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.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI. Downloaded on February 3, 2010 at 01:13 from IEEE Xplore. Restrictions apply.

In the research area of cochlear implants, the effect of bands in AM+FM processing have been analysed through human listening tests using stimuli where the speech was reconstructed with combined AM and FM cues [3, 4]. In these human listening experiments, a monotonic increase of accuracy was observed with number of bands.

This paper investigates the effect of the number of filters used in the FM based front-end in automatic speaker recognition. The main motivation for this work is the extremely small inter-band correlation between FM features reported in [8]. This suggests the use of a larger number of more closely spaced sub-bands, and is the basis for experimental work to find an upper bound for the number of bands that can be employed in a speech front-end before the inter-band correlation becomes significant. The objective of this work is thus to find a suitable range for the number of filters, to trade off losing information by using fewer filters with introducing redundancy through larger numbers of filters, based on performance criteria. In order to find the range, automatic speaker recognition experiments were carried out on cellular NIST 2001 database, varying the number of filters in the front-end.

II FM Feature Extraction

a) FM Extraction using the Least Squares Differential Ratio (LSDR)

A newly developed FM extraction method, based on a least squares differential ratio (LSDR), is used in this work [10]. Initially, speech is passed through a bank of band pass filters and then FM components are extracted in each sub band. The instantaneous digital frequency θ is extracted using equation (1), as explained in [10].

$$\theta = 2\arcsin\left(\frac{\sqrt{a}}{2}\right) \qquad (1)$$

where

$$a = -[\mathbf{X}^T\mathbf{X}]^{-1}\mathbf{X}^T\mathbf{D} \qquad (2)$$

with vector $\mathbf{X}$ and matrix $\mathbf{D}$ defined for a sub-band signal $x(n)$ over a window length of N as in equation (3) and (4).

$$\mathbf{X} = \begin{bmatrix} x[n-1] \\ \vdots \\ x[n-N] \end{bmatrix} \qquad (3)$$

$$\mathbf{D} = \begin{bmatrix} x[n]-2x[n-1]+x[n-2] \\ \vdots \\ x[n-N+1]-2x[n-N]+x[n-N-1] \end{bmatrix} \qquad (4)$$

The FM component FM is extracted by subtracting the center frequency f_c of the band pass filter from the instantaneous frequency θ as in equation (5).

$$FM(n) = \frac{\theta}{2\pi} f_s - f_c \qquad (5)$$

b) Effect of Window Length on FM Estimate

FM estimates using equation (5) for different window lengths N are given in Figure 1, for a sub-band speech signal with a bandwidth of 2310 Hz to 2690 Hz. The figure illustrates that for longer window lengths, the FM estimate becomes more smooth because the overlap between adjacent windows is larger. This observation is utilised to form a computationally efficient feature extraction process in Section II.c.

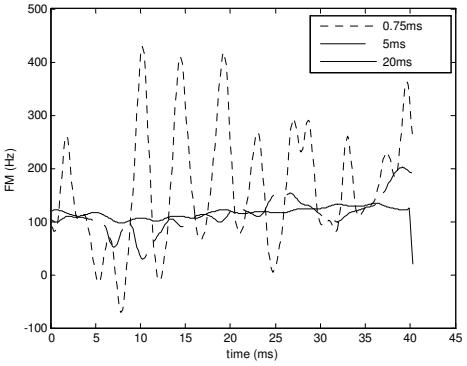

Figure 1. Instantaneous FM estimate of a sub-band speech signal for different window lengths.

c) FM Feature Extraction

In this work, each element of the FM feature represents the FM estimate in each band over a duration of 20 ms, thus the dimension of the feature vector is same as the number of sub bands. Conventionally each element in the FM feature vector is characterised for an entire frame by a measure of central tendency, such as the mean or median, applied to the instantaneous FM estimate of each band [1, 8]. To avert the need for redundant calculation of instantaneous FM estimates, and exploiting the observation in Section II.b that the FM estimate becomes more smooth for a longer window length, we calculated the FM estimate just once per 20 ms duration, with an analysis window length of 20 ms. This forms a very computationally efficient method of FM feature extraction that is also more robust than using a shorter window. The block diagram for the FM feature extraction process is given in Figure 2.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI. Downloaded on February 3, 2010 at 01:13 from IEEE Xplore. Restrictions apply.

