Evaluation of a Fused FM and Cepstral-Based Speaker Recognition System on the NIST 2008 SRE

Mohaddeseh Nosratighods [1,2], *Tharmarajah Thiruvaran* [1,2], *Julien Epps* [1,2], *Eliathamby Ambikairajah* [1,2], *Bin Ma* [3], *Haizhou Li* [3,1]

[1] School of Electrical Engineering and Telecommunications
The University of New South Wales, Sydney, NSW 2052, Australia
[2] National ICT Australia, Australian Technology Park, Eveleigh, NSW 2015, Australia
[3] Human Language Technology Dept, Institute for Infocomms Research, A*STAR, Singapore

hadis@unsw.edu.au, thiruvaran@student.unsw.edu.au, j.epps@unsw.edu.au, ambi@ee.unsw.edu.au, mabin@i2r.a-star.edu.sg, hli@i2r.a-star.edu.sg

ABSTRACT

In this paper, the fusion of two speaker recognition subsystems, one based on Frequency Modulation (FM) and another on MFCC features, is reported. The motivation for their fusion was to improve the recognition accuracy across different types of channel variations, since the two features are believed to contain complementary information. It was found that the MFCC-based subsystem outperformed the FM-based subsystem on telephone conversations from NIST SRE-06 dataset, while the opposite was true for NIST SRE-08 telephone data. As a result, the FM-based subsystem performed as well as the MFCC-based subsystem and their fusion gave up to 23% relative improvement in terms of EER over the MFCC subsystem alone, when evaluated on the NIST 2008 core condition.

Index Terms Speaker Recognition, Frequency Modulation, MFCC, Fusion

1. INTRODUCTION

The fusion of complementary subsystems has become common practice in achieving good speaker verification accuracy [1]. This motivation suggests the investigation of features that are not explicitly based on spectral magnitude information. Some phase-based features, such as frequency modulation (FM) features [2,5], have shown promise in robust speech recognition [2], and there is psychoacoustic evidence to suggest that FM components are processed differently to amplitude information in the human auditory system [3]. Recently, an improved technique for the extraction of FM components from a sub-band decomposition of the speech signal was developed, and was shown to be effective in speaker recognition [4]. In this paper, the fusion of speaker recognition sub-systems based on MFCCs and FM features extracted using this technique is proposed. Comparative evaluation of the FM- and MFCC-based subsystems on the NIST 2006 and 2008 SRE data sets demonstrates the complementary properties of the two feature sets.

2. FM FEATURE EXTRACTION

Modeling of the speech signal in terms of frequency modulation components in this work is based on the AM-FM model of speech signals 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. The authors experimented with resonance-based FM extraction for automatic speaker recognition, and results of informal experiments were poor, due to the imperfect formant estimation. This motivated us to instead estimate FM components from the sub-bands of the speech signal. In the proposed FM sub-system, following a Bark-spaced Gabor filter bank analysis, each kth sub-band signal is modeled according to an AM-FM model [5]:

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

where $a_k[n]$ is the time-varying AM component, $q_k[r]$ is the FM component, f_s is the sampling frequency and f_{ck} is the center frequency of the kth band pass filter. To determine $q_k[r]$, we employ second-order all-pole modeling [4] to estimate the instantaneous frequency θ_k of the windowed sub band signal $p_k[n]$ as

$$\theta_k = \frac{2\pi f_{ck}}{f_s} + \frac{2\pi}{f_s} q_k[r] \ , \quad (2)$$

Practically, we estimate θ_k from the pole angle of the second-order linear predictor coefficients of the windowed sub band signal $p_k[n]$. The estimated FM component $q_k[r]$ at instant n is then obtained from the estimated θ_k by rearranging (2). We use one FM estimate per frame per sub-band. This recently proposed FM extraction technique has been shown to outperform the DESA [5] and Hilbert transform-based extraction methods for speaker recognition [4]. Finally, we note that since MFCCs are derived from Mel-spaced spectral magnitudes (related to $a_k[n]$) and the FM features here are based on θ_k, the two features can be considered to be complementary.

