INTERSPEECH 2009 BRIGHTON

Analysis of band structures for speaker-specific information in FM feature extraction

Tharmarajah Thiruvaran, Eliathamby Ambikairajah and Julien Epps

School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052 Australia.
National Information Communication Technology (NICTA),
Australian Technology Park, Eveleigh 1430, Australia

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

Abstract

Frequency modulation (FM) features are typically extracted using a filterbank, usually based on an auditory frequency scale, however there is psychophysical evidence to suggest that this scale may not be optimal for extracting speaker-specific information. In this paper, speaker-specific information in FM features is analyzed as a function of the filterbank structure at the feature, model and classification stages. Scatter matrix based separation measures at the feature level and Kullback-Leibler distance based measures at the model level are used to analyze the discriminative contributions of the different bands. Then a series of speaker recognition experiments are performed to study how each band of the FM feature contributes to speaker recognition. A new filter bank structure is proposed that attempts to maximize the speaker-specific information in the FM feature for telephone data. Finally, the distribution of speaker-specific information is analyzed for wideband speech.

1. Introduction

Perhaps in the majority of speech front-ends, the speech signal is decomposed into subbands, whose spacing and bandwidths are often motivated by auditory frequency scales. Despite the success of front-ends simulating the process of auditory perception, it has been shown that machine performance in speaker recognition tasks is very competitive with human performance, as far back as about a decade [1]. Here, although the machine did not *exactly* replicate human perception of speech, it is still competitive. This may be because the machine can be tailor-made specific to speaker recognition, which further implies that each part of the feature extraction process could be tested and tuned to speaker-specific information.

There are psychophysical explanations for the speaker-specific information being concentrated in particular frequency regions. The effect of the hypopharynx (laryngeal tube and the piriform fossa) was found to be relatively stable for vowel production but varied widely among different speakers [2]. This inter-speaker variation affects the frequency spectra above approximately 2.5 kHz. In addition, the higher formants such as F4 and F5 have been found to be rather stable for continuous speech [3]. F4 is determined almost entirely by the geometry of the laryngeal cavity, which has been suggested as an organ to transmit non linguistic information. Further, the fundamental frequency in the range between 50 Hz to 400Hz depends on the stiffness and length of the vocal folds[4]. Apart from these psychophysical explanations, an experimental analysis was performed using MFCC features on a speaker recognition database with 18 kHz sampling frequency [4]. The authors experimentally found that the three frequency regions from 100 Hz to 300Hz, from 4 kHz to 5.5 kHz and from 6.5 kHz to 7.8 kHz contain higher speaker-specific information. Further, the frequency region from 500Hz to 3.5 kHz was found to have less speaker discriminative information. Unfortunately the bandwidth of the major speaker recognition databases, the NIST SRE databases, lies mainly within this region. This paper studies how speaker-specific information in frequency modulation features is distributed among different bands mainly on telephone bandwidth of NIST SRE database using FM feature not only at the feature level as in [4], but also at model and classification levels. FM is selected as each dimension of FM directly corresponds to each subband [5]. Psychophysical and experimental motivation for the use of FM in speaker recognition was introduced in [5-6].

FM can characterize the phase of the signal, and its importance has emerged recently as a complementary feature to amplitude-based features [5]. Nonlinearities in speech production, together with the evidence for modulation in speech production led to an AM-FM modulation model being proposed in [7]. Here, modulation around the vocal tract resonances is explained as "the air jets flowing through the vocal tract during speech production is highly unstable and oscillates between its walls, attaching or detaching itself, and thereby changing the effective cross-sectional areas and air masses, which affects the frequency of the cavity resonator" [7]. FM feature extraction involves subband filtering and usually auditory motivated Bark or Mel scale filters are used.

This paper analyses how speaker-specific information is distributed across different bands in FM feature extraction, with a view to reallocate filter bank centre frequencies and bandwidths specifically for speaker recognition in the telephone bandwidth. Further, this paper will assess whether the auditory motivated filter bank is a good choice for speaker recognition. For reasons of computational convenience, the relatively small NIST2001 database has been used extensively and then selected experiments are conducted on NIST 2006 male database. Finally this study is extended to wideband speech sampled at 44.1 kHz using the CHAINS corpus [8].

2. FM feature extraction

Based on the evidence for the existence of modulations in speech, the resonances are each modeled as AM-FM signals in, and summed, can be used to model the speech as in (1) [7].

$$x(t) = \sum_{k=1}^{K} A_k(t) \cos\left[2\pi f_{c_k} t + 2\pi \int_0^t q_k(\tau) d\tau \right] + \theta_k \qquad (1)$$