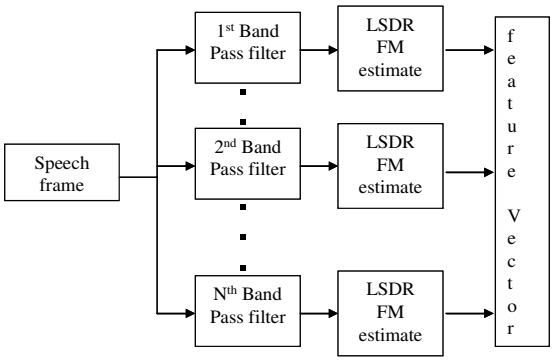

Figure 2. Block diagram for FM feature extraction using LSDR

III EXPERIMENTAL SETUP

a) Database

The database used in these experiments was the NIST 2001 cellular speaker recognition evaluation database. The training and testing phases use different handsets and channels, under different environmental conditions, either indoors, outdoors or in-car. The development database consists of 38 male speakers and 22 female speakers and the evaluation database comprises 74 male speakers and 100 female speakers for training and 850 male test segments and 1188 female test segments for testing. The training time for each speaker is 2 minutes and the test segment duration is less than 60 seconds.

b) Performance Measures

NIST evaluations use the detection cost function (DCF) as one of their primary performance measures. The DCF is defined in [11] as a weighted sum of the miss and false alarm probabilities. In addition to this, the equal error rate (EER), was employed. Smaller values of DCF and EER represent a more accurate speaker recognition system.

c) Front-end Filter Bank Configuration

Gabor filters were used for the design of filter bank because of their time-frequency compactness and good side lobe suppression [5]. Center frequencies and bandwidths were chosen to be spaced uniformly in Mel scale. The number of sub-bands used in the front-end of the speaker recognition system was varied from 6 to 26 in steps of 2, and in each configuration, separate experiments were conducted. No dimensionality reduction or DCT were employed, so the feature dimension thus also varied from 6 to 26.

d) Speaker Recognition System

The back-end of the recognition system was based on statistical modeling with Gaussian Mixture Models (GMMs). Initially two gender-dependent, 512-mixture universal background models (UBMs) [12] were trained on the NIST 2001 development set. The training data from the evaluation set was then used to adapt the speaker models from the UBM.

IV EXPERIMENTAL RESULTS AND DISCUSSION

a) Performance of the Speaker Recognition System

The equal error rate (EER) and detection cost function (DCF) of the system are given in Figure 3, graphed against the number of filters in the filter bank. These performance measures were obtained with combined male and female scores according to the NIST evaluation standard protocols.

The performance significantly improves from 6 to 10 filters and then significantly degrades after 18 filters. For 10 to 18 filters, the performance does not differ very much. The overall best performance is obtained with 16 bands.

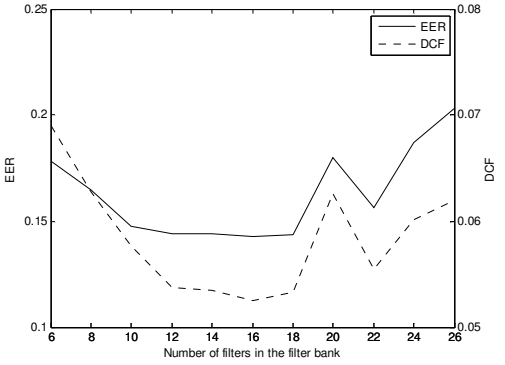

Figure 3. System performance (EER and DCF) for the entire NIST 2001 Speaker Recognition Evaluation database

In addition to the above results, we provide the system performance for male and female speakers separately in Figure 4 and Figure 5 respectively in order to gain more insight into the trend of the performance against the number of bands.

The performance of female speakers is similar to the combined performance. This is because in the NIST database, the number of female speakers is higher than the number of male speakers. For male speakers the performance degradation begins from 22 filters. However, in both cases the trend of the performance is similar such that degradations are observed for fewer than around 10 filters and for more than around 18 filters. Further, the degradation after 18 filters for female speakers is more significant than the slight improvement from 18 to

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI. Downloaded on February 3, 2010 at 01:13 from IEEE Xplore. Restrictions apply.

22 filters for male speakers. Therefore, it can be empirically concluded that the suitable range of the number of filters as 10 to 18 for automatic speaker recognition.