3. SYSTEM DESCRIPTION

The year 2008 NIST Speaker Recognition Evaluation [6] was distinguished from other recent evaluations by including in the training and test conditions for the core test not only conversational telephone speech data, but also conversational speech data recorded over a microphone channel involving an interview scenario, and additionally, conversational telephone speech recorded over a microphone channel. UNSW participated in the NIST 2008 speaker recognition evaluation as a collaborator with the Institute for Infocomms Research (IIR) from Singapore and four other universities [7]. Our contribution was providing a set of scores for two subsystems based on FM and MFCC front-ends for the core condition, termed the *short3-short2* condition [6].

3.1. Corpus and Tasks

The development database was designed to evaluate the system under different channel mismatches. The training condition was either a telephone conversation, termed *1conv4w* in NIST SRE-06, or *mixer5* data[1]; and the test condition was either a telephone conversation termed *1conv4w*, or a conversation, with the auxiliary microphone data known as *1convmic* from the NIST SRE-06 dataset, or *mixer5* data. The combinations of training and test conditions were used as development tasks for NIST SRE 2008. For instance, the combination of the cross-testing of mixer5 against 2006 telephone data is termed the mixer5-1conv4w task. The evaluation task was the NIST 2008 speaker recognition core training/test condition *short2-short3*, which can be further divided into 4 different tasks as shown in Table 1.

3.2. Feature Extraction

The Speech activity detection (SAD) employed in this experiment for NIST 2006 and 2008 SRE was based on energy, and accepted any frame with the energy above -50 dB, and within 30 dB of the maximum energy, while the number of accepted frames was within 40% - 60% of the total number of frames available.

MFCC subsystem: Features were extracted from 30 ms hamming-windowed frames, overlapped by 10 ms. The features employed by this subsystem comprised 13 MFCCs (*C*0 to *C*12), and the deltas and delta-deltas were appended following RASTA. Finally, mean-variance normalization was applied to all MFCC-based features.

FM subsystem: Features were extracted from 30 ms frames, overlapped by 20 ms. The features employed by this subsystem comprised frequency modulation (FM) features extracted according to the algorithm described in [4]: 14 Bark-spaced, Gabor filters were used to decompose the speech signal sub-band signals, from which FM components were estimated using the all-pole method [4]. Following feature warping of the FM features, delta-FM features were calculated, to produce 28-dimensional feature vectors.

3.2. Classification

Considering the benefits of using both discriminative and generative modeling, both subsystems employed GMM-SVM classifiers. For each sub-system, two 512-mixture gender-dependent GMM-UBMs were trained from NIST 2004 data (male: 2619 utterances; female: 2599 utterances) to model the background speakers, and GMM mean supervectors were computed by MAP adaptation with a relevance factor of 19. UBM training and MAP adaptation were performed using HTK. A GMM mean supervector kernel was used for SVM. The SVMs were trained using the NIST 2006 training data for the development and NIST 2008 training data for the evaluation tasks, with background data taken from the same NIST 2004 segments used in UBM training. SVM classification was implemented using SVMTorch [8].

Table 1: Training and test conditions comprising the 'short2-short3' condition, the core condition of NIST SRE 2008

		Test Segment Condition		
		telephone	microphone	interview
Training condition	telephone	tel-tel	tel-mic	
	interview	Interview-tel		interview-interview

3.4. NAP Parameter Optimization and Normalization

In order to model inter-session variations, separate NAP lists (comprising multiple recordings of several speakers) were prepared for telephone and microphone recorded data. Initially, the NAP list of telephone data was generated from NIST 2004 to model the inter-telephone[2] variation (3347/4498 male/female utterances from 124/185 male/female speakers) and tested on the development tasks of *1conv4w-1conv4w*. The NAP list of non-interview microphone recorded data was generated from NIST 2005 microphone recorded speech to model the Inter-microphone variation (374/437 utterances from 47/56 male/female speakers) and appended to the list for telephone data and tested on the *1conv4w-1convmic* task. The NAP list for interview data from mixer5 data was then created to fully model the inter-microphone variation (860/921 utterances from 3/3 male/female speakers), appended to the NAP list for telephone data and non-interview microphone data and finally tested on the *mixer5-mixer5* and *short2-short3* evaluation tasks. All the above NAP lists were selected to have at least seven recordings per speaker and a minimum duration of three seconds per utterance. A NAP rank of 80 was used in both sub-systems. TNorm was used to normalize the scores prior to fusion. For development and evaluation tasks, the 100- and 75-longest segments from the NIST 2005 training data for male and female were used respectively to provide the statistics for gender-dependent TNorm.

