Odyssey 2010
The Speaker and Language Recognition Workshop
28 June – 1 July 2010, Brno, Czech Republic

Investigation of Spectral Centroid Magnitude and Frequency for Speaker Recognition

Jia Min Karen Kua[1,2], Tharmarajah Thiruvaran[1], Mohaddeseh Nosratighods[1]
Eliathamby Ambikairajah[1,2], Julien Epps[1,2]

[1]School of Electrical Engineering and Telecommunications,
The University of New South Wales, Sydney, NSW 2052, Australia
[2]ATP Research Laboratory, National ICT Australia (NICTA), Eveleigh 2015, Australia

jmkua@student.unsw.edu.au, thiruvaran@student.unsw.edu.au, hadis@unsw.edu.au,
ambi@ee.unsw.edu.au, j.epps@unsw.edu.au

Abstract

Most conventional features used in speaker recognition are based on spectral envelope characterizations such as Mel-scale filterbank cepstrum coefficients (MFCC), Linear Prediction Cepstrum Coefficient (LPCC) and Perceptual Linear Prediction (PLP). The MFCC's success has seen it become a de facto standard feature for speaker recognition. Alternative features, that convey information other than the average subband energy, have been proposed, such as frequency modulation (FM) and subband spectral centroid features. In this study, we investigate the characterization of subband energy as a two dimensional feature, comprising Spectral Centroid Magnitude (SCM) and Spectral Centroid Frequency (SCF). Empirical experiments carried out on the NIST 2001 and NIST 2006 databases using SCF, SCM and their fusion suggests that the combination of SCM and SCF are somewhat more accurate compared with conventional MFCC, and that both fuse effectively with MFCCs. We also show that frame-averaged FM features are essentially centroid features, and provide an SCF implementation that improves on the speaker recognition performance of both subband spectral centroid and FM features.

1. Introduction

Speaker recognition depends on the isolation of speaker-dependent characteristics from speech signals, and the speaker's vocal tract configuration has been recognized to be extremely speaker-dependent because of the anatomical and behavioral differences between subjects [1]. The most successful vocal tract-related acoustic feature is the Mel-frequency cepstral coefficients (MFCC). However during the MFCC extraction procedure, information related to the distribution of energy across the band is not effectively captured. For a subband speech signal MFCC carries mainly the average energy of the subband as a single dimension (the overlapped triangular filters capture some information from neighbouring bands, but this can be considered an inter-band rather than an intra-band information). In this paper, we investigate expanding this single dimensional information into two dimensional information that captures both the average energy and additional information concerning the distribution of energy within each subband.

Research reported in [2, 3, 4, 5] suggests that phase or frequency related features are potentially complementary to MFCCs. One problem with using frequency modulation (FM) extraction in practical implementations is computational complexity [6]. Recently, the effectiveness of the frame-averaged FM components extracted using second order all pole method [2] on speaker recognition and its complementary nature to magnitude based information was demonstrated [3]. A comparison between these frame-averaged FM components and the deviation of subband spectral centroid [7] from the center frequency of the subband, as shown in Figure 1, reveals that both subband spectral centroid and frame-averaged FM components carry similar information. However, estimation of subband spectral centroid is more efficient than the estimation of frame-averaged FM components.

In [7] it was shown that spectral centroid frequency carries formant-related information. It was further argued that though formant locations are robust to additive noise, formant frequencies should not be directly used as features due to the problem of accurate estimation. This problem can be overcome using other features that carry formant related information such as spectral centroid frequency, as in [7]. Spectral centroid frequency was earlier used in [7] for speech recognition and the use of subband spectral centroid in recent literature have shown some success in noisy speech recognition [8, 9]. Recently, spectral centroid frequency was also used for speaker recognition [10, 11] to complement cepstral based features with very slight success in contrast to FM features. Considering the similarity with frame-averged FM seen in Figure 1, however, the slight improvements over MFCC in speaker recognition applications seems something of an anomaly.

In this paper, we investigate the effectiveness of the combination of Spectral Centroid Frequency (SCF) and Spectral centroid Magnitude (SCM) features for speaker recognition, and demonstrate an improved implementation of subband spectral centroid. Here SCM carries the magnitude related information similar to MFCC while SCF carries the frequency bias of the SCM as shown on Figure 2. These features will be evaluated on the NIST2001 and NIST2006 speaker recognition databases.

