Extraction of FM components from speech signals using all-pole model

T. Thiruvaran, E. Ambikairajah and J. Epps

Frequency modulation has recently emerged as a promising model for characterising the phase of a speech signal. Proposed is a novel technique for extracting the frequency modulation (FM) components from the subband speech signal, using a second-order all-pole model. Evaluation of a speaker recognition system employing FM features, extracted using the proposed technique, on the NIST 2001 database reveals improvement over MFCC baseline and significant improvements over the discrete energy separation algorithm and a Hilbert transform based approach in terms of equal error rate.

Introduction: Conventionally, amplitude-based features have been used in the front-ends of speech processing systems. Since these features alone do not seem adequate for speech and speaker recognition systems, recently phase based features have received increased research attention [1, 2]. One phase based feature used in recent years is frequency modulation [1], which have shown promise in robust speech recognition. In particular, the frequency modulation feature is motivated by an AM-FM model of the speech signal, in which vocal tract resonances are modelled by AM-FM signal, based on evidence of such modulation during speech production [3]. The most popular frequency modulation extraction methods for AM-FM signals used in speech processing applications are the discrete energy separation algorithm (DESA) [3] and Hilbert transform based algorithms [4], while alternative approaches are based on an iterative Hilbert transform [5], linear prediction [6].

The main obstacle in using the above methods for speech processing applications is the previously observed variability of the FM estimates [1], which result in degraded classification accuracy when FM estimates are used as features. In this Letter, we address the problem by proposing a frequency modulation extraction technique using a second-order all-pole model that produces a considerably more consistent FM estimate.

AM-FM model of speech signal: In the AM-FM model of speech signal, the vocal tract resonances are modelled as AM-FM signals and the speech $s[n]$ is taken as the sum of all resonances [3], represented in discrete form as:

$$s[n] = \sum_{k=1}^{K} a_k[n]\cos(\phi_k[n]) \tag{1}$$

where K is the total number of resonances, $a_k[n]$ is the time-varying AM component, $\phi_k[n]$ is the phase of the kth resonance and n is the sample index. A number of bandpass filters can be used to isolate these resonances [1, 2]. The kth bandpass filter output $p_k[n]$ can be represented according to the AM-FM model [3], represented in discrete form as:

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

where $q_k[n]$ is the FM component, f_s the sampling frequency and f_{ck} the centre frequency of the kth bandpass filter.

Proposed method of FM extraction: FM component(s) are modelled in each subband using second-order all-pole resonators, which provide a simple but effective characterisation of bandpass signals. The resonator parameters are estimated using linear prediction, and the FM estimate is derived from the pole angle of the resulting all-pole model. Effectively, this assumes that the windowed subband signal can be approximated by the impulse response of the all-pole resonator, given by

$$h[n] = \frac{r^n \sin[(n+1)\theta]}{\sin\theta} \quad n \geq 0 \tag{3}$$

where $\pm\theta$ and r are the angles (digital frequency) and radius of the conjugate poles from the origin. This has been found to be a robust assumption in empirical work to date. Thus, each kth windowed subband signal $p_k[n]$ is approximated by a resonator of the form shown in (3)

$$p_k[n] \simeq \left[\frac{r_k^n}{\sin\theta_k}\right]\cos\left[(n+1)\theta_k - \frac{\pi}{2}\right] \tag{4}$$

By comparing (4) with (2), the AM component $a_k[n]$ and the total phase of the windowed subband signal can be expressed as follows:

$$a_k[n] = \frac{r_k^n}{\sin\theta_k} \tag{5}$$

$$(n+1)\theta_k - \frac{\pi}{2} = \frac{2\pi f_{ck^n}}{f_s} + \frac{2\pi}{f_s}\sum_{r=1}^{n} q_k[r] \tag{6}$$

Using (6) and calculating the difference between the successive samples as a first-order approximation to differentiation (with respect to n), we obtain an expression in terms of the pole frequencies

$$\theta_k = \frac{2\pi f_{ck}}{f_s} + \frac{2\pi}{f_s}q_k[n] \tag{7}$$

which can be interpreted as the instantaneous frequency (IF) of the windowed subband signal $p_k[n]$. Note that the notion of the instantaneous frequency of a short-term window of a non-stationary signal such as speech is only rigorous if the window length includes a full period of the signal. By rearranging (7), the FM estimate $q_k[n]$ of $p_k[n]$ is obtained as

$$q_k[n] = \theta_k\frac{f_s}{2\pi} - f_{ck} \tag{8}$$

where θ_k is the pole angle derived from the linear predictor coefficients of the windowed subband signal $p_k[n]$. The FM estimate $q_k[n]$ at the instant n is obtained under the assumption that the AM-FM model is valid over the duration of a window centred around n.