3.5. Fusion

The fusion weights were optimized using the NIST 2004-2006 development datasets. Fusion was achieved using a linear weighting scheme, with weights determined by linear search using the Focal software [9]. The combined system is shown in Figure 3.

[1] The mixer5 data consists of conversations of 6 speakers recorded over microphone for each speaker involved in an interview scenario. It was specially released by NIST for development purposes for the NIST 2008 evaluation.

[2] The intersession variation when both training and test sessions are telephone data.

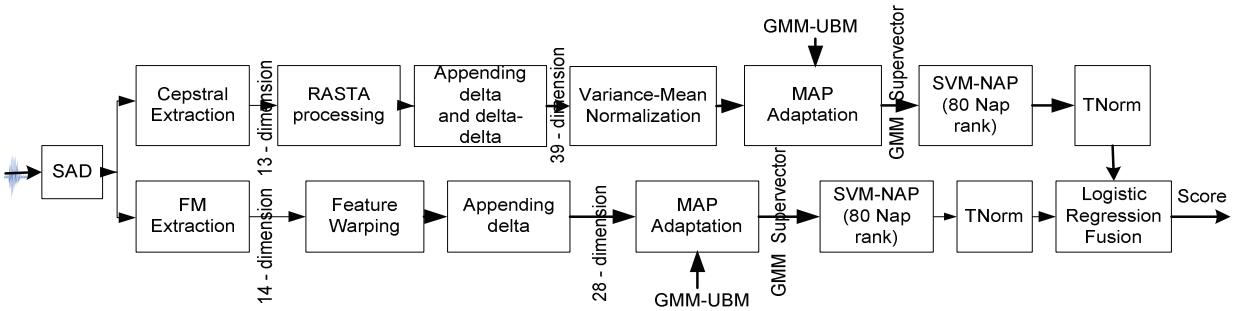


Figure 3: Proposed system, fusing FM-based and MFCC-based subsystems

4. NIST 2008 EXPERIMENTS

4.1. Development Experiments – 2006 Data

Evaluation results for different configurations of the MFCC/FM subsystem, tested on the NIST 2006 conditions (*1conv4w-1conv4w*, *1conv4w-1convmic*) and *mixer5-mixer5* development tasks are summarized in Table 2. These indicate that the system performance on *mixer5-mixer5* was poorer than that on telephone (i.e.*1conv4w-1conv4w*) and microphone (i.e.*1conv4w-1convmic*) tasks, irrespective of the features. The reason may be the sensitivity of the interview data to the speech activity detection. In each task (i.e. consistently), the fused system significantly outperformed both the MFCC and FM subsystems when taken on their own. The performance of the system for the telephone/telephone task outperformed that of the cross-channel conditions, i.e. telephone/microphone and interview/interview tasks, and the FM-based subsystem outperformed the MFCC-based subsystem on the interview data, while the opposite result was seen for the non-interview microphone and telephone data.

4.2. Evaluation Experiments – 2008 Data

NIST 2008 evaluation results for different configurations of the MFCC/FM and fused based system, tested on NIST 2008 core condition *(short2-short3*), are summarized in Table 3. Note that all the results include NAP and TNorm, as explained in section 3.3. The core condition evaluation results indicate that contrary to the development results, FM-based features alone show the same performance as MFCC-based features and even outperform the MFCC subsystem in terms of *min DCF*. Although the MFCC and FM subsystems showed similar performances, their fusion improved the accuracy by 17% and 14% in terms of relative reduction in *EER* when compared with the MFCC and FM stand-alone subsystems respectively. To analyse the effect of each channel or session variation in the NIST 2008 data set, evaluation results for the defined sub-tasks (See Table 1) for different configurations of MFCC/FM subsystems and their fused system are given in Table 4 and Figure 4. In the *interview-interview* condition (See Table 4, greyed rows), MFCC-based features performed better than FM when the data in training and test came from both the same and different microphones. Although the fusion enormously improved the error rate for every operating point compared with the FM/MFCC subsystems alone (i.e., 40% and 23% relative improvements in EER over FM and MFCC subsystems, respectively), the improvement mainly came from the interview-interview condition with different microphones in training and test (Table 4, row 3). However, contrary to the development results, the system performance in the *interview-interview* sub-task was better than that of the *tel-tel* sub-task where the training/test come from telephone channels. The reason might be that the NIST 2008 telephone data contains various languages in addition to English. This hypothesis is supported by Table 4 (rows 4-6). As can be clearly seen, the EER of the MFCC, FM and Fused systems of the English *tel-tel* task improved relatively by 30%, 35% and 40% respectively, compared with the *tel-tel* task including all languages.