2. Spectral centroid feature extraction

The proposed SCF and SCM are extracted according to the schematic diagram shown in Figure 3. Let $s[n]$, for $n \in [0, N-1]$, represent a frame of speech and let $S[f]$ represent the spectrum of this frame. Then, $S[f]$ is divided into K subbands, where each subband is defined by a lower frequency edge (l_k) and an upper frequency edge (u_k). The frequency-sampled fre-

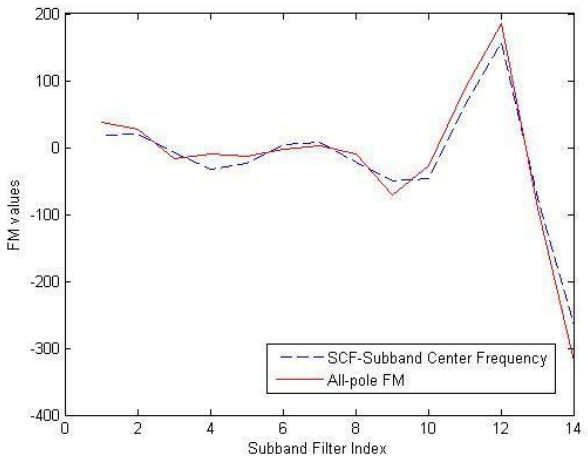

Figure 1: *Frame-averaged Frequency Modulation, based on the all-pole method [2], compared with spectral centroid frequency for a frame of voiced speech signal*

quency response of the filter is $w_k[f]$.

2.1. Spectral centroid frequency

Spectral centroid frequency (SCF) is the weighted average frequency for a given subband, where the weights are the normalized energy of each frequency component in that subband. Since this measure captures the center of gravity of each subband, it can detect the approximate location of formants, which are manifested as peaks in neighbouring subband [1, 7]. However, the center of gravity of a subband is also affected by the harmonic structure and pitch frequencies produced by the vocal source particularly for narrow bandwidths. Hence, the SCF feature is affected by changes in pitch and harmonic structure. The kth subband spectral centroid frequency F_k is defined as follows [7]:

$$F_k = \frac{\sum_{f=l_k}^{u_k} f\,|S\,[f]\,w_k[f]|}{\sum_{f=l_k}^{u_k} |S\,[f]\,w_k[f]|} \tag{1}$$

Spectral centroid frequency is commonly known as subband spectral centroid [7, 10], however, we use the term spectral centroid frequency in order to avoid the ambiguity with spectral centroid magnitude, proposed herein.

2.2. Implementation of spectral centroid frequency

In the preliminary experiments, subband spectral centroid features based on mel-scaled triangular filters as proposed in [7] did not outperform second order all pole FM [2], achieving an EER around 2% poorer than FM. Since SCF is a frequency-based feature, we experimented with extracting SCF using Bark scale Gabor filterbank which is motivated by the extraction of second order all-pole FM [2]. In addition, we increased the number of FFT points by an order of magnitude (from 160 to 2048 for f_s = 8 kHz by zero-padding) to better approximate the speech power spectrum and filterbank frequency response, which was found to have a significant effect on the SCF performance.

2.3. Spectral centroid magnitude

Spectral centroid magnitude (SCM) is the weighted average magnitude for a given subband, where the weights are the frequency of each magnitude component in that subband as computed in equation (2). SCM captures, to a first order approximation, the distribution of energy in a subband, as shown in Figure 4, for two arbitrary signals with the same average energy. Due to the weighting function, the two signals would each be represented by different SCF and SCM values. The different steepness of the weighting function with respect to the subband bandwidth may also be noted; this results in different feature element variances. Average energy could be computed using equation (2) by simply setting $f = 1$. As the spectral centroid magnitude is the magnitude at the position of the spectral centroid frequency, it will carry formant related information which is useful for speaker recognition.

