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.

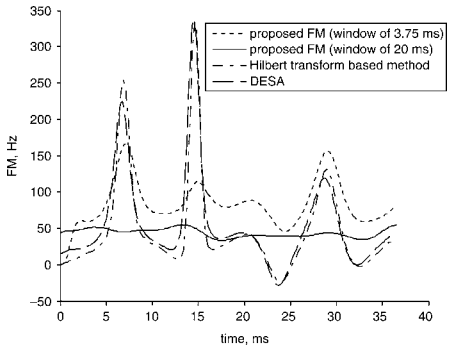

Fig. 1 *FM estimates from DESA, Hilbert transform and proposed technique for 510 to 630 Hz subband speech signal*

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