A Text–Independent Speaker Identification System Based on the Zak Transform

Abdulnasir Hossen abhossen@squ.edu.om
Department of Electrical and Computer Engineering
College of Engineering
Sultan Qaboos University
P.O. Box 33, PC 123, Al-Khoud, Muscat, Oman

Said Al-Rawahi s_raw@yahoo.com
Department of Electrical and Computer Engineering
College of Engineering
Sultan Qaboos University
P.O. Box 33, PC 123, Al-Khoud, Muscat, Oman

Abstract

A novel text-independent speaker identification system based on the Zak transform is implemented. The data used in this paper are drawn from the ELSDSR database.
The efficiency of identification approaches 91.3% using a single test file and 100% using two test files. The method shows comparable efficiency results with the well known MFCC method with an advantage of being faster in both modeling and identification.

Keywords: Speaker identification, Zak transform, Feature extraction, Classification

1. INTRODUCTION

Speaker recognition systems are classified into speaker verification (SV) systems and speaker identification (SI) systems. The task of SV system is to verify the claimed identity of a person from his voice, while the SI system decides who the speaking person is from a database of speakers [1-2].
SI systems can be also classified into text-dependent or text-independent depending on whether the identification based on known utterance or for any given utterance.The SI system consists of two stages:

- Speaker enrolment or speaker modeling, in which features are extracted from all speakers and models are built for them.
- Speaker recognition, in which features are extracted from the speaker under test and a model is built for and compared with models of all speakers in the data set to find the closest speaker (matching).

One of the most successful techniques in speaker recognition is the technique that uses the mel-frequency cepstrum coefficients (MFCC) as a feature set and the Gaussian mixture model (GMM) for matching [3-6].

In this paper, the feature set used for modeling the speakers is based on the Zak transform, while the matching is performed using the Euclidean distance measure.
The new technique is compared with the MFCC technique in terms of identification accuracy and complexity. Both techniques are to be introduced in the next section.

2. METHODS

2.1 MFCC Technique

The perception system of human has some interesting facts. Within the human perception system there is a low-pass filter that blocks high frequency signals from being received. That is why human beings can not percept high frequency signal, where some animals do. Another interesting fact about human perception system is being a non-linear system. Human perception system perceives speech signals in a mel scale. A mel is a unit of measure of perceived pitch or frequency of a tone. It does not correspond linearly to the physical frequency of the tone. The mapping between real frequencies and mel scale is approximately linear below 1 KHz and logarithmic above [7].

Based on the non-linearity of the human perception system, mel-frequency emerged as to mimic the human perception system. The feature set named Mel-frequency Cepstrum Coefficients is obtained using information wrapped in a mel-frequency scale. The speech signal is decomposed into series of frames using a suitable window function, and then MFCC is applied to each frame. The coefficients are obtained through mel-scale filters and collectively named a mel-scale filter banks. These filters follow the mel-scale whereby band edges and centre frequencies of the filters are linear for low frequencies (< 1000) and logarithmically increase with increasing frequency [3]. Figure 1 illustrates the process involving the calculation of MFCC [6].

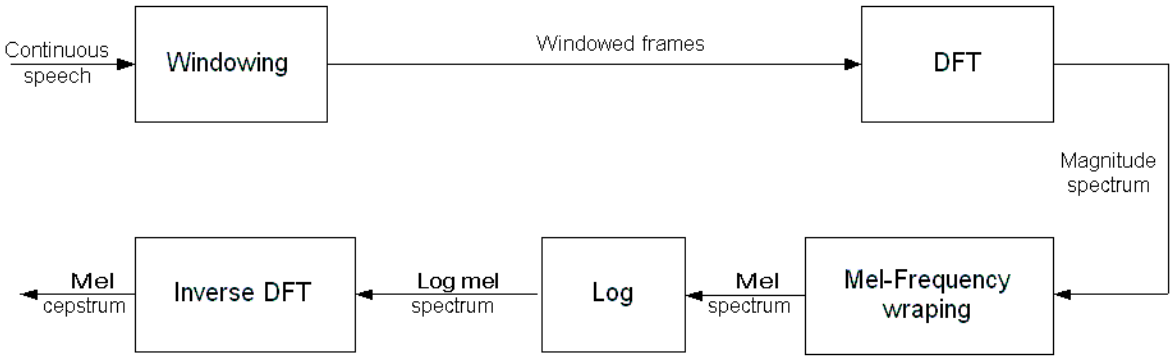

FIGURE 1: Block-Diagram for Computation of Mel-Cepstrum

The pattern-matching task involves computing a match score, which is a measure of the similarity between the input feature vectors and some model. Speaker models are constructed from the features extracted from the speech signal [1]. The comparison could be tested for maximum or minimum score or for a threshold value, dependable on the application.

