Bispectrum Mel-frequency Cepstrum Coefficients for Robust Speaker Identification

Ufuk Ülüğ[1], *Tolga Esat Özkurt*[2], *Tayfun Akgül*[1]

[1]Department of Electronics and Communications Engineering, Istanbul Technical University, Istanbul, Turkey
[2]Department of Electrical and Computer Engineering, University of Pittsburgh, Pittsburgh, PA, USA

ulug@itu.edu.tr, tolga@neuronet.pitt.edu, tayfun.akgul@itu.edu.tr

Abstract

In this paper, we introduce the use of bispectrum slice for mel-frequency cepstrum coefficients as robust text-independent speaker identification. The main advantage of using the bispectrum is to be able to suppress additive Gaussian noise while preserving the phase information of the signal. In order to obtain cepstral coefficients, features of the speech signal are extracted by mel-frequency filter banks, the cosine transform and the logarithm operator. Under various noisy test utterances, we compare and present the performances of the methods which use the bispectrum and the classical mel-frequency cepstrum coefficients.

1. Introduction

Speech signals yield information about the identity of the speaker as well as the content of the speech. Speaker recognition methods have found various applications such as security, voiced internet applications and telephone banking. Such systems consist mainly two parts: Speaker identification and verification. While speaker identification determines the identity of the speaker among a group of people from text dependent or independent speech signal, speaker verification is utilized as a second step to ensure the validity of the resultant speaker obtained by the identification process.

In speaker identification systems, speech signals are recorded and saved into a database. Training sets, which consist of the feature vectors of previously archived speech signals, are compared with the test sets. The conditions for obtaining the training set and the test set can be tremendously different. While the former can usually be obtained in noiseless environments, the test set may not. This may lead to an important decrease in the performance of a speaker identification system. Various methods have been proposed in literature in order to prevent this problem. They mainly include robust feature extraction, speech enhancement techniques and noise compensation [1].

The main advantage of using higher order statistics is to be able to suppress Gaussian noise unlike the classical autocorrelation-based (power spectrum-based) methods. For example, in [2] a particular part of bispectrum is suggested for feature extraction for speaker identification which is shown to be robust to additive Gaussian noise compared to the classical cepstrum.

In this study, we propose to use a bispectrum slice for the computation of mel-frequency cepstrum coefficients as robust features in a text-independent speaker identification system. The organization of the paper is as follows: Section 2 is for a brief explanation of feature extraction, bispectrum and sum-of-cumulants. In Section 3, the speaker identification systems and Gaussian mixture models are summarized. Simulation results are presented in Section 4. Conclusion is given in Section 5.

2. Feature Extraction Using Bispectrum Slice

Feature extraction of a speech signal is mostly based on the spectrum because any information about the characteristics of the vocal track can be obtained from the spectrum [3]. Although the spectrum of a speech signal can be defined by different models, using filter banks instead of linear prediction analysis provides more robust speech features. In this study, we extend to use bispectrum slice instead of the classical spectrum cepstrum coefficients obtained by the mel-frequency filter banks. The brief introduction of the bispectrum slice is given below.

2.1. Bispectrum Slice

If the autotripplecorrelation of any discrete signal $x(n)$ is

$$c(\tau_1, \tau_2) = E[x(n)x(n+\tau_1)x(n+\tau_2)] \qquad (1)$$

then the bispectrum $B(\omega_1, \omega_2)$ is defined as the 2-D Fourier transform of its autotripplecorrelation [4]:

$$B(\omega_1,\omega_2) = F\{c(\tau_1,\tau_2)\} \qquad (2)$$

where $E[.]$ is the expected value and $F\{.\}$ is Fourier transform. 1-D inverse Fourier transform of the bispectrum, *q(n),* on the $\omega_1 = \omega_2$ line is defined as the sum of cumulants [4]:

$$q(n) = x(n)*x(n)*k(n) \tag{3}$$

where * denotes convolution operator and

$$k(n)= \begin{cases} x(N\text{-}1\text{-}N/2) & n \text{ odd} \\ 0 & n \text{ even} \end{cases} \tag{4}$$

for the signal with *n=0,1,...,N-1*. The sequence, *q(n),* has *4N-3* samples with *n=-2(N-1),...,-1,0,1,...,2(N-1).*

2.2. Feature Extraction

Before the analysis, speech signal is divided into segments of 16 ms length and 10 ms overlap. After the estimation of the sum of cumulants using (3) for these frames, Mel-Frequency Cepstrum Coefficients (MFCC) are obtained by utilizing mel-frequency filter banks. Figure 1 shows the block diagram of the steps for the extraction of features.

