Computationally efficient frame-averaged FM feature extraction for speaker recognition

T. Thiruvaran, M. Nosratighods, E. Ambikairajah and J. Epps

Recently, subband frame-averaged frequency modulation (FM) as a complementary feature to amplitude-based features for several speech based classification problems including speaker recognition has shown promise. One problem with using FM extraction in practical implementations is computational complexity. Proposed is a computationally efficient method to estimate the frame-averaged FM component in a novel manner, using zero crossing counts and the zero crossing counts of the differentiated signal. FM components, extracted from subband speech signals using the proposed method, form a feature vector. Speaker recognition experiments conducted on the NIST 2008 telephone database show that the proposed method successfully augments mel frequency cepstrum coefficients (MFCCs) to improve performance, obtaining 17% relative reductions in equal error rates when compared with an MFCC-based system.

Introduction: Frequency modulation (FM) information has been shown to be very useful in augmenting amplitude information in several speech processing applications, such as in cochlear implant applications [1] and speaker recognition [2]. The use of FM features is motivated by AM-FM modelling of the speech signal [3]. In this model, speech is represented as a sum of resonances, each of which is represented by an AM-FM signal:

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

where 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 resonant frequency and f_s is the sampling frequency.

There are numerous FM extraction methods available, such as the digital energy separation algorithm (DESA) [3], the smoothed energy operator separation algorithm (SEOSA) [4], the FAME method [1] and methods based on the Hilbert transform [4]. There are two main problems with these methods: (i) the previously observed spiky nature of the FM estimate; (ii) the computational complexity [2]. A second-order all pole method was shown to outperform various alternative approaches, owing to its smooth estimates [2], providing a solution to (i). We propose a method as a solution for the computational complexity problem. Most FM extraction methods are computationally complex, making subband FM extraction unfeasible in real-world implementations and difficult in experiments with large databases, such as NIST evaluations. Furthermore, these methods extract sample-by-sample FM estimates, which perform poorly in speaker recognition experiments [2] when central tendency is used. To find a computationally efficient method, we exploit the fact that although the instantaneous FM estimate can be estimated, only the frame-averaged FM is usually used as a feature for classification.

Proposed frame-averaged FM feature: We propose a frame-averaged FM feature extraction method using zero crossing counts (ZC) and zero crossing counts of the differentiated signal (DZC), where the DZC enhances the robustness of the FM feature. As stated in [5], the discrimination of signals can be economically expressed through ZC and DZC, which we use in the proposed feature. As the proposed FM feature directly estimates the frame-averaged FM estimate, avoiding instantaneous estimation, it is highly computationally efficient.

For feature extraction purposes, K in (1) is taken as the number of subband filters used to analyse the speech signal, to allow a fixed feature dimension and avoid dependence on formant estimation accuracy. Then, for each subband, the frame-averaged FM estimate is used as the feature, assuming a piecewise constant FM component between two zero crossings. If the number of zero crossings over a frame length of L samples with sampling frequency f_s is N_{ZC} (number of intervals is $N_{ZC}-1$), then the frame-averaged estimate of the FM component for a subband signal of centre frequency f_c is

$$f_{ZC} = \left(\frac{N_{ZC}-1}{2L}\right) f_s - f_c \quad (2)$$

When differentiating a signal, the maxima and minima become zero crossings. Thus, frequency estimation along the lines of (2) using the differentiated signal zero-crossings (DZC) instead of ZC gives another frame-averaged FM estimate (f_{DZC}) based on the maxima and minima of the original subband signal.

Here we propose extraction of the frame-averaged FM estimate as an average of f_{ZC} and f_{DZC} for each subband signal, which is termed f_{AZC}. The motivation for averaging is to reduce feature variability for classification purposes. The accuracy of the frame-averaged FM estimation of f_{AZC} over f_{ZC} and f_{DZC} is demonstrated using the simple synthetic FM signal given in (3), of a length of 5 s:

$$s(n) = \sin\left[\frac{2\pi f_c n}{f_s} + k\cos\left(\frac{2\pi f_a n}{f_s}\right)\right] \quad (3)$$

where k is -0.4, f_a is 400 Hz and f_s is 8000 Hz. Mean squared errors (MSE) of the frame-averaged FM estimates of f_{AZC}, f_{ZC} and f_{DZC} were calculated for typical speech analysis frame sizes of 20 and 30 ms for two different centre frequencies (see Table 1). Results from this comparison show that for this signal, the averaging of f_{ZC} and f_{DZC} to produce an f_{AZC} frame-averaged FM estimate produces a substantial reduction in MSE. This is due to f_{ZC} being estimated slightly lower than the actual value, and f_{DZC} being estimated slightly higher than the actual value in these particular cases. Thus the averaging provides a more accurate estimate.

Table 1: Mean-square error comparison of frame-averaged FM estimates for signal shown in (3)

f_c	700 Hz		1100 Hz	
Frame size	20 ms	30 ms	20 ms	30 ms
MSE_ZC	4.05	1.72	4.05	1.72
MSE_DZC	1.69	0.73	1.69	0.73
MSE_AZC	0.13	0.05	0.13	0.05

The expression for the frame-averaged FM estimate using f_{ZC} and f_{DZC} is given in (4):

$$f_{AZC1} = \frac{1}{2}\left[\frac{N_{ZC}-1}{2L} + \frac{N_{DZC}-1}{2L}\right] f_s - f_c \quad (4)$$

where N_{DZC} is the number of zero crossing counts of the differentiated signal. Finer estimation of the above estimate can be achieved by considering the length between the first and the last zero crossings n_{ZC}^{first} and n_{ZC}^{last}, respectively, and the first and the last zero crossings of the differentiated signal n_{DZC}^{first} and n_{DZC}^{last}, respectively, as follows:

$$f_{AZC2} = \frac{1}{4}\left[\frac{(N_{ZC}-1)}{n_{ZC}^{last} - n_{ZC}^{first}} + \frac{(N_{DZC}-1)}{n_{DZC}^{last} - n_{DZC}^{first}}\right] f_s - f_c \quad (5)$$

Frequency estimation based on zero crossings was initially proposed in [6], where the instantaneous frequency between two zero crossings was approximated as a linear function of time. In speech processing zero crossings have been used to estimate instantaneous FM using polynomial interpolation [7] and to estimate logarithmic pseudo-instantaneous frequencies [8]. However, these techniques do not utilise the DZC and are designed to extract instantaneous estimates.

Evaluation: Speech reconstruction was performed to check the validity of our piecewise constant assumption and then speaker recognition experiments were performed. For the speech reconstruction experiment, one reconstructed signal used only instantaneous amplitude, setting the FM ($q_k[r] = 0$) component to zero in (1) as the baseline. In the second reconstruction, in addition to the instantaneous amplitude, frame-averaged FM estimates are also used according to

$$x[n] = \sum_{k=1}^{K} A_k[n] \cos\left[\frac{2\pi f_{ck} n}{f_s} + \frac{2\pi}{f_s}\sum_{r=1}^{n} q'_k[\lceil \tfrac{r}{L} \rceil]\right] \quad (6)$$

where $q'_k[\lceil \frac{r}{L} \rceil]$ is the frame-averaged FM estimate (piecewise constant), taken as either f_{AZC1} or f_{AZC2} in (4) or (5), respectively, and L is the frame size-20 ms in this experiment. Objective perceptual evaluation of speech quality (PESQ) estimates for the first and second reconstructed signals are 2.4 and 3.4, respectively. The significant improvement supports our piecewise constant assumption.

A comparative speaker recognition experiment was performed on the NIST 2001 database to examine the trade-off between error rate and