INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

A Non-Uniform Filterbank for Speaker Recognition

Jia Min Karen Kua[1,2], *Tharmarajah Thiruvaran*[1], *Eliathamby Ambikairajah*[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, t.thiruvaran@unsw.edu.au, ambi@ee.unsw.edu.au

Abstract

It is known that speaker-specific information is distributed non-uniformly in the frequency domain. Current speaker recognition systems utilize auditory-motivated scales for extracting acoustic features. These scales, however, are not optimised to exploit the spectral distribution of speaker-specific information and hence may not be the optimal choice for speaker recognition. In this paper, the authors studied the distribution of speaker-specific information for Spectral Centroid Frequency feature, and a non-uniform filter bank is proposed to capture the information effectively for spectral centroid feature. The F-ratio and Kullback-Leibler (KL) distance were used to measure distribution of speaker-specific information and it was empirically shown that the KL distance is better than F-ratio in measuring discriminative ability. The proposed filterbank emphasises the high KL distance regions by allocating more filters in those regions. Experimental results showed a relative EER reduction of 8.8% over the Mel-scale filterbank on NIST2006 SRE database.

Index Terms: speaker recognition, F-ratio, Kullback-Leibler distance, Spectral centroid frequency

1. Introduction

For speaker recognition, filterbank modelling inspired by auditory findings is a widely used approach for the parameterization of a speech signal. The Mel-frequency cepstral coefficients (MFCC) are the most evident example of a feature set that is extensively used in speaker recognition. When the MFCC front-end is used in a speaker recognition system, one makes an implicit assumption that the human hearing mechanism is the optimal speaker recognizer. Recently, Lu et al. [1] showed that an auditory scale is not an optimal scale for designing a speaker identification scale based on MFCC, as it does not take into account the distribution of speaker discriminative information across the spectrum.

The psychological explanation and experimental analysis concludes that speaker information is deeply concerned with physiological and morphological differences of the speech organs [1]. Speaker physiology-dependent features extracted with their proposed non-uniform subband filters with emphasis on the frequency region of glottis and piriform fossa, achieved better speaker identification performance [1]. This is because the vibration frequency of the vocal folds in the range of 100 – 400 Hz caused by the length and stiffness of the vocal folds are speaker-dependent characteristics [1]. Next, the effect of the laryngeal tube and bilateral cavities of the piriform fossa were found to be relatively stable, regardless of vowel type, in contrast to relatively large inter-speaker variation in the frequency range of 2.5kHz [2]. In addition, the higher formants such as F4 and F5 have been found to be rather stable for continuous speech. The formant F4 is determined almost entirely by the geometry of the laryngeal cavity, which has been suggested as an organ to transmit non-linguistic information [3].

Distribution of speaker-specific information across different frequency bands was investigated by Lu et al. [1] for log energies and by Thiruvaran et al. [4] for frequency modulation (FM), and optimum filters were designed for those respective features. Recently, the effectiveness of the Spectral Centroid Frequency (SCF) for speaker recognition and its similarity to frame-averaged FM was demonstrated [5] as a complementary feature to traditional MFCC feature. This paper investigates the distribution of speaker-specific information of SCF features and designs a non-uniform filterbank to effectively capture this information. Further, a comparison between the effectiveness of KL distance and F-ratio in measuring speaker specific information is conducted. In addition, it empirically shows that to effectively capture speaker specific information, more filters need to be assigned by reducing their bandwidth in prominent speaker specific regions rather than allocating a fewer number of filters with larger bandwidth.

2. Spectral centroid frequency

Spectral centroid frequency (SCF) [6] is the weighted average frequency for a given subband, where the weights are the normalized energy of each frequency component in that subband. Since this measure captures the center of gravity of each subband, it can detect the approximate location of formants, which manifest as peaks in neighbouring subbands [6, 7]. However, the center of gravity of a subband is also affected by the harmonic structure and pitch frequencies produced by the vocal source, particularly for narrow bandwidths. Hence, the SCF feature is affected by changes in pitch and harmonic structure. The k^{th} subband spectral centroid frequency F_k is defined as follows [6]:

$$F_k = \frac{\sum_{f=l_k}^{u_k} f|S[f]\omega_k[f]|}{\sum_{f=l_k}^{u_k} |S[f]\omega_k[f]|} \qquad (1)$$

where |S[f]| represent the magnitude spectrum of a frame of speech and $\omega_k[f]$ is the frequency response of the k^{th} subband filter that is defined by a lower frequency edge (l_k) and an upper frequency edge (u_k). For all the experiments reported in this paper, SCFs are extracted with 14 Gabor filterbank. Furthermore, the number of FFT points are increased by an order of magnitude (from 160 to 2048 for f_s=8 kHz by zero-padding) to better approximate the speech power spectrum and filterbank frequency