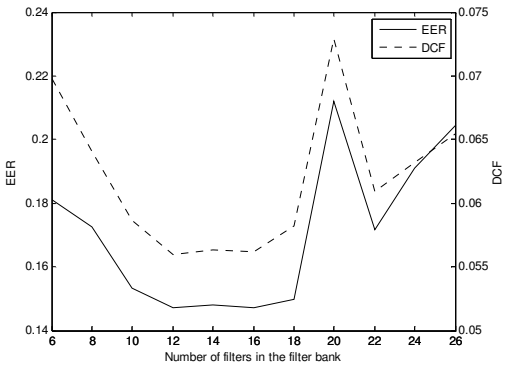

Figure 4. System performance (EER and DCF) for the female speaker subset of the NIST 2001 Speaker Recognition Evaluation database

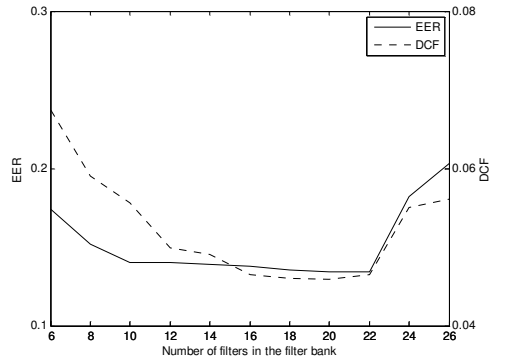

Figure 5. System performance (EER and DCF) for the subset of male speakers of the NIST 2001 Speaker Recognition Evaluation database

b) Discussion of EER and DCF Results

The degraded performance for low numbers of bands can be attributed to the fact that the FM feature contains less information. Further, for the degraded performance for higher numbers of filters could be caused by three reasons. Firstly, the bandwidth of each filter may not allow adequate spectral content (particularly at bands with center frequencies higher than the third formant frequency) to extract a meaningful FM component because when the number of bands increases, the bandwidth of each individual band decreases. The second reason may be that when the number of bands increases, the FM feature contains more redundancy due to increased correlation between adjacent bands of FM. This hypothesis is tested in section IV.c. A third possible reason is that for higher numbers of bands, the number of mixtures of GMM may become the limiting factor. That is, without increasing the number of mixtures of GMM, possibly the extra information in the higher dimension FM features cannot be modelled accurately enough. This hypothesis is tested in section IV.d.

c) Correlation Analysis

As an attempt to explain the results in terms of the correlation between the adjacent bands of the FM, these correlations were calculated separately for male and female speakers, and then averaged over all male and female speakers respectively in the training database. The correlation coefficient ρ between the k^{th} and its adjacent band was calculated as in equation (6).

$$\rho_{k,k+1} = \frac{E\{(X_k - \mu_k)(X_{k+1} - \mu_{k+1})\}}{\sigma_k \sigma_{k+1}} \qquad (6)$$

where X_k is the vector containing the k^{th} band FM feature of all frames , μ_k is the mean and σ_k is the variance of the X_k.

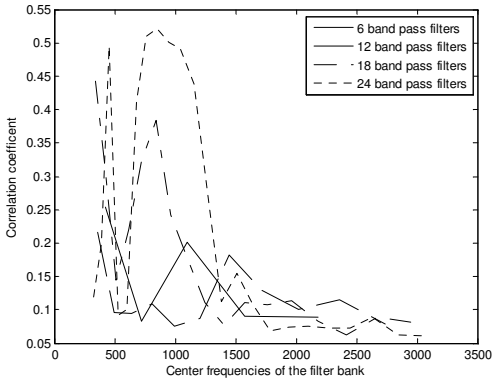

Figure 6. Correlation coefficients of adjacent bands of FM feature averaged over all female speakers.

The variations of the average correlation coefficient against the center frequency of band k is given in Figure 6 for female speakers. Similar variations are observed for male speakers as well. As expected, when the number of bands increases, the correlation between the adjacent bands also increases. Thus, this increase in correlation introduces redundancy in the feature vector resulting in a degradation of performance for higher numbers of filters. A possible reason for higher correlation in lower bands and lower correlation in higher bands is the increase in leakage from the adjacent bands. In Mel scale the bandwidth of the filters increases with the center frequency and in our filter design the number of coefficients are kept same for all the filters.

d) Effect of the Number of Mixtures in GMM

In order to test whether the number of mixtures in the GMM was the key limiting factor for the performance degradation of higher bands, another

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI. Downloaded on February 3, 2010 at 01:13 from IEEE Xplore. Restrictions apply.

experiment was performed with the number of GMM mixtures increased to 1024 from 512. This experiment was performed only for the 26-band FM feature. The performance, in terms of DCF and EER, is given in Table 1. The performance improved with an increased number of mixtures, however the improvement is not very significant compared with the performance degradation for a larger number of bands seen in figures 3, 4 and 5. Though the number of mixtures can be further increased, the results strongly suggest that there is an upper bound (in this work it is 22) for the FM features.