Due to variability within a speech signal, multi-dimensional Gaussian probability density function (PDF) can be used to represent the signal probabilistically. The Gaussian pdf is state-dependent in that there is assigned a different Gaussian pdf for each acoustic sound class [3]. In this technique the feature set is modelled through GMM, where a speaker model is represented as the measure of the coefficients in terms of means, variances and weights of each mixture. The Gaussian pdf of a feature vector $\vec{x}$ for the *i* th state is written as [3]:

(1)

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{\frac{D}{2}}|\sum_i|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\vec{x}-\vec{\mu}_i)^T \sum_i^{-1}(\vec{x}-\vec{\mu}_i)\} \quad (1)$$

where $\vec{\mu}_i$ is the state mean vector, $\sum_i$ is the state covariance matrix, and D is the dimension of the feature vector. The vector $(\vec{x}-\vec{\mu}_i)^T$ denotes the matrix transpose of $\vec{x}-\vec{\mu}_i$, where $|\sum_i|$ and $\sum_i^{-1}$ indicate the determinant and inverse of matrix $\sum_i$ respectively. The mean vector $\vec{\mu}_i$ is the expected value of the elements of the feature vector $\vec{x}$, while the covariance matrix $\sum_i$ represents the cross-correlations (*off-diagonal terms*) and the variance (*diagonal terms*) of the elements of the vector [3].

The speaker model λ represents the set of GMM mean, covariance and weight parameters [3]:

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad (2)$$

where p_i is the weight, $\vec{\mu}_i$ is the mean and $\sum_i$ is the covariance, for each acoustic class (*mixture*). The probability of a feature vector being in any one of *I* states (*or acoustic classes*) can be represented as the union of different Gaussian pdfs for a particular speaker [3]:

$$p(\vec{x}|\lambda) = \sum_{i=1}^{I} p_i b_i(\vec{x}) \quad (3)$$

The job of the GMM is to cluster the coefficients and elaborate all the coefficients extracted from all the frames into a mixture of Gaussian models having different means and variances.

2.2 Discrete Zak Transform

The Fourier transform has been recognized as the great tool for the study of stationary signals and processes where properties are statistically invariant over time.
However, it can not be used for the frequency analysis that is local in time. In recent years, several useful methods have been developed for the time--frequency signal analysis. They include the Gabor transform, Zak transform, and the wavelet transform.
Decomposition of a signal into a small number of elementry waveforms that are localized in time and frequency plays a remarkable role in signal processing. Such a decomposition reveals important structures in analyzing nonstationary signals such as speech and music [8].

Zak transform has been implemented in efficient computaion of Gabor's expansion coefficients in a most reliable and completely non--invasive biometric method in iris recognition system [9-10].
For numerical implementations, a Zak transform that is discrete in both time and frequency is required [11].

The discrete Zak transform (DZT) is a linear signal transformation that maps a discrete time signal *x[n]* onto a 2-D function of the discrete time index *n* and the continuous normalized frequency variable *w*. The DZT of *x[n]* sampled equidistantly at N points can be expressed as the 1-D DFT of the sequence *x [n+kP]* [12] :

$$Z_x(P,M) = \sum_{k=0}^{M-1} x(n+kP)e^{-jkw2\pi/M}, \quad (4)$$

where *MP*=N. Since the DZT is periodic both in frequency domain *w* (with the period 2π/*M*) and in time domain *n* (with period *P*), the fundamental zak interval is selected to be (*n* = 0,1,…,*P*-1) and (*w* = 0,1,…, *M*-1).

3. DATA

The speech data used in this work are drawn from the ELSDSR (English Language Speech Database for Speaker Recognition) [13]. This ELSDSR database, which is created by the technical university of Denmark, is English spoken by non-native speakers. It contains 23 speakers, 13 male and 10 female. Each speaker has 7 speech files associated with the training and 2 speech files associated with the testing. On average, the duration for reading the training data is 83 s and for reading the test files is 17.6 s.

4. IMPLEMENTATION AND RESULTS

4.1 Implementation

The system uses DZT with *P* = 8192 and *M* = 8. Each speaker is modeled with a matrix of size (8192, 8) by its Zak transform of a speech file of length (N = 8192*8) samples. During the recognition phase, a comparison of the Zak transform matrix of the speaker under test with all matrices (models) of the speakers in the data set is done. The closest speaker (the matched speaker) is identified by the speaker with the minimum Euclidean distance from the speaker under test. The Euclidean distance between two points is defined as:

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_m - y_m)^2} \quad (5)$$

The average of M minimums of M columns of the (M, M) matrix of distances is obtained and used to represent a score to show how close the speaker under test from the speaker of the compared model. The speaker with the minimum score is the identified speaker.

4.2 Results

Table 1 shows the results of identification (efficiency rate) in case of using the first test file or the second test file or both files in the test phase. The mis-identified speakers and the correct order of the speakers are also shown in this table.
Table 2 shows a comparison between MFCC method and the Zak method in speaker identification while either all 7 training files are used in the training phase or only 4 out of them. The Zak method shows the same efficiency rate 100% as the MFCC method if two test files are used.

Table 3 and Table 4 show the complexity comparison between MFCC and Zak methods in terms of modeling and identification times in seconds, respectively. It is clear that the Zak method is faster than MFCC method in both modeling and identification. The time-measurements are obtained using a Fujitsu Siemens computer with Intel Pentium M processor running at 2.00 GHz.