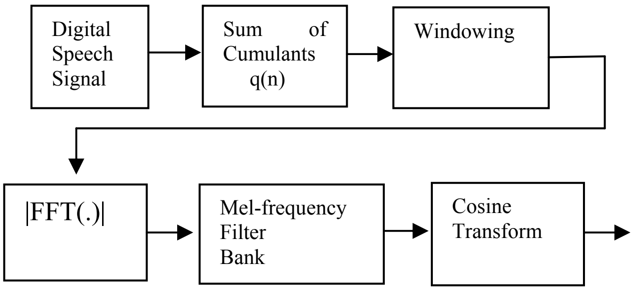

Figure 1: Feature Extraction Block Diagram

3. Speaker Identification System

Gaussian Mixture Model (GMM) is used for speaker identification in this study. In text-independent speaker identification systems, the performance of GMM is known to be relatively high. Let *M* be the order, then GMM is expressed as:

$$p(\mathbf{x}\,|\,\lambda) = \sum_{i=1}^{M} w_i b_i(\mathbf{x}) \tag{5}$$

where

$$b_i(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}\,|\mathbf{C}_i|^{1/2}} \exp(-0.5(\mathbf{x}-\boldsymbol{\mu}_i)^T \mathbf{C}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)) \tag{6}$$

In that expression ***x*** is a *D* dimensional random vector, w_i *(i=1,...,M)* are weight coefficients and $b_i(\boldsymbol{x})$ are component density functions. The mixture weight coefficients satisfy the below equation:

$$\sum_{i=1}^{M} w_i = 1 \tag{7}$$

GMM is defined with $\lambda = \{w_i, \boldsymbol{\mu}_i, \boldsymbol{C}_i\}$, where $\boldsymbol{\mu}_i$ is the mean value vectors and $\boldsymbol{C}_i$ is the covariance matrix. For speaker identification systems, λ can be represented to model the speaker. Then the model parameters are estimated by expectation maximization method, which maximizes the model likelihood iteratively [6].

4. Simulation Results

For the training and test sets we use TIMIT database. The training set contains 50 male speech excerpts with the same dialect for various lengths. Test data contains 5 different sentences from different speech segments for each training member. The order of GMM is chosen as *M=40.* The performance of the proposed bispectrum-based mel-frequency cepstrum coefficients are compared with the classical mel-frequency cepstrum coefficients.

The evaluation is done by the normalized total score which is the logarithmic extraction of the actual speaker's likelihood from the maximum likelihood of the other speakers apart from the actual speaker:

$$\log(L(X)) = \log p(X\,|\,S = S_c, \lambda) - \max(p(X\,|\,S \neq S_c, \lambda)) \tag{9}$$

Here, X is the features in the test set, S is the speakers in the training set and S_c is the actual speaker with λ model parameters. If the normalized score is positive, the speaker is estimated correctly.

First, we simulated our speaker identification system with noiseless training sets and noise-free test sets and reached almost %100 correct identification with both spectral and bispectral methods. Then, in order to compare the robustness of bispectrum to spectrum, we add white Gaussian noise with SNR = 40, 20, 15, 10, 5 dBs to the speech signal. Speaker identification performances for both methods are given in Figure 2. Below 10 dB, the performance results of both the spectrum and the bispectrum slice is low but above 10 dB, the speaker identification rate is better when the bispectral method is used.

