Speaker Identification using FM Features

Tharmarajah Thiruvaran[1], Eliathamby Ambikairajah[1], Julien Epps[2]

[1]School of Electrical Engineering and Telecommunications, Faculty of Engineering,
The University of New South Wales, Sydney, Australia
[2]ATP Research Laboratory, National ICT Australia,
Sydney, Australia
thiruvaran@student.unsw.edu.au ambi@ee.unsw.edu.au julien.epps@nicta.com.au

Abstract

The AM-FM modulation model of speech is a nonlinear model that has been successfully used in several branches of speech-related research. However, the significance of the AM-FM features extracted from this model has not been fully explored in applications such as speaker identification systems. This paper shows that frequency modulation (FM) features can improve speaker identification accuracy. Due to the similarity between amplitude modulation (AM) feature and the conventional Mel frequency cepstrum coefficients (MFCC), this paper mainly focuses on the FM feature. The correlation between FM feature components is shown to be very small compared with that of Mel filterbank log energies, thus reducing the need for decorrelation. FM feature components are shown to be very nearly Gaussian distributed. Further, speech synthesis using AM-FM features is performed to compare four existing AM-FM demodulation methods based on the perceptual quality of the synthesized speech. Of these, Digital Energy Separation Algorithm (DESA) gives the best synthesized speech, and is thus used as a front-end in our speaker identification system. Evaluation of speaker identification using FM features on the NIST 2001 database shows a relative improvement in speaker identification accuracy of 2% for male speakers and 9% for female speakers over the conventional MFCC-based front-end.

1. Introduction

Speech analysis using the conventional source-filter model alone appears to provide only a certain level of performance in speech processing applications. One alternative approach is the modeling of speech resonances as AM-FM signals. In the formant AM-FM model (Maragos, Kaiser & Quatieri, 1993), each formant is represented as an AM-FM signal and the total speech is the sum of all the formants, as in equation (1).

$$s(t) = \sum_{k=1}^{K} a_k(t)\cos(\omega_{ck}t + \phi_k(t)) \quad (1)$$

$$\phi_k(t) = \omega_{dk}\int_0^t q_k(\tau)d\tau \quad (2)$$

where, $a_k(t)$ is the AM component, $\cos(\omega_{ck}t + \phi_k(t))$ is the FM modulated component, $\phi_k(t)$ is the FM component, ω_{dk} is the modulation index, K is the total number of formants and f_{ck} is the center frequency of the k^{th} formant. Several examples of evidence for the existence of modulation in speech are summarized by Maragos et al., (1993). One such cause of the modulation is the cavities in the vocal tract. In other words, the modulation characteristics depend on the physical properties of the speaker. These properties can be captured in AM-FM features, which are the demodulated AM and FM components of the speech.

In speaker identification systems, the front-end must extract features that characterize the speaker. Based on the argument relating the AM-FM features to the speaker's physical properties, we can hypothesize that the speaker identification system can be improved using these AM-FM features. In work by Jankowski, Quatieri and Reynolds (1994), AM-FM features were applied to speaker identification, but the authors found that FM features gave poor performance, and focused mainly on "Teager Energy" (Maragos et al., 1993) as a feature. They used Digital Energy Separation Algorithm (DESA) with formant tracking.

The most common method of using AM-FM features in speech processing is estimating speech formants and then extracting the AM-FM components from them. Formant estimation for AM-FM extraction can be achieved using Linear Predictive Coding (LPC) analysis (Jankowski et al., 1994) or using an iterative approach as suggested in Maragos et al., (1993). The iterative approach involves repeatedly changing the center frequency of the band pass filter to the average of the extracted instantaneous frequency. All these formant estimation techniques introduce additional complexity to the AM-FM extraction methods. A method to avoid this formant estimation in AM-FM extraction is to use the band that gives maximum normalized energy (Bovik, Maragos &

Proceedings of the 11th Australian International Conference on Speech Science & Technology, ed. Paul Warren & Catherine I. Watson. ISBN 0 9581946 2 9
University of Auckland, New Zealand. December 6-8, 2006.

Accepted after full paper review