As the length of the sliding window used to estimate the resonator parameters increases, the FM estimate becomes considerably smoother, a desirable property for speaker recognition front-ends [1, 7]. This can be seen from the FM estimates from the proposed method with two different sliding window lengths (3.75 and 20 ms) shown in Fig. 1, for a 510 to 630 Hz subband speech signal with a sampling frequency of 8 kHz. Fig. 1 also provides comparison with the DESA and Hilbert transform-based methods, implemented as described in [3] and [4], respectively. Both the DESA (window length approximately 1 ms) and Hilbert-transform-based method (window length approximately 40 ms) produce substantially more variability than the proposed technique. Even when averaging was applied to the DESA FM estimate in informal experiments (to produce an effective window length of greater than 4 ms), the proposed technique was found to produce more consistent estimates for this example signal.

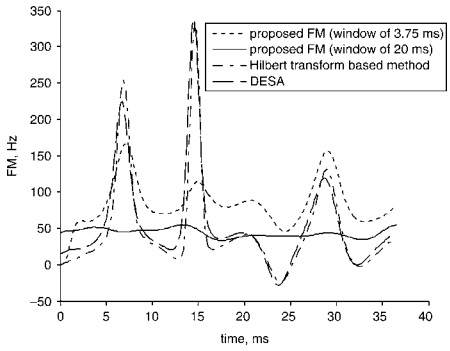

Fig. 1 *FM estimates from DESA, Hilbert transform and proposed technique for 510 to 630 Hz subband speech signal*

Evaluation: The proposed FM extraction method was evaluated in a speaker recognition system using the NIST 2001 cellular speaker recognition evaluation (SRE) database. The back-end of the system used Gaussian mixture models (GMMs) with 512 mixtures which were adapted from a Universal Background Model (UBM) using maximum *a posteriori* (MAP) adaptation. In this application, the objective is to enhance the performance of the system by combining FM information with the traditional amplitude information, in the form of Mel frequency cepstral coefficients (MFCCs).

For FM feature extraction, a continuous FM component can be extracted from each band with a sliding window shifted by one sample and an estimate of central tendency can be taken as the FM feature for that frame if desired, or the window can be advanced by a full window length to produce a single feature for each frame. The advantages of the latter approach, which is used in the first experiment

of this evaluation are: (i) computational efficiency, since only one calculation per frame is performed for each band; and (ii) larger window sizes produce smoother FM estimates that are more suitable for representation by a single value per band. The proposed FM estimation technique avoids the need for post-extraction measures of central tendency such as the intensity weighted average [8], mean [1] or median [7] to summarise the instantaneous FM information as a single value for each frame, since smoothing is effectively applied here through the use of a long analysis window.

16-Dimensional MFCCs with 16 delta coefficients were used with feature warping as channel normalisation. In FM feature extraction, 20 ms frames of speech are decomposed into subband signals using Gabor bandpass filters with centre frequencies and bandwidths spaced according to critical band specifications. As the database is cellular speech data, with a bandwidth of 0.3 to 3.4 kHz, this requires 14 critical bands. In principle it is possible to decompose the speech signal into components due to each time varying resonance [3], however in practice formant tracking approaches pose two problems for speech front-ends: (i) inaccuracies in formant frequency estimates cause problems in the resulting FM extraction; and (ii) in most pattern recognition approaches, a fixed-dimension feature vector is required, while the number of formants (and hence the feature dimension) may vary for a fixed bandwidth. These problems were presumably also anticipated by the fixed six-band Mel-spaced Gabor filter bank proposed in the AM-FM front-end of [1]. The detection error trade-off (DET) curves for the speaker recognition system with MFCC, FM and the combination of both features (fused using a linear weighting) are given in Fig. 2. The combination of MFCC and FM improves the equal error rate (EER) from 9.52 to 8.94% and the detection cost function (DCF) from 0.0379 to 0.0363 over the MFCC baseline.