Table 1. Performance comparison with different GMM mixtures for the entire speakers in NIST 2001 database and the subset of male and female database.

	Entire speakers	Male speakers	Female speakers
EER (512-GMM)	0.2031	0.2034	0.2037
EER (1024-GMM)	0.2017	0.2012	0.2045
DCF (512-GMM)	0.1784	0.17	0.184
DCF (1024-GMM)	0.1771	0.1685	0.1839

V CONCLUSION

The effect on speaker recognition equal error rate of the number of sub-bands in an FM based speech front-end has been analyzed for an automatic speaker recognition system. For a 4 kHz speech bandwidth, a number of sub-bands from 10 to 18 was empirically found suitable for automatic speaker recognition purposes, with the upper bound appearing dependent upon the correlation between FM estimates in adjacent sub-bands. The performance degradation for larger numbers of filters in this automatic speaker recognition experiment, which can be attributed to the increased correlation among FM feature, contrasts with the monotonic performance improvement with number of bands in human listening tests reported in [4]. Future research directions include using other auditory scales as an alternative to the Mel scale, or even other empirically determined scales to achieve (on average at least) a relatively constant correlation between adjacent bands as a function of the band number.

REFERENCES

[1] D. V. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM Features for Speech Recognition," *IEEE Signal Processing Letters,* vol. 12, pp. 621-624, 2005.

[2] Y. Wang, S. Greenberg, J. Swaminathan, R. Kumaresan, and D. Poeppel, "Comprehensive modulation representation for automatic speech recognition," in *INTERSPEECH*, Lisbon, Portugal, 2005, pp. 3025-3028.

[3] G. S. Stickney, K. Nie, and F.-G. Zeng, "Contribution of frequency modulation to speech recognition in noise," *Journal of the Acoustical Society of America,* vol. 118, pp. 2412-2420, 2005.

[4] F. G. Zeng, K. Nie, G. S. Stickney, Y. Y. Kong, M. Vongphoe, A. Bhargave, W. Chaogang, and C. Keli, "Speech recognition with amplitude and frequency modulations," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 102, pp. 2293-8, 2005.

[5] P. Maragos, J. F. Kaiser, and T. F. Quatieri, "Energy separation in signal modulations with application to speech analysis," *IEEE Transactions on Signal Processing,* vol. 41, pp. 3024-3051, 1993.

[6] C. R. Jankowski Jr, T. F. Quatieri, and D. A. Reynolds, "Formant AM-FM for speaker identification," in *Proceedings of IEEE International Symposium on Time-Frequency and Time-Scale Analysis*, 1994, pp. 608-611.

[7] K. K. Paliwal and B. S. Atal, "Frequency-related representation of speech," in *EUROSPEECH-2003*, 2003, pp. 65-68.

[8] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Speaker Identification using FM Features.," in *Proceedings of Eleventh Australasian International Conference on Speech Science and Technology*, 2006, pp. 148-152.

[9] R. Kumaresan and W. Yadong, "On representing signals using only timing information," *Journal of the Acoustical Society of America,* vol. 110, pp. 2421-39, 2001.

[10] J. Epps, E. Ambikairajah, and T. Thiruvaran, "Robust FM demodulation of discrete-time signals using least squares differential ratio," *Electronics Letters,* vol. 43, pp. 727-729, 2007.

[11] "The NIST Year 2001 Speaker Recognition Evaluation Plan," *http://www.nist.gov/speech/tests/spk/2001/doc/,* 2001.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing,* vol. 10, pp. 19-41, 2000.

Authorized licensed use limited to: INDIAN INSTITUTE OF TECHNOLOGY GUWAHATI. Downloaded on February 3, 2010 at 01:13 from IEEE Xplore. Restrictions apply.