Table 2. Development results for MFCC, FM and fused systems in terms of *EER (%),* left and *min DCF (%),* right.

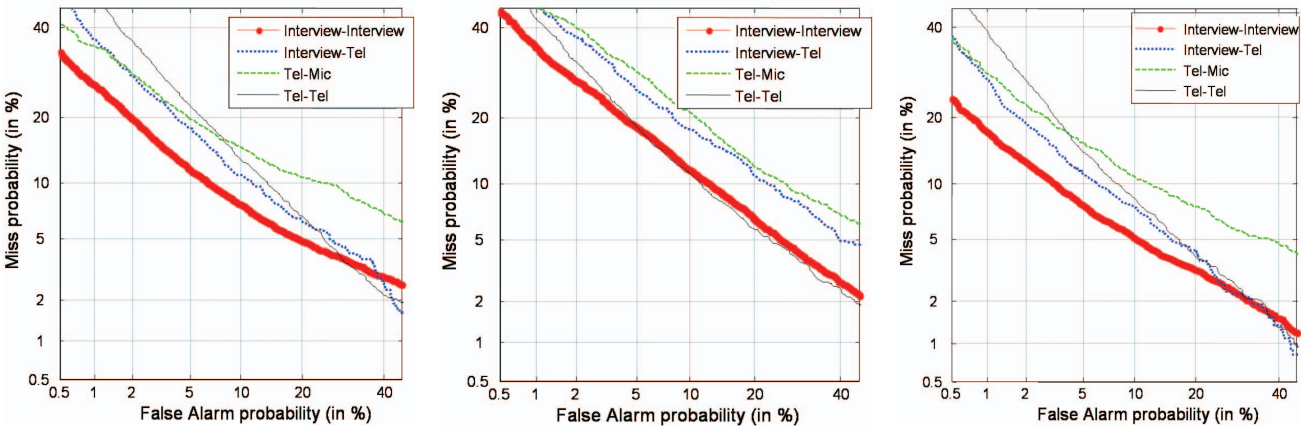
Systems	1conv4w-1conv4w		1conv4w-1convmic		mixer5-mixer5	
MFCC	5.33	2.62	5.78	2.51	23.26	7.98
FM	7.86	3.61	10.37	3.86	16.36	6.99
Fused	4.57	2.43	5.51	2.14	11.91	4.90

Table 3. Summary of results of MFCC, FM and fused systems, tested on the NIST2008 core condition (*short2-short3*), in terms of *EER (%),* left and *min DCF (%),* right.

Systems	Male		Female		All	
MFCC	9.74	4.80	14.45	6.05	12.78	5.70
FM	9.77	4.07	14.18	5.92	12.43	5.18
Fused	7.92	3.45	11.90	5.19	10.63	4.76

Another source of channel or session variation can be attributed to the cross-channel condition when the training and test data might come from different channel conditions. Comparing the results, the system for *Interview-tel* (Interview in training /telephone in test) performed significantly better, i.e. 30% relative reduction in EER, when compared with that of the *tel-mic* task (telephone in training/microphone in test). The reason may be the fact that the target models trained on interview data are more accurate than those trained on telephone data. Table 4 (rows 7-8) also indicates that the MFCC-based features outperform FM-based features for the two cross-channel tasks, while the fused system resulted in 22% and 18% relative reductions in terms of EER over the best MFCC subsystem in *Interview-tel* and *tel-microphone* tasks, respectively.