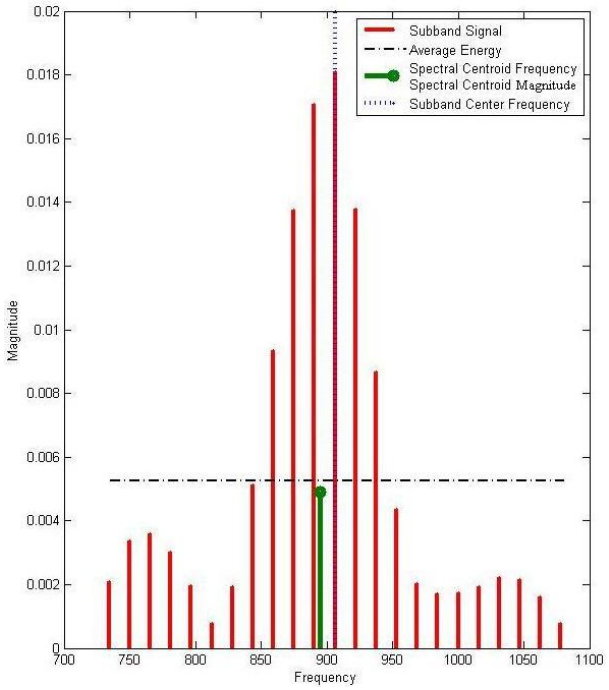

Figure 2: *Subband signal, average energy, spectral centroid frequency and spectral centroid magnitude for a subband of center frequency 906Hz*

$$M_k = \frac{\sum_{f=l_k}^{u_k} f\,|S\,[f]\,w_k[f]|}{\sum_{f=l_k}^{u_k} f} \tag{2}$$

In equation (2), the denominator is not speaker dependent, unlike for the SCF. In order to increase the speaker dependency of the SCM, an alternative formulation, using only the P most significant frequency components within each subband can be proposed, as follows:

$$M_{sc,k} = \frac{\sum_{f_i' \in I_k} f_i' \left|S\left[f_i'\right] w_k[f_i']\right|}{\sum_{f_i' \in I_k} f_i'} \tag{3}$$

where I_k is a set of frequencies corresponding to the P largest values of $|S[f]w_k[f]|$. We refer to this alternative method of SCM as SCM based on significant components (SCM_SC). As shown in Figure 5, when SCM is plotted against SCF, it provides a better approximation to the LPC spectrum compared with average energy plotted against the center frequency of each subband. To confirm this result, the average MSEs of average energy, SCM and SCM_SC of 100 speakers

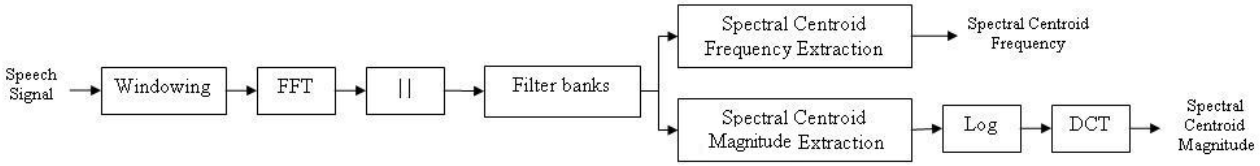

Figure 3: *Proposed feature extraction scheme*

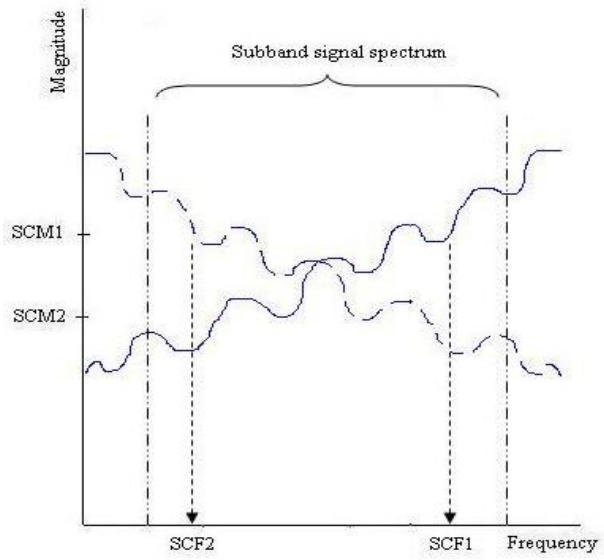

Figure 4: *SCF and SCM extraction for two different example subband signals (solid (1) and dashed (2)) with equal average energy. Due to the SCM frequency weighting, SCM1>SCM2.*

