An Investigation of Sub-band FM Feature Extraction in Speaker Recognition

Tharmarajah Thiruvaran[1,2], **Julien Epps**[1], **Eliathamby Ambikairajah**[1,2], **Edward Jones**[3]

[1]*School of Electrical Engineering and Telecommunications,*
The University of New South Wales, Sydney
NSW 2052 Australia.
thiruvaran@student.unsw.edu.au,
j.epps@unsw.edu.au,
ambi@ee.unsw.edu.au

[2]*National Information Communication Technology (NICTA),*
Australian Technology Park, Eveleigh 1430,
Australia.
[3]*Department of Electronic Engineering*
National University of Ireland, Galway
edward.jones@nuigalway.ie

***Abstract*— Following recent evidence that FM features extracted from a sub-band decomposition of speech are highly uncorrelated, this paper investigates the effect of the number of auditory scale sub-bands in FM based front-end processing. For this study, a newly developed robust FM extraction method based on the least square differential ratio is used to extract features, comprising one FM component per sub-band. Automatic speaker recognition experiments were conducted on the cellular NIST 2001 database, with the number of filters in the front-end varied from 6 to 26. Performance degradation was observed for very low numbers of filters and very high numbers of filters. Results show that for a 4 kHz speech bandwidth, a minimum of 10 and a maximum of 18 sub-bands is a suitable choice for speech front-end applications such as automatic speaker recognition.**

***Keywords* – Frequency modulation, automatic speaker recognition, Mel scale, filter bank.**

I INTRODUCTION

Automatic speaker recognition is a biometric recognition system using the speech signal to recognise a person’s claimed identity. It is preferred in security systems to authenticate for secure access, as it is a natural signal produced by humans that can be easily accessed remotely through a phone line. However, the accuracy of the speaker recognition system with conventional amplitude based features such as Mel frequency cepstral coefficients (MFCCs) alone does not satisfy the requirements of such security applications, leading to the consideration of phase based features for front-end processing. Frequency modulation (FM) is one such phase based feature that has recently received research attention [1]. As might be expected, FM features produce a significant improvement when combined with amplitude based features [1-4].

This FM feature is extracted based on the AM-FM model of the speech signal 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. This approach was used in [6], where formants were identified using linear prediction, and a band pass filter with the same center frequency and bandwidth as the formant was used to isolate that formant. The authors experimented informally with this approach of FM extraction on automatic speaker identification, and results were poor.

Subsequent work on FM features has used a filter bank with fixed bandwidths and center frequencies to decompose the speech, from which FM components are extracted in each sub-band [1, 2, 7-9]. This fixed filter bank is preferred to formant tracking because: (i) the fixed filter bank removes band mismatch in the FM feature (ii) in most pattern recognition tasks fixed dimension features are preferred, while the number of formants (hence the feature dimension) may vary for a given speech bandwidth (iii) formant tracking itself is an imperfect process which may introduce errors in FM extraction. However, in fixed filter bank processing there is no theoretical basis from which to select the number of filters. The number of filters used in FM extraction varies widely, for example 6 [1] and 200 [9] in Mel scale, 17 in Bark scale [8] and 32 in uniform scale [9] for various automatic speech processing applications.