Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition

Jia Min Karen Kua[1,2], Tharmarajah Thiruvaran[1], Mohaddeseh Nosratighods[1]
Eliathamby Ambikairajah[1,2], Julien Epps[1,2]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Eveleigh 2015, Australia

jmkua@student.unsw.edu.au, thiruvaran@student.unsw.edu.au, hadis@unsw.edu.au,
ambi@ee.unsw.edu.au, j.epps@unsw.edu.au

Abstract

Most conventional features used in speaker recognition are based on spectral envelope characterizations such as Mel-scale filterbank cepstrum coefficients (MFCC), Linear Prediction Cepstrum Coefficient (LPCC) and Perceptual Linear Prediction (PLP). The MFCC's success has seen it become a de facto standard feature for speaker recognition. Alternative features, that convey information other than the average subband energy, have been proposed, such as frequency modulation (FM) and subband spectral centroid features. In this study, we investigate the characterization of subband energy as a two dimensional feature, comprising Spectral Centroid Magnitude (SCM) and Spectral Centroid Frequency (SCF). Empirical experiments carried out on the NIST 2001 and NIST 2006 databases using SCF, SCM and their fusion suggests that the combination of SCM and SCF are somewhat more accurate compared with conventional MFCC, and that both fuse effectively with MFCCs. We also show that frame-averaged FM features are essentially centroid features, and provide an SCF implementation that improves on the speaker recognition performance of both subband spectral centroid and FM features.

1. Introduction

Speaker recognition depends on the isolation of speaker-dependent characteristics from speech signals, and the speaker's vocal tract configuration has been recognized to be extremely speaker-dependent because of the anatomical and behavioral differences between subjects [1]. The most successful vocal tract-related acoustic feature is the Mel-frequency cepstral coefficients (MFCC). However during the MFCC extraction procedure, information related to the distribution of energy across the band is not effectively captured. For a subband speech signal MFCC carries mainly the average energy of the subband as a single dimension (the overlapped triangular filters capture some information from neighbouring bands, but this can be considered an inter-band rather than an intra-band information). In this paper, we investigate expanding this single dimensional information into two dimensional information that captures both the average energy and additional information concerning the distribution of energy within each subband.

Research reported in [2, 3, 4, 5] suggests that phase or frequency related features are potentially complementary to MFCCs. One problem with using frequency modulation (FM) extraction in practical implementations is computational complexity [6]. Recently, the effectiveness of the frame-averaged FM components extracted using second order all pole method [2] on speaker recognition and its complementary nature to magnitude based information was demonstrated [3]. A comparison between these frame-averaged FM components and the deviation of subband spectral centroid [7] from the center frequency of the subband, as shown in Figure 1, reveals that both subband spectral centroid and frame-averaged FM components carry similar information. However, estimation of subband spectral centroid is more efficient than the estimation of frame-averaged FM components.

In [7] it was shown that spectral centroid frequency carries formant-related information. It was further argued that though formant locations are robust to additive noise, formant frequencies should not be directly used as features due to the problem of accurate estimation. This problem can be overcome using other features that carry formant related information such as spectral centroid frequency, as in [7]. Spectral centroid frequency was earlier used in [7] for speech recognition and the use of subband spectral centroid in recent literature have shown some success in noisy speech recognition [8, 9]. Recently, spectral centroid frequency was also used for speaker recognition [10, 11] to complement cepstral based features with very slight success in contrast to FM features. Considering the similarity with frame-averged FM seen in Figure 1, however, the slight improvements over MFCC in speaker recognition applications seems something of an anomaly.

In this paper, we investigate the effectiveness of the combination of Spectral Centroid Frequency (SCF) and Spectral centroid Magnitude (SCM) features for speaker recognition, and demonstrate an improved implementation of subband spectral centroid. Here SCM carries the magnitude related information similar to MFCC while SCF carries the frequency bias of the SCM as shown on Figure 2. These features will be evaluated on the NIST2001 and NIST2006 speaker recognition databases.

2. Spectral centroid feature extraction

The proposed SCF and SCM are extracted according to the schematic diagram shown in Figure 3. Let $s[n]$, for $n \in [0, N-1]$, represent a frame of speech and let $S[f]$ represent the spectrum of this frame. Then, $S[f]$ is divided into K subbands, where each subband is defined by a lower frequency edge (l_k) and an upper frequency edge (u_k). The frequency-sampled fre-