4.3 Modifications

The new technique is implemented also on short-length segments and the Zak transform is found as an average of the Zak transforms of all segments. The signal with P=8192 samples is divided into R segments of length P/R. The DZT is found for each segment of length P/R and then the average of the DZT of those R segments is obtained.

The modified Zak method shows equal efficiency results as R varies from 2 to 16. Table 5 shows such consistent results of efficiency of identification, which are better than the results obtained on the full-length P=8192.

The modeling time is increased from 0.312 s to 1.06 s, if R=8 segments is used instead of full-length signal. The modified method is still much faster than the MFCC method in modeling.

Table 6 shows the results of identification execution-time as R varies from 2 to 16. The identification times is less than that obtained with full-length Zak transform method and MFCC method.

Test Files	Efficiency Rate	Identified Speaker	Correct Speaker
File 1	91.3 %	2, 20	3,15
File 2	86.96 %	19, 22, 22	5,13,19
Both	100 %		

TABLE 1. .Results of Identification.

Test Files	All Training Files	All Training Files	First 4 Training Files	First 4 Training Files
	MFCC	Zak	MFCC	Zak
File 1	100 %	91.30 %	95.65 %	91.30 %
File 2	100 %	86.96 %	95.65 %	82.61 %
Both	100 %	100 %	95.65 %	95.65 %

TABLE 2. Comparison with MFCC in terms of Efficiency Rate

Method	All Training Files	First 4 Training Files
Zak	0.320	0.185
MFCC	9.122	6.010

TABLE 3. Comparison with MFCC in terms of Modeling Complexity

Method	Test File 1	Two Test Files
Zak	0.5278	0.5643
MFCC	0.812	1.888

TABLE 4. Comparison with MFCC in terms of Identification Complexity

Test Files	Efficiency Rate	Identified Speaker	Correct Speaker

File 1	95.65 %	10	16
File 2	91.3 %	11, 18	6, 21
Both	100 %		

TABLE 5. Results of Identification of the Modified Zak Method

R	1 Test File	2 Test Files
2	0.1126	0.2048
4	0.1335	0.2126
8	0.1575	0.2377
16	0.2113	0.2874

TABLE 6. Effects of Number of Segments on the Identification Time in Seconds

5. CONCLUSIONS

A novel text-independent speaker identification system is implemented using Zak transform coefficients as a feature set. The method is simple and shows 100% efficiency rate in case of using two files in the test phase in identifying 23 speakers forming the ELSDSR database.
Compared to MFCC method, the Zak methods shows a clear advantage in both modeling and identification complexity.

The new method is also improved by applying it on a basis of short--length segments and then an average DZT is computed. The improvement is achieved in both identification efficiency and time while more time is needed for the modeling compared to the full-length Zak method.

6. ACKNOWLEDGMENTS

The authors would like to thank Mrs. Ling Feng from the Technical University of Denmark for providing them with the ELSDSR data.

REFERENCES

1. Campbell, J. P.: "*Speaker recognition: A tutorial*", Proceedings of the IEEE, Vol.85(9), 1437-1462, 1997.
2. Naik, J. M.: "*Speaker verification: A tutorial*", IEEE Communications Magazine, 42-48, January 1990.

3. Quatieri, T. F.: "*Discrete-time speech signal processing: Principles and practice*", Prentice Hall, 2002.

4. Benesty, J., Sondhi, M. M., and Huang, Y.: "*Springer handbook of speech processing*", Springer, 2007.

5. Keshet, J., and Bengio, S.: "*Automatic speech and speaker recognition: Large margin and kernel methods*", John Wiley, 2009.

6. Karpov, E.: "*Real-time speaker identification*", Master Thesis, Department of Computer Science, University of Joensuu, 2003.

7. Deller, J.R., Hansen, J.H.L., Proakis, "*Discrete-Time Processing of Speech Signals*", IEEE Press, New York, NY, 2000.

8. Debnath, L.: "*Wavelet transforms and their applications*", Springer Verlag, 2001.

9. Czajka, A., and Pacut, A.: "*Zak's transform for automatic identity verification*", 4th International Conference on Recent Advances in soft computing RASC2002, Nottingham, United Kingdom, 374-379, December 2002.

10. Daugman, J.: "*Wavelet demodulation codes, statistical independence, and pattern recognition*", Institute of Mathematics and its Applications, Proc. 2nd IMA-IP, 244-260, London, Horwood, 2000.

11. Boelcskei, H., and Hlawatsch, F.: "*Discrete Zak transforms, polyphase transforms, and applications*", IEEE Transaction on Signal Processing, Vol. 45(4), 851--866, April 1997.
12. Janssen, A.: "*The Zak transform: A signal transform for sampled time-continuous signals*", Phillips J. Res., 43, 23--69, 1988.

13. ELSDSR database for speaker recognition, 2004, http://www.imm.dtu.dk/~lf/eLSDSR.htm