of			
		1 s	1.5 s	2 s	3 s
100	6	58.33	58.33	75.00	66.67
167	11	83.33	66.67	83.33	75.00
233	15	83.33	75.00	83.33	75.00
300	19	91.67	75.00	100	75.00
367	23	75.00	75.00	91.67	75.00
433	28	83.33	75.00	91.67	75.00

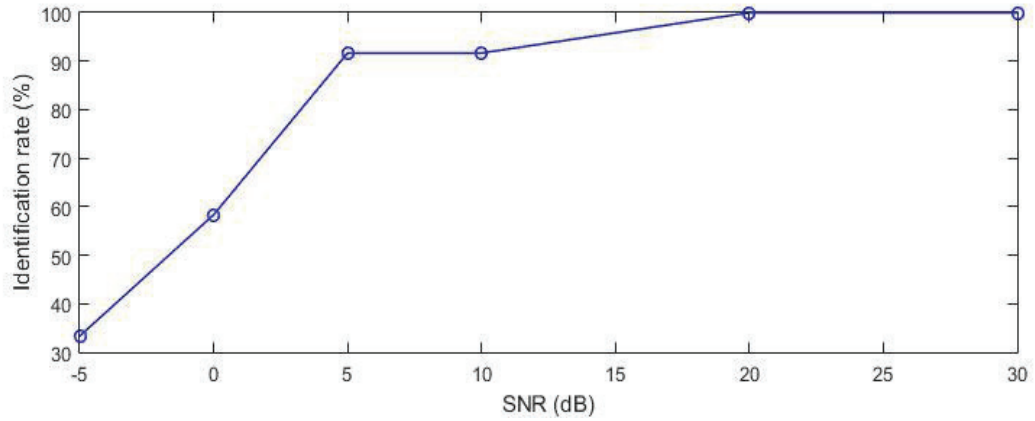

Figure 3 – Identification rate versus SNR, 20 coefficients extracted, reverberation time = 2s.

6. CONCLUSIONS

The paper discusses the problem of automatic speaker identification in various noisy environments. The autocorrelation method and the discrete cosine transform are applied to extract the speech feature vectors and those vectors are employed by Gaussian mixture model to determine the speaker identity.

Clean speech signals are used for the training phase, whereas reverberant speech signals are used for the test phase. The reverberation time is varied to construct various sets of testing signals, and the identification rate is calculated for each testing set. The paper specifies extracting the autocorrelation-based DCT coefficients that correspond to the low frequency band of cutoff frequency of 300 Hz as it yields to the highest identification rate.

Moreover, the effect of both reverberation and additive noise is investigated for various SNR. It is found that employing the autocorrelation-based DCT coefficients that correspond to the specified low frequency band improves the performance of the identification process for SNR equals or greater than 5 dB.

REFERENCES

1. Korany N, Speaker identification in reverberant environments. Proc ICA 2013; 2-7 June 2013; Montreal, Canada 2013.p.1-8.
2. Gong Y, Speech recognition in noisy environments: A survey. Speech Communications 1995;16:261-291.
3. Korany N, Measuring sound coloration due to synthesized room reverberation. Fortschritte der Akustik – Deutsche Gesellschaft fuer Akustik DAGA 2008; Dresden, Germany 2008. p. 607-8.
4. Korany N, Classification of ships using autocorrelation technique for feature extraction of the underwater acoustic noise. Proc Inter-noise 2016; Hamburg, Germany 2016. p. 7097-7102.

5. Mueller Ch., Speaker Classification I - Fundamentals, Features, and Methods. Berlin Heidelberg, Germany: Springer; 2007.
6. Carroll J, Zeng F, Fundamental frequency discrimination and speech perception in noise in cochlear implant simulations. Hearing research 2007;231:42-53.
7. Schroeder M.R., Natural sounding artificial reverberator, AES 1962; 10 (62):219-223.
8. Reynolds D, Rose R., Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models. IEEE Transactions on Speech & Audio Processing 1995; 3(1):72-83.
9. Do CB, Batzogloo S, What is the Expectation Maximization Algorithm? Nature Biotechnology 2008; 26: 897-9.
10. Korany N, Abd Elzaher M, Khater H, Investigation about the performance of GMM for recognition of underwater acoustics signals. Fortschritte der Akustik – Deutsche Gesellschaft fuer Akustik DAGA 2012; Darmstadt, Germany 2012. p. 655-6.