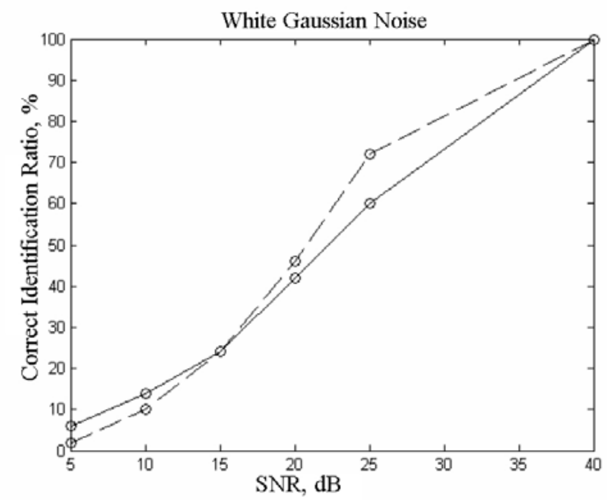

Figure 2: Correct identification ratios vs. SNR for spectral (straight-line) and bispectral methods (dashed-line) are used when white Gaussian noise added.

Various noises are also added to the speech signals in test set to show the effect of the proposed methods. Real noise samples, such as babble, car and factory noises, are collected from NOISEX database of NATO Speech Signal Workgroup [8]. We down sampled that noise samples from 19.98 kHz to 16 kHz and added them to speech signals with SNR = 30, 25, 15, 10, 5, 0 dBs. Although, the distributions of these noises may not be Gaussian, they are symmetric since their skewnesses are closer to zero. The histograms of the noises can be seen in Figure 3. Test results for real noise experiment are provided in Figures 4, 5, 6, which show that, at any SNR above 0 dB, the performance is higher when the bispectrum slice is utilized for feature extraction in our speaker identification system.

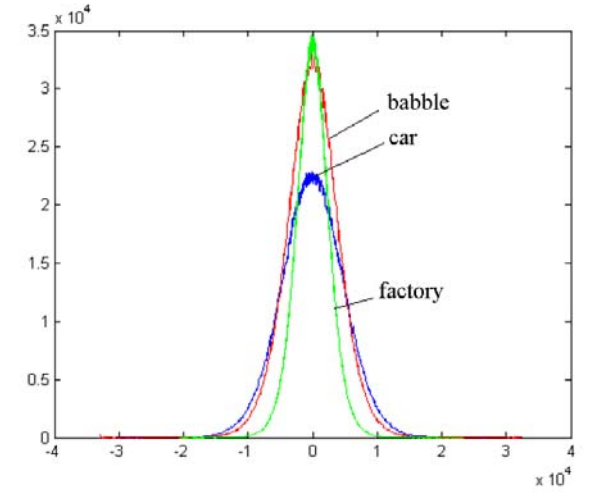

Figure 3: Noise histograms for babble, car and factory

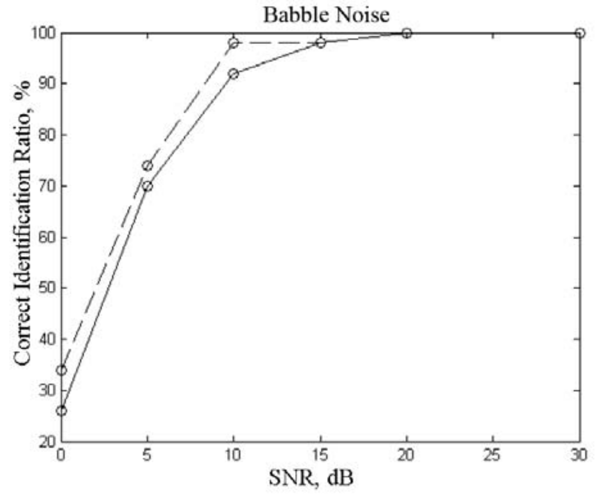

Figure 4: Correct identification ratios vs. SNR for spectral (straight-line) and bispectral methods (dashed-line) are used when babble noise added.