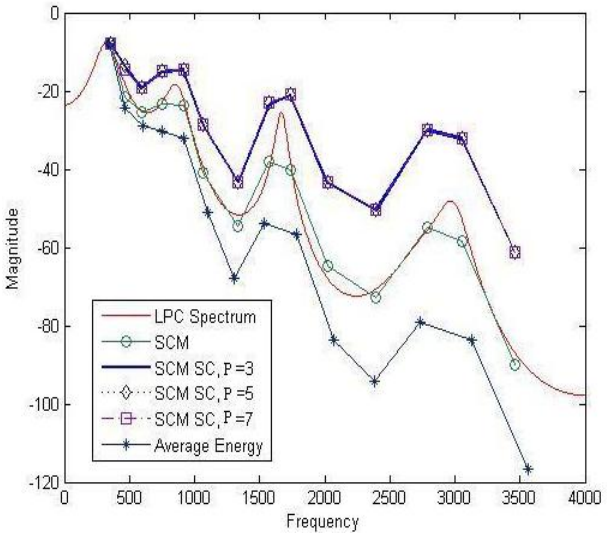

Figure 5: *LPC spectrum, SCM vs SCF and Average energy vs subband center frequency*

(50 male and 50 female from NIST2001 SRE) were computed against the LPC spectrum. The resulting MSEs were 3.17 for the average energy and 3.13 for SCM. The MSEs of SCM_SC for P=3, 5 and 7 were 6.38.

3. Experiment

3.1. Database

Speaker recognition experiments were conducted using the NIST 2001 SRE database and core condition of the NIST 2006 SRE database (1conv4w-1conv4w). Due to the extensive nature of the initial investigative experiments, the NIST 2001 SRE database was used for the intial experimental analysis. Finally the NIST 2006 SRE database was used to evaluate a selection of feature combination.

The NIST 2001 SRE development database consists of 38 male speakers and 22 female speakers. The evaluation database comprises 74 male speakers and 100 female speakers for training, 850 male speakers and 1188 female speakers for testing. The training time for each speaker was 2 minutes and the testing segment duration was less than 60 seconds.

The final evaluation data is the core test condition (1conv4w-1conv4w) of the NIST 2006 SRE where 51 448 trials are tested, which includes 3612 true trials and 47 836 false trials. The background data consists of 3079 speech utterances from the NIST 2004 SRE, which cover a number of speakers (female and male). The Nuisance Attribute Projection (NAP) [12] training data includes approximately 10000 speech utterances from the NIST 2004 and 2005 SRE corpus. The training data in the NIST 2004 SRE corpus and NIST 2005 SRE corpus are used for training cohort models in ZNorm and Tnorm score normalization [13] respectively.

3.2. Speaker verification system

The front-end of the recognition system includes an energy-based speech detector which is applied to discard silence and noise frames and delta coefficients are appended, as dynamic features, to the static features.

The back-end of the recognition system for the NIST 2001 SRE database was based on Universal Background Model - Gaussian Mixture Models (UBM-GMMs) for simplicity due to extensive nature of initial investigation experiments. Initially, two gender-dependent UBMs were created with 512 GMMs. For UBM creation, the development set of the NIST 2001 SRE database was used. Then the training data from the evaluation set was used to adapt speaker models from the UBM. Finally, the system was tested with the testing data of the evaluation set, by detecting the target speaker as the model having the maximum likelihood for the given test segment.

The back-end of the final evaluation on the 1con4w-1con4w database was based on the GMM-SVM technique. This system used GMM supervectors to construct kernels of support vector machines (SVMs). Given a speaker's speech data, a speaker model is estimated by using MAP adaptation on the means of the UBM. The means of mixture components in the speaker model are then concatenated to form a GMM supervec-

Table 1: *The speaker recognition results for spectral centroid frequency with various normalisation approaches, on the NIST 2001 SRE database*

Normalisation techniques	EER (%)
No normalisation	12.17
CMS	10.11
Feature warping	9.47

tor, which is used as an SVM kernel.

4. Results

4.1. Comparison of normalization

Usually feature warping is used as a feature normalization for magnitude based features. As SCF is a frequency based feature we empirically studied the behaviour of the feature with different feature normalization techniques. In these experiments 14 uniformly spaced Gabor filter banks across the bandwidth of 0.3 to 3.4 kHz were chosen to analyse the cellular telephone speech data. Table 1 shows the EERs from speaker recognition experiments on the NIST 2001 SRE database with several normalization techinques.

These experiments verify that the feature warping is the best normalization technique for SCF. Hereafter in all subsequent experiments, feature warping was used as the feature normalization for both SCF and SCM.

