Available online at www.sciencedirect.com

SciVerse ScienceDirect

Speech Communication 54 (2012) 543–565

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Design, analysis and experimental evaluation of block based transformation in MFCC computation for speaker recognition

Md. Sahidullah *, Goutam Saha

Department of Electronics and Electrical Communication Engineering, Indian Institute of Technology, Kharagpur, Kharagpur 721 302, India

Received 18 April 2011; received in revised form 14 November 2011; accepted 18 November 2011
Available online 26 November 2011

Abstract

Standard Mel frequency cepstrum coefficient (MFCC) computation technique utilizes discrete cosine transform (DCT) for decorrelating log energies of filter bank output. The use of DCT is reasonable here as the covariance matrix of Mel filter bank log energy (MFLE) can be compared with that of highly correlated Markov-I process. This full-band based MFCC computation technique where each of the filter bank output has contribution to all coefficients, has two main disadvantages. First, the covariance matrix of the log energies does not exactly follow Markov-I property. Second, full-band based MFCC feature gets severely degraded when speech signal is corrupted with narrow-band channel noise, though few filter bank outputs may remain unaffected. In this work, we have studied a class of linear transformation techniques based on block wise transformation of MFLE which effectively decorrelate the filter bank log energies and also capture speech information in an efficient manner. A thorough study has been carried out on the block based transformation approach by investigating a new partitioning technique that highlights associated advantages. This article also reports a novel feature extraction scheme which captures complementary information to wide band information; that otherwise remains undetected by standard MFCC and proposed block transform (BT) techniques. The proposed features are evaluated on NIST SRE databases using Gaussian mixture model-universal background model (GMM-UBM) based speaker recognition system. We have obtained significant performance improvement over baseline features for both matched and mismatched condition, also for standard and narrow-band noises. The proposed method achieves significant performance improvement in presence of narrow-band noise when clubbed with missing feature theory based score computation scheme.

Keywords: Speaker recognition; MFCC; DCT; Correlation matrix; Decorrelation technique; Linear transformation; Block transform; Narrow-band noise; Missing feature theory

1. Introduction

Speaker recognition is a biometric authentication process where the characteristics of human voice are used as the attribute (Kinnunen and Li, 2010; Campbell et al., 2009). A state-of-the art speaker recognition system has three fundamental sections: a feature extraction unit for representing speech signal in a compact manner, a modeling scheme to characterize those features using statistical approach (Campbell, 1997), and lastly a classification scheme for characterizing the unknown utterance. Most of the feature extraction techniques use low level spectral information which conveys vocal tract characteristics. The spectral information is extracted from 20–30 ms of speech signal using squared magnitude of discrete Fourier transform (DFT). As vocal tract is a slowly varying system, speech signal is nearly stationary over this analysis window. Hence, DFT based spectrum estimation technique is quite suitable. A systematic study of various spectral features can be found in (Kinnunen, 2004). Out of all existing features, *Mel frequency cepstral coefficient* (MFCC) is the

* Corresponding author. Tel.: +91 3222 283556/1470; fax: +91 3222 255303.
E-mail addresses: sahidullahmd@gmail.com (Md. Sahidullah), gsaha@ece.iitkgp.ernet.in (G. Saha).

0167-6393/$ - see front matter
doi:10.1016/j.specom.2011.11.004

most popular and has become standard in speaker recognition system. MFCC is popular also due to the efficient computation schemes available for it and its robustness in presence of different noises.

In MFCC computation process, the speech signal is passed through several triangular filters which are spaced linearly in a perceptual Mel scale. The Mel filter bank log energy (MFLE) of each filters are calculated. Finally, cepstral coefficients are computed using linear transformation of MFLE. The linear transformation is essential here. The major reasons are as follows: (a) *improving the robustness*: the MFLEs are not much robust. They are very much susceptible to a small change in signal characteristics due to noise and other unwanted variabilities, (b) *decorrelation*: the log energy coefficients are highly correlated whereas uncorrelated features are preferred for statistical pattern recognition systems, specially for diagonal covariance based *Gaussian mixture model* (GMM) which is employed in today's speaker recognition system.