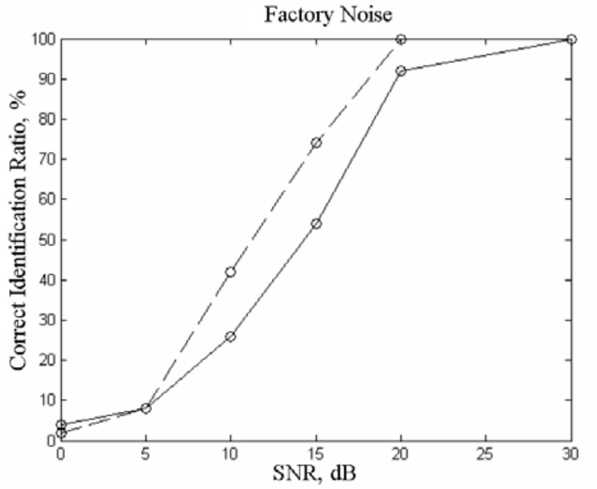

Figure 5: Correct identification ratios vs. SNR for spectral (straight-line) and bispectral methods (dashed-line) are used when factory noise added.

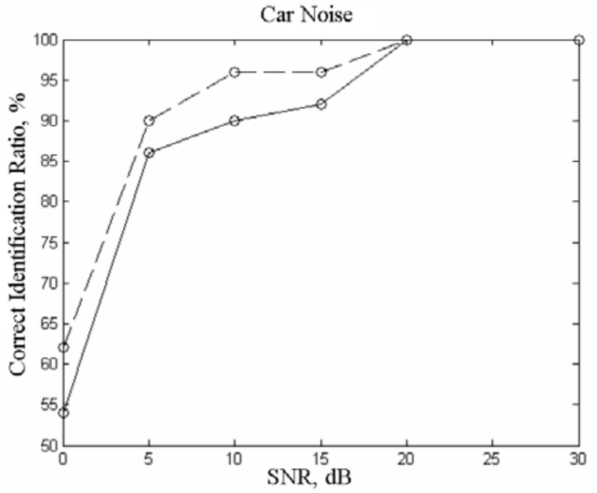

Figure 6: Correct identification ratios vs. SNR for spectral (straight-line) and bispectral methods (dashed-line) are used when car noise added.

The next step in our study is to compare the robustness of spectral and bispectral methods to colored noise produced by wavelet-based methods. We will present the results and the comparisons during the presentation and in the final version of the paper.

5. Conclusion

It is known that additive Gaussian noise deteriorates the identification rate seriously in speaker identification systems. We propose to use bispectrum slice mel-frequency cepstrum features as robust and efficient features for text-independent speaker identification systems. We show the comparisons between the classical and bispectrum based methods.

6. References

[1] **Gong, Y.**, "Speech Recognition In Noisy Environments: A Survey," *Speech Communication*, vol. 16, pp. 261-291, 1995.

[2] **Wenndt, S., Shamsunder, S.,** "Bispectrum Features for Robust Speaker Identification," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing*, 1095-1098, 1997.

[3] **Campbell J.,** "Speaker Recognition: A Tutorial," *Proceedings of the IEEE*, vol. 85, no. 9, 1437-1462, 1997.

[4] **Akgül T., El-Jaroudi A.,** "Reconstruction of Mixedphase Signals From Sum-of-Autotripple-correleations Using Least Squares," *IEEE Transactions on SignalProcessing*, vol.46 , no.1, 250-254, 1998.

[5] **Reynolds D.,** "Robust Text-Independent Speaker Identification Using Gaussian Mixture Models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no.1, 4072-4075, 1995.

[6] **Bilmes J.,** "A Gentle Tutorial of the EM Algorithm and Its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models," International Computer Science Institute, 1998.

[7] **Furui S.,** *Digital Speech Processing, Synthesis and Recognition*, Marcel Dekker, New York, 2001.

[8] **Varga M. G., Steeneken H. J. M.,** "Assessment for Automatic Speech Recognition: II. NOISEX-92:A Database and an Experiment to Study the Effect of Additive Noise on Speech Recognition Systems," *Comput. Speech Lang.*, vol. 12, pp. 247-251, 1993.