4.2. Comparison of SCF and FM

After finalizing the normalization techniques in the previous section, we investigate the effects of different frequency scales and filter banks on SCF. In these comparative experiments three different frequency scales: Bark, uniform and mel scales, and two different filter shapes: Gabor and triangular were chosen. Uniform-scaled and Bark-scaled Gabor were chosen for comparative studies between SCF and FM, since SCF and FM carry similar information as discussed in Section 1, and also because of the significant improvement of the uniform scale over Bark scale observed for FM features in [14]. In all these experiments, the number of filters was fixed at 14. Results for speaker recognition experiments based on the NIST 2001 SRE database SCF are given in Table 2.

According to the results, the Gabor filterbank with a mel scale produced the best results for SCF, outperforming the mel-scale triangular filterbank SCF implementation proposed in [7]. One reason might be that SCF is a frequency based feature and previously for another frequency based feature, FM feature, the Gabor filter bank was chosen for its optimum time, frequency sensitivity and the absence of large side lobes [15]. Further, Bark scale filters performed slightly better than uniform scale filters for SCF in contrast to the results in [14] for FM features on NIST 2001 database.

4.3. Comparison of filterbanks for SCM

The same filterbank configuration as mentioned in Section 4.2 were used for SCM extraction. Results for speaker recognition experiments based on the NIST 2001 SRE database for SCM are given in Table 3. For SCM, mel scale triangular filters performed the best among our comparisons. This result is perhaps expected since MFCCs also employ triangular mel scale filters, for which SCM is equivalent to a frequency-weighted MFCC feature.

Table 2: *The speaker recognition results for spectral centroid frequency on the NIST 2001 SRE database*

Filterbank		SCF EER (%)	FM EER [14] (%)
Mel Scale	Gabor	8.83	-
	Triangular	11.19 [7]	-
Bark Scale Gabor		9.42	12.71
Uniform Scale Gabor		12.17	10.45

Table 3: *The speaker recognition results for spectral centroid magnitude on the NIST 2001 SRE database*

Filterbank		SCM EER (%)
Mel Scale	Gabor	9.12
	Triangular	8.88
Bark Scale Gabor		9.53
Uniform Scale Gabor		9.62

4.4. Combination of SCM and SCF

In this section, we investigate the effectiveness of the combination of SCF and SCM features for speaker recognition. First, SCM and SCF were combined using score level fusion with results as given in Table 4. Linear fusion was used, with weights calculated using the same NIST 2001 database. The fusion can thus be considered optimum. When the filter banks were fixed to the same shape and scale, the best fused results were obtained with mel scale triangular fitlers. Keeping the same filterbanks for both SCF and SCM is preferred as it reduces the computational complexity significantly. It could be observed from equations (1) and (2) that only the denominator of equation (1) and (2) differs when using the same filter banks for both SCM and SCF. This is one advantage of using the combination of the (FFT-based) SCM and SCF over the alternative feature combination of MFCC and FM, where FM extraction occurs in the time domain and is very computationally demanding. System performance was further improved by fusing the best SCF and best SCM features as shown on Table 4.

Though score level fusion is usually used to combine different subsystems, in our case as both features are extracted using the same filterbanks, feature level concatenation is a reasonable alternative. The advantage of feature level concatenation over score level fusion is that a development database is not required, while score level fusion is biased by the choice of development data for computing the fusion weights. Although the performance of concatenation is slightly less than that of fusion, it should be noted that the fusion is the optimum fusion trained using the same evaluation database.

It can be observed that both fused and concatenated SCF + SCM systems perform better than the MFCC system (EER = 8.49%) [11], with an increment in the feature dimension from 32 (16 MFCCs + 16 Δs) to 56 (14 SCFs + 14 Δs + 14 SCMs + 14 Δs).

