# Evaluation of a Fused FM and Cepstral-Based Speaker Recognition System on the NIST 2008 SRE

*Mohaddeseh Nosratighods* [1,2], *Tharmarajah Thiruvaran* [1,2], *Julien Epps* [1,2], *Eliathamby Ambikairajah* [1,2], *Bin Ma* [3], *Haizhou Li* [3,1]

[1] School of Electrical Engineering and Telecommunications
The University of New South Wales, Sydney, NSW 2052, Australia
[2] National ICT Australia, Australian Technology Park, Eveleigh, NSW 2015, Australia
[3] Human Language Technology Dept, Institute for Infocomms Research, A*STAR, Singapore

hadis@unsw.edu.au, thiruvaran@student.unsw.edu.au, j.epps@unsw.edu.au, ambi@ee.unsw.edu.au, mabin@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg

## ABSTRACT

In this paper, the fusion of two speaker recognition subsystems, one based on Frequency Modulation (FM) and another on MFCC features, is reported. The motivation for their fusion was to improve the recognition accuracy across different types of channel variations, since the two features are believed to contain complementary information. It was found that the MFCC-based subsystem outperformed the FM-based subsystem on telephone conversations from NIST SRE-06 dataset, while the opposite was true for NIST SRE-08 telephone data. As a result, the FM-based subsystem performed as well as the MFCC-based subsystem and their fusion gave up to 23% relative improvement in terms of EER over the MFCC subsystem alone, when evaluated on the NIST 2008 core condition.

***Index Terms*** Speaker Recognition, Frequency Modulation, MFCC, Fusion

## 1. INTRODUCTION

The fusion of complementary subsystems has become common practice in achieving good speaker verification accuracy [1]. This motivation suggests the investigation of features that are not explicitly based on spectral magnitude information. Some phase-based features, such as frequency modulation (FM) features [2,5], have shown promise in robust speech recognition [2], and there is psychoacoustic evidence to suggest that FM components are processed differently to amplitude information in the human auditory system [3]. Recently, an improved technique for the extraction of FM components from a sub-band decomposition of the speech signal was developed, and was shown to be effective in speaker recognition [4]. In this paper, the fusion of speaker recognition sub-systems based on MFCCs and FM features extracted using this technique is proposed. Comparative evaluation of the FM- and MFCC-based subsystems on the NIST 2006 and 2008 SRE data sets demonstrates the complementary properties of the two feature sets.

## 2. FM FEATURE EXTRACTION

Modeling of the speech signal in terms of frequency modulation components in this work is based on the AM-FM model of speech signals proposed in [5] to accommodate the modulations during speech production. The AM-FM model treats each vocal tract resonance as an AM-FM signal, and models speech as the sum of all such resonances. This implies that a front-end employing FM features needs to identify the resonances (formants) from which the FM components can be extracted. The authors experimented with resonance-based FM extraction for automatic speaker recognition, and results of informal experiments were poor, due to the imperfect formant estimation. This motivated us to instead estimate FM components from the sub-bands of the speech signal. In the proposed FM sub-system, following a Bark-spaced Gabor filter bank analysis, each $k$th sub-band signal is modeled according to an AM-FM model [5]:

$$p_k[n] = a_k[n]\cos\left(\frac{2\pi f_{ck} n}{f_s} + \frac{2\pi}{f_s}\sum_{r=1}^{n} q_k[r]\right), \quad (1)$$

where $a_k[n]$ is the time-varying AM component, $q_k[r]$ is the FM component, $f_s$ is the sampling frequency and $f_{ck}$ is the center frequency of the $k$th band pass filter. To determine $q_k[r]$, we employ second-order all-pole modeling [4] to estimate the instantaneous frequency $\theta_k$ of the windowed sub band signal $p_k[n]$ as

$$\theta_k = \frac{2\pi f_{ck}}{f_s} + \frac{2\pi}{f_s} q_k[r] \ , \quad (2)$$

Practically, we estimate $\theta_k$ from the pole angle of the second-order linear predictor coefficients of the windowed sub band signal $p_k[n]$. The estimated FM component $q_k[r]$ at instant $n$ is then obtained from the estimated $\theta_k$ by rearranging (2). We use one FM estimate per frame per sub-band. This recently proposed FM extraction technique has been shown to outperform the DESA [5] and Hilbert transform-based extraction methods for speaker recognition [4]. Finally, we note that since MFCCs are derived from Mel-spaced spectral magnitudes (related to $a_k[n]$) and the FM features here are based on $\theta_k$, the two features can be considered to be complementary.