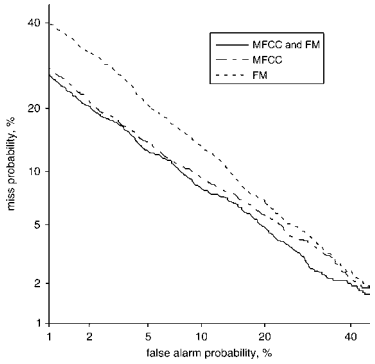

Fig. 2 *DET curves comparing speaker recognition performance of MFCC with proposed FM feature*

In the second experiment, we compared the speaker recognition performance of the proposed FM extraction technique with two other popular FM extraction methods used in speech processing: DESA and a Hilbert transform based method. For the proposed method a sliding window size of 3.75 ms, shifted by one sample, was used to estimate the FM component, consistent with the DESA and Hilbert transform approaches. The FM feature in each band was represented by the median of FM estimates over the duration of the frame [11] for all three methods. The EER results, given in Table 1, show that the system based on the proposed FM features improves by 9% compared with DESA and 4% compared with the Hilbert transform-based method. Similar improvements are seen in terms of the detection cost function (DCF).

Table 1: Comparison of FM feature extraction techniques for speaker recognition, on NIST 2001 SRE database

Feature	DCF	EER
DESA	0.0760	0.2179
Hilbert	0.0629	0.1638
Proposed method	0.0560	0.1277

Conclusion: The proposed technique for FM extraction produces more consistent instantaneous FM estimates that provide improved classification performance when used as features for speaker recognition, when compared with existing methods. Future work will include the novel application of FM based features to forensic speaker recognition.

16 January 2008
Electronics Letters online no: 20080147
doi: 10.1049/el:20080147

T. Thiruvaran, E. Ambikairajah and J. Epps (*School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney, 2052 NSW, Australia*)

E-mail: thiruvaran@student.unsw.edu.au

T. Thiruvaran and E. Ambikairajah: Also with the National Information and Communication Technology Australia (NICTA), Australian Technology Park, Eveleigh 1430, Australia

References

1 Dimitriadis, D.V., Maragos, P., and Potamianos, A.: 'Robust AM-FM features for speech recognition', *IEEE Signal Process. Lett.*, 2005, **12**, (9), pp. 621–624
2 Wang, Y., Greenberg, S., Swaminathan, J., Kumaresan, R., and Poeppel, D.: 'Comprehensive modulation representation for automatic speech recognition'. Proc. INTERSPEECH, Lisbon, Portugal, 2005, pp. 3025–3028
3 Maragos, P., Kaiser, J.F., and Quatieri, T.F.: 'Energy separation in signal modulations with application to speech analysis', *IEEE Trans. Signal Process.*, 1993, **41**, (10), pp. 3024–3051
4 Potamianos, A., and Maragos, P.: 'A comparison of the energy operator and the Hilbert transform approach to signal and speech demodulation', *Signal Process.*, 1994, **37**, (1), pp. 95–120
5 Francesco, G., Giorgio, B., Paolo, C., and Claudio, T.: 'Multicomponent AM-FM representations: an asymptotically exact approach', *IEEE Trans. Audio Speech Lang. Process.*, 2007, **15**, (3), pp. 823–837
6 Griffiths, L.: 'Rapid measurement of digital instantaneous frequency', *IEEE Trans. Acoustics Speech Signal Process.*, 1975, **23**, (2), pp. 207–222
7 Thiruvaran, T., Ambikairajah, E., and Epps, J.: 'Speaker identification using FM features'. Proc. 11th Australasian Int. Conf. on Speech Science and Technology, Auckland, New Zealand, 2006, pp. 148–152
8 Potamianos, A., and Maragos, P.: 'Time-frequency distributions for automatic speech recognition', *IEEE Trans. Speech Audio Process.*, 2001, **9**, (3), pp. 196–200