Table 4: *Fused and Concatenated EER for speaker recognition on the NIST 2001 SRE database*

Features	Fused EER (%)
SCF (Mel Scale Gabor) + SCM (Mel Scale Gabor)	8.05
SCF (Bark Scale Gabor) + SCM (Bark Scale Gabor)	8.43
SCF (Mel Scale Triangular) + SCM (Mel Scale Triangular)	7.99
SCF (Mel Scale Gabor) + SCM (Mel Scale Triangular)	7.90

Features	EER (%)
Concatenation of SCF (Mel Scale Gabor) and SCM (Mel Scale Gabor)	8.19

Table 5: *EER of SCM_SC for speaker recognition on the NIST 2001 SRE database*

Features	N	EER
SCM	-	8.88
SCM_SC	3	9.57
SCM_SC	5	9.51
SCM_SC	7	9.08
SCM_SC	9	9.19
SCM_SC	11	12.22

4.5. SCM based on significant components

In this section, the alternative expression of equation (2) to calculate SCM based on significant components is briefly explored. As shown on Table 5, the EER for SCM based on significant components did not outperform SCM for which all frequency components are taken into consideration. However, the performance of SCM_SC is close to that of SCM even when we use just a few frequency components.

4.6. SCF and SCM performance for 1conv4w-1conv4w

Finally, the combination of SCF and SCM were evaluated using the larger and more contemporary NIST 2006 database, in order to see the database independency of the results. Based on the results in Section 4.4, we selected mel scale triangular filters SCM and SCF to compare with MFCC since this combination gave the best results when they are extracted using the same filters. Further, SCF performed better when mel scale Gabor filter was used. So these filter banks were used to extract SCF and SCM when evaluating on NIST 2006 database. The performance of SCM and SCF when used alone is given in Table 5 together with the MFCC baseline, and their fusion results are given in Table 6.

It can be observed that SCF extracted using the Gabor filter performed significantly better than SCF extracted using triangular filter as proposed in [7] or all-pole based FM [14]. In addition, triangular filter extracted SCF performs worse than all-pole based FM as mentioned in Section 2.2.

Interestingly the fusion of MFCC with SCM extracted using mel scale triangular filters improved the individual subsystems.

Table 6: *Speaker recognition results for spectral centroid features on the NIST 2006 SRE database*

	Features EER				
System	Mel Scale Triangular SCM	Mel Scale Triangular SCF	Mel Scale Gabor SCF	MFCC	FM
Baseline	7.58	10.04	10.55	7.13	12.26
Baseline + NAP	5.92	9.36	7.14	5.78	7.93
Baseline + NAP + ZNorm	5.36	9.11	6.42	5.15	7.23
Baseline + NAP + TNorm	5.90	9.66	6.98	5.73	7.81
Baseline + NAP + ZTNorm	5.40	9.23	6.45	5.09	7.01

This could be attributed partly to the different number of filters used for MFCC (26 filters) and SCM (14 filters) which 'partitions' the acoustic space in a slightly different way, and partly to the different extraction methods that is, MFCC is based on average energy while SCM is based on weighted average energy.

Results from this experiment showed that the improvements discussed in Section 4.4, fusion of SCM + SCF outperforms MFCC, were also found for the more contemporary NIST2006 database, where SCM and SCF improved on a 5.09% EER MFCC baseline to 4.4% after fusion as shown in Table 6 and 7. When SCM and SCF is further fused with MFCC, the EER dropped to 3.73% (Baseline + NAP + ZTNorm) as shown in Figure 6. These results provide strong encouragement that SCM and SCF carries complementary information to MFCCs.

5. Conclusion

In this paper, we have proposed an alternative centroid feature extraction method to extract subband magnitude-based and frequency-based features from the speech spectrum. Evaluation on the NIST 2006 database using a fusion of SCM-based and SCF-based subsystems, demonstrated relative improvements of 13% over the performance of an MFCC-only system. This strongly supports the hypothesis that the combination of SCM and SCF carries more information than MFCC alone. SCF was also shown to perform significantly better than the previously proposed subband spectral centroid and frame-averaged FM features for speaker recognition. For future study, other methods of characterizing the distribution of energy within a subband and the usage of the proposed features under adverse conditions will be explored.

6. References

[1] D. Hosseinzadeh and S. Krishnan, "On the use of complementary spectral features for speaker recognition," *Eurasip Journal on Advances in Signal Processing*, vol. 2008, pp. 258184, 2008.