Amongst all linear transformation discrete cosine transform (DCT) is most popular and widely used for MFCC computation. The motivations behind the usage of DCT can be stated as follows. Firstly, the DCT is the sub-optimal approximation of the basis function of Karhunen–Loéve transform (KLT) when the correlation matrix of the sample closely approximates the correlation matrix of Markov-I process (Ahmed et al., 1974). The correlation matrix of MFLE data is fairly similar to the correlation matrix of first order Markov process. Secondly, DCT has the best energy compaction property for arbitrary data length compared to DFT and other sinusoidal transform like discrete sine transform (DST), discrete Hadamard transform (DHT), etc. (Oppenheim and Schafer, 1979). Though DCT based MFLE transformation technique is very popular, some studies have been carried out recently on further processing schemes of cepstral coefficient to improve the robustness against channel and other variabilities (Garreton et al., 2010; Hung and Wang, 2001; Nadersharif and Akbari, 2007). *Principal component analysis* (PCA) (Takiguchi and Ariki, 2007), *linear discriminant analysis* (LDA) (Kajarekar et al., 2001), *independent component analysis* (ICA) (Kwon and Lee, 2004), etc. are some traditional techniques which are also applied for formulating decorrelated features for speech processing applications.

Our proposed work is focused to design a linear transformation technique which can effectively preserve speech related information to improve the speaker recognition performance. Being motivated by the fact that the block wise filter bank outputs are more suitable for transformation using DCT, we have investigated block based transformation approach in case of traditional full-band based DCT which is applied to all the MFLE at a time. Earlier block based cosine transform has been applied for speech recognition (Jingdong et al., 2000). Recently, DCT is applied in a distributed manner (Sahidullah and Saha, 2009) to formulate feature for speaker identification. In image processing applications, DCT has also been applied in blocked manner (Jain, 2010). Subband DCT based coding method has been shown to be effective in image coding, image resizing schemes where DCT is computed for different block of subband (Jung et al., 1996; Mukherjee and Mitra, 2002). Here the signal is first divided into two parts: a high pass and a low pass and DCT is computed for each signals separately. On the other hand, subband based speaker recognition technology also gained attention as an alternative of conventional MFCC. In (Sivakumaran et al., 2003), different experimental results are reported based on subband DCT. During the last decade, several works have been carried out in subband processing based speaker recognition (Besacier and Bonastre, 2000; Finan et al., 2001; Damper and Higgins, 2003; Vale and Alcaim, 2008). The mathematical relationship between multi-band and full-band based MFCC coefficient are established in (Mak, 2002). In (Kim et al., 2008), subband DCT based MFCC is shown to perform better than full-band MFCC for different additive noises. There exists a number of other work on subband DCT or multi-band MFCC where it is shown to outperform existing baseline MFCC specially for partially corrupted speech signal (Besacier and Jean-Frantois, 1997; Ming et al., 2007; Jingdong et al., 2004). Though, it has played an effective role in improving performance of speech processing applications still multi block DCT is not much used in state-of-the art speaker recognition system. The main reason is that most of the existing works are at experimental level and the design issues related to multi-block configuration (i.e. number of bands, size of band, etc.) are yet to be precisely addressed. This is one of the main issue behind its unpopularity in spite of its superior empirical performance for speech and speaker recognition.

In our present work, the design issues related to block based MFCC computation scheme is addressed carefully along with a thorough experimental evaluation. The cepstral coefficient using multi-block DCT approach is systematically formulated. The scheme is also restructured for improving the performance of speaker recognition. The block transform (BT) based approach is shown to carry several levels of information. A novel block based approach is also proposed which has complementary information to the formerly proposed methods. The strengths of both the systems are combined using weighted linear fusion to get better performance. We have evaluated the performance of speaker recognition system with NIST SRE 2001 (for matched condition) and NIST SRE 2004 (for both matched and mismatched condition). The experimental result shows the superiority of our proposed block-based MFCC computation scheme for both the databases. As a final point, the paper proposes a technique where significant performance improvement is obtained for multi-block approach using linear transformation only. The system also performs better than standard MFCC for different types of noise. Additionally this system is significantly better than baseline system in case of narrow-band noise when