Interspeech 2008 Special Session: Forensic Speaker Recognition - Traditional and Automatic Approaches.
Brisbane, Queensland, Australia. September 2008

# FM Features for Automatic Forensic Speaker Recognition

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

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney NSW 2052 Australia.
[2]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) information from the speech signal is herein proposed to complement the conventional amplitude based features for automatic forensic speaker recognition systems. In addition to presenting the AM-FM model of speech used to generate the proposed frequency modulation features, the significance of frequency modulation for speaker recognition is discussed. Evaluation results from an automatic forensic speaker recognition system combining FM and MFCC features are shown to out-perform those of a system employing MFCC features alone, in terms of all typical metrics, such as detection error trade-off curves, Tippett curves and applied probability of error curves.

**Index Terms**: frequency modulation, automatic forensic speaker recognition.

## 1. Introduction

Expert opinion may be required about a recording from a crime scene relating to the suspect in judicial proceedings. In that situation automatic forensic speaker recognition (FSR) system can be used by a forensic scientist to produce a meaningful estimate of the strength of the evidence (information extracted from the questioned recording) in the form of Likelihood Ratio (LR). To accomplish this task, the front-end of the automatic FSR system should utilize all possible information from the speech signal. Studies of human auditory perception suggest that frequency modulation information from the speech signal is complementary to the conventional amplitude based features such as MFCC.

Psychophysical evidence for the existence of human auditory pathways tuned to frequency modulated tones was first reported in [1]. Further experiments, supporting the claim that information pertaining to changes in amplitude and information pertaining to changes in frequency are processed in separate psychophysical channels, are reported in [2]. While these findings reveal that frequency modulation plays a significant part in human perception, several sources of evidence for the existence of frequency modulation in the speech signal are reported in [3], based on vocal tract air velocity measurements conducted in [4]. Further, a model for the speech signal, incorporating this frequency modulation information in terms of an AM-FM model, was also proposed in [3] as an AM-FM model. Later, several human perception experiments showed that an FM signal, combined with amplitude information, enhanced human perception [5, 6]. Further, the FM properties of the speech signal have been successfully exploited in cochlear implant applications [7], automatic speech recognition [8] and automatic speaker recognition [9], in each case as a complement to amplitude information. Thus, there is a significant and diverse body of research to support the notion that FM components in the speech signal provide complementary information to amplitude information in features such as MFCCs.

The primary motivation for using FM features for speaker recognition is the explanation for the modulation observed in the speech signal in [3], based on Teager's experiment. In particular, "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 affected cross-sectional areas and air masses, which affects the frequency of the cavity resonator" [3]. It is the vocal tract walls that cause the modulation of the oscillating air flow, thus the modulation in speech can be expected to carry speaker-specific information. In addition to this, the initial volume and mass of the air flow entering through the vocal cord depends on the size and properties of the vocal cord, thus it can be assumed to contain speaker-specific information. Further, using cochlear implant subjects, a speaker recognition experiment [5] comparing AM only and AM+FM found that the performance is improved with FM. This experiment also showed that the FM contains speaker-specific information.

Based on the above motivations, we propose to use FM components from speech to improve the performance of an automatic FSR system. Evaluation of the FM features, MFCCs and the combination of both on the NIST 2001 cellular database shows that FM can be used to improve the performance of the automatic FSR system.

## 2. Frequency Modulation of Speech

### 2.1. AM-FM Model

Based on Teager's [4] experimental findings of the existence of modulations in speech, the resonances are each modeled as AM-FM signals in [10]. Then the total speech is taken as the sum of all the resonances as given in equation 1.

$$x[n] = \sum_{k=1}^{K} A_k[n] \cos\left[ \frac{2\pi f_{c_k} n}{f_s} + \frac{2\pi}{f_s} \sum_{r=1}^{n} q_k[r] + \theta \right], \qquad (1)$$

where $K$ is the total number of resonances, $A_k[n]$ is the time-varying amplitude component, $q_k[r]$ is the time-varying frequency component, $f_{ck}$ is the center frequency of the resonance, $f_s$ is the sampling frequency and $\theta$ is the initial phase. This model is the basis for FM extraction from speech.

### 2.2. FM Feature Extraction from Speech

In this work, long term average FM features are extracted over a 20 ms window length using the second-order all-pole method [9]. In this method, the speech is initially filtered using Bark scaled Gabor band pass filters. Then each sub band signal is approximated as the impulse response of a second order resonator, from which the pole frequency is estimated. From the pole frequency, taken as an estimate of