Table 7: *Fused speaker recognition results for spectral centroid features on the NIST 2006 SRE database*

System	Fusion EER					
	Mel Scale Triangular SCM + Mel Scale Triangular SCF	Mel Scale Triangular SCM + Mel Scale Gabor SCF	MFCC + Mel Scale Gabor SCF	MFCC + Mel Scale Triangular SCM	MFCC + FM	MFCC + Mel Scale Triangular SCM + Mel Scale Gabor SCF
Baseline	7.41	7.26	7.13	6.89	7.13	6.86
Baseline + NAP	5.46	4.92	5.14	4.92	5.49	4.66
Baseline + NAP + ZNorm	5.24	4.85	4.87	4.86	5.09	4.43
Baseline + NAP + TNorm	4.73	4.41	4.41	4.31	4.59	3.79
Baseline + NAP + ZTNorm	4.82	4.40	4.31	4.26	4.57	3.73

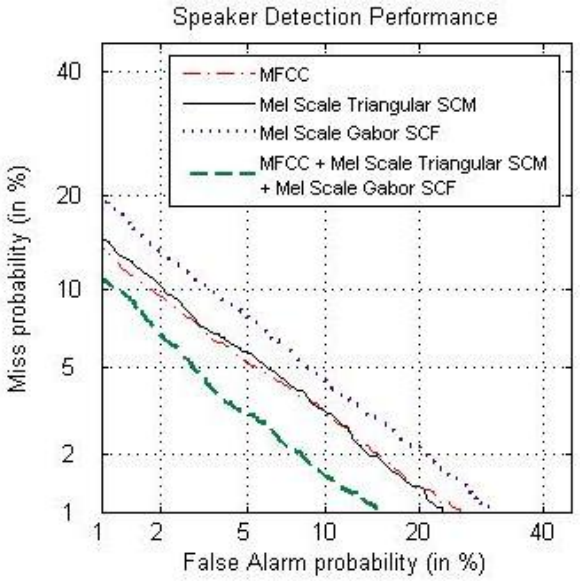

Figure 6: *DET curves showing the speaker recognition results (Baseline+NAP+ZTNorm) on the NIST 2006 SRE database*

[2] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Extraction of fm components from speech signals using all-pole model," *Electronics Letters*, vol. 44, no. 6, pp. 449–50, 2008.

[3] M. Nosratighods, T. Thiruvaran, J. Epps, E. Ambikairajah, B. Ma, and H. Li, "Evaluation of a fused fm and cepstral-based speaker recognition system on the nist 2008 sre," *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, vol. 0, pp. 4233–4236, 2009.

[4] H. A. Murthy and B. Yeganarayana, "Speech processing using group delay functions," *Signal Processing*, vol. 22, no. 3, pp. 259–67, 1991.

[5] B. Yegnanarayana, Hema A. Murthy, and V. R. Ramachandran, "Processing of noisy speech using modified group delay functions," 1991, vol. 2 of *Proceedings - ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 945–948.

[6] T. Thiruvaran, M. Nosratighods, E. Ambikairajah, and J. Epps, "Computationally efficient frame-averaged fm feature extraction for speaker recognition," *Electronics Letters*, vol. 45, no. 6, pp. 335 – 337, 2009.

[7] Kuldip K. Paliwal, "Spectral subband centroids as features for speech recognition," *IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 124 – 131, 1997.

[8] J. Chen, Y. Huang, Q. Li, and K.K. Paliwal, "Recognition of noisy speech using dynamic spectral subband centroids," *IEEE Signal Processing Letters*, vol. 11, no. 2, pp. 258 – 61, 2004.

[9] B. Gajic and K. K. Paliwal, "Robust speech recognition in noisy environments based on subband spectral centroid histograms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, no. 2, pp. 600 – 608, 2006.

[10] N.P.H. Thian, C. Sanderson, and S. Bengio, "Spectral subband centroids as complementary features for speaker authentication," Berlin, Germany, 2004, pp. 631 – 9.

[11] T. Kinnunen, B. Zhang, J. Zhu, and Y. Wang, "Speaker verification with adaptive spectral subband centroids," Seoul, Korea, Republic of, 2007, vol. 4642 LNCS, pp. 58 – 66.

[12] A. Solomonoff, W. M. Campbell, and I. Boardman, "Advances in channel compensation for svm speaker recognition," 2005, vol. Vol. 1 of *2005 IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp. 629–32, IEEE.

[13] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems," *Digital Signal Processing: A Review Journal*, vol. 10, no. 1-3, pp. 42–54, 2000.

[14] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Analysis of band structures for speaker-specific information in fm feature extraction," *Proc. INTERSPEECH*, 2009.

[15] 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, no. 10, pp. 3024–51, 1993.