Figure 4(a) illustrates the performance of different variations of the core condition in NIST 2008 for the MFCC subsystem individually. It shows that the *interview-interview* task achieved 24% relative reduction in EER when compared to that of the *tel-tel* task in the MFCC-based subsystem; the *Interview-tel* task also outperformed that of *tel-microphone* task in higher false alarm probability regions, however they behave almost the same in low false alarm probability regions. The performance of different variations of the core condition in NIST 2008 for the FM subsystem is shown in Figure 4(b). It can be seen that, unlike the



(a) (b) (c)

Figure 4: DET curves showing the performance of various conditions of the NIST 2008 SRE core condition for (a) the MFCC subsystem, (b) FM subsystem and (c) the fused system.

Table 4: Breakdown of results for the MFCC, FM and fused systems, tested on the NIST 2008 core condition (*short2-short3*), in terms of *EER* (%), left, and *min DCF* (*100), right.

Sub-tasks	MFCC		FM		Fused	
Interview-Interview	8.55	3.59	11.0	4.43	6.54	2.70
same microphone in Train/Test	4.63	1.25	6.64	1.89	3.60	0.94
different microphone in Train/Test	8.75	3.66	11.2	4.52	6.68	2.73
Tel/Tel	11.6	5.52	10.7	5.04	9.11	4.64
English language in Tel./Tel	8.14	3.58	6.94	3.17	5.54	2.56
Native US English in Tel/Tel	8.22	3.77	7.24	3.47	5.59	8.32
Inter-Tel	10.7	4.63	14.7	5.50	8.32	3.53
Tel-mic	13.0	4.38	15.3	5.47	10.6	3.86

MFCC-based subsystem, the FM subsystem performed equally well in *tel-tel* and *interview-interview* tasks; Furthermore, the system performance of *interview-tel* was slightly better than that of the *tel-microphone* task using FM features. Figure 4(c) shows that the trend of the fused DET-curve is very similar to the trend of the MFCC subsystem. However, significant improvement at every operating point is achieved by fusing the MFCC with the FM subsystem, which again confirms that MFCC and FM features contain complementary information.

5. CONCLUSION

In this paper two subsystems, based on FM and MFCC features respectively, were fused together in a discriminative framework and were compensated using NAP. The fusion of the two subsystems was found to improve the results significantly under different channel or speaker circumstances, a result attributed to the complementary information carried by the two types of features. Interestingly, the individual FM and MFCC subsystems proved to be complementary across different training/test conditions. In the case of the NIST 2008 core condition, the FM-based subsystem performed as well as the MFCC-based subsystem on average and the channel-compensated fused system produced up to 23% relative EER improvement over the channel-compensated MFCC subsystem alone.

5. REFERENCES

[1] N. Brummer, L. Burget, *et al*., "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006, *IEEE Trans. on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072-2084, 2007.

[2] D. Dimitriadis, P. Maragos, and A. Potamianos, "Robust AM-FM features for speech recognition", *IEEE Signal Process. Letters*, vol. 12, no. 9, pp. 621_624, 2005.

[3] D. Regan and B. W. Tansley, "Selective adaptation to frequency-modulated tones: Evidence for an information-processing channel selectively sensitive to frequency changes," *JASA*, vol. 65, pp. 1249_1257, May 1979.

[4] T. Thiruvaran, E. Ambikairajah, and J. Epps, "Extraction of FM components from speech signals using an all-pole model," *IET Electronics Letters*, vol. 44, no. 6, March 2008, pp. 449-450.

[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, no. 10, pp. 3024-3051, 1993.

[6] "The NIST Year 2008 Speaker Recognition Evaluation Plan", http://www.nist.gov/speech/tests/sre/2008/index.html, 2008.

[7] B. Ma, H. Sun, *et al*., "The I4U System in NIST 2008 Speaker Recognition Evaluation," apears in *Proc. ICASSP,* Taipei, Taiwan, April 2009.

[8] R. Collobert and S. Bengio, "SVMTorch: Support vector machines for large-scale regression problems," *Journal of Machine Learning Research*, vol. 1, pp. 143_160, February 2001.

[9] N. Brummer, "Tools for fusion and calibration of automatic speaker detection systems," http://www.dsp.sun.ac.za/nbrummer/focal/, 2005.