A COMPARISON OF FUSION TECHNIQUES IN MEL-CEPSTRAL BASED SPEAKER IDENTIFICATION *

S.Slomka, S.Sridharan and V. Chandran

Speech Research Laboratory, Queensland University of Technology. GPO Box 2434, Brisbane, QLD, 4001, Australia

s.slomka@qut.edu.au, s.sridharan@qut.edu.au, v.chandran@qut.edu.au

ABSTRACT

Input level fusion and output level fusion methods are compared for fusing Mel-frequency Cepstral Coefficients with their corresponding delta coefficients. A 49 speaker subset of the King database is used under wideband and telephone conditions. The best input level fusion system is more computationally complex than the output level fusion system. Both input and output fusion systems were able to outperform the best purely MFCC based system for wideband data. For King telephone data, only the output level fusion based system was able to outperform the best purely MFCC based system. Further experiments using NIST'96 data under matched and mismatched conditions were also performed. Provided it was well tuned, we found that the output level fused system always outperformed the input level fused system under all experimental conditions.

1. INTRODUCTION

Traditionally closed-set Speaker Identification (SI) systems use LPC or FFT derived spectral coefficients as input features. Transitional spectral representations such as first order differences of LPC or FFT derived Cepstral coefficients may provide additional uncorrelated information for SI. First order delta coefficients [1] have been investigated for text-independent SI and in that study it was found that the LPC-derived Cepstral coefficients and first order delta coefficients could be linearly combined at the classifier output (output level fusion) to improve performance for telephone speech [2]. Results obtained in this study were believed to be equally applicable to filterbank and FFT derived cepstral coefficients. Appending the delta coefficients to the cepstral coefficients at the classifier input (input level fusion) [3][5][6] has been recently investigated and found to improve SI performance in telephone speech.

In this paper, we investigate and compare input level fusion and output level fusion of Mel-Frequency warped FFT derived Cepstral Coefficients (MFCC) and their corresponding delta coefficients (DeltaMFCC) in the framework of text-independent SI. The SI system used is based on Gaussian Mixture Models (GMMs) [6]. A GMM based SI system is chosen because it is well known and has previously been used with input fusion [3][6]. The investigation is carried out mainly on the KING database but some supplementary experiments are also performed using the 1996 NIST Speaker Recognition Evaluation database. The main foci of the comparison are SI system computational complexity, and SI.

2. THE SPEAKER IDENTIFICATION SYSTEM

The SI system used for this study is based on Gaussian Mixture Model [7]. The concept of a GMM is to model a target Probability Density Function (PDF) with multiple weighted gaussian component PDFs (typically referred to as mixtures). The probability of a D dimensional test vector $\vec{x}$ belonging to target model λ is given by:

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} w_i \, g(\vec{x}, \vec{\mu}_i, \Sigma_i) \qquad (1)$$

where w_i is the weight of the ith gaussian component PDF and $g(\vec{x}, \vec{\mu}_i, \Sigma_i)$ is the likelihood of $\vec{x}$ belonging to gaussian component PDF i. The value of the latter is given by:

$$g(\vec{x}, \vec{\mu}_i, \Sigma_i) = \frac{1}{(2\pi)^{D/2} \mid \Sigma_i \mid^{1/2}} \exp\left\{ -\frac{1}{2} (\vec{x} - \vec{\mu}_i)' \Sigma_i^{-1} (\vec{x} - \vec{\mu}_i) \right\} \qquad (2)$$

where $\vec{\mu}_i$ and Σ_i are the mean vector and covariance matrix of gaussian component PDF i. The number of component PDFs is referred to as the order of the model M. In this study we use nodal, diagonal covariance matrices in line with [6]. Given S sample training vectors, the parameters of a GMM are generally estimated by iterating through the Expectation-Maximisation (EM) algorithm [4] for I iterations, it is relativly simple to show that the computational complexity of GMM training is O(SxDxMxI) and GMM testing is O(SxDxM) as S,D,M and I become large. In this study GMM training was terminated when either a fixed number of iterations I_{fixed} was reached or when the change in training between two consecutive iterations i-1 an i for $1 \le i \le I_{fixed}$, was less than a set threshold $\Delta_{ethresh}$. In an input level fusion based SI system, the two parameter types MFCC and DeltaMFCC are simply concatenated, thus increasing D. One GMM model λ_i (i=1,2,…,N) is then trained for each of the N speakers. For a given test sequence of T vectors, the input level fusion SI system computes a speaker identity using:

$$X = \arg \max_{1 \le n \le N} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_n) \qquad (3)$$

In output fusion, for each speaker, one GMM is trained and tested solely with DeltaMFCC coefficients and a second GMM

*This project is partially funded by an Defence Science and Technology Organisation of Australia research contract.

solely with MFCC coefficients. For each test vector, the GMM outputs (in the form of log-likelihoods) are equalised by their corresponding averages of pooled intraspeaker log-likelihoods $\bar{l}_{MFCC}$ and $\bar{l}_{\Delta MFCC}$ over the whole test sequence and then linearly combined on a frame-by-frame basis as in [2]. The output level fusion based SI system thus computes a speaker identity:

$$X = \arg \max_{1 \le n \le N} \sum_{t=1}^{T} \{ \quad \alpha \frac{\log p(\vec{x}_t \mid \lambda_n^{MFCC})}{\bar{l}_{MFCC}} \tag{4}$$

$$+ (1-\alpha) \frac{\log p(\vec{x}_t \mid \lambda_n^{\Delta MFCC})}{\bar{l}_{\Delta MFCC}} \}$$

where $0 \le \alpha \le 1$. Fusing the outputs increases the number of GMMs (two for every speaker). The estimation of GMM model parameters may be differently affected by the two approaches (even for diagonal covariance matrices).

3. SPEECH PARAMETERISATION

In this study we use 15 dimensional Mel-Frequency FFT derived Cepstral Coefficients and their corresponding Delta Coefficients [1] as input features. MFCC Coefficients were derived with a frame size of 32ms and a frame advance of 10ms. DeltaMFCC are calculated using the formula:

$$d_t = \frac{\sum_{k=1}^{K} k(c_{t+k} - c_{t-k})}{2\sum_{k=1}^{K} k^2} \tag{5}$$

where d_t is a delta coefficient at time t computed from the static coefficients c_{t-k} to c_{t+k} . The value of K determines the size of the "window" of static coefficients used. Except for results in Section 4.5 the value of K used is always 5.

Silence and low energy frames are removed from all speech. Cepstral mean subtraction is performed on the telephone speech only. The telephone speech is also bandlimited in the range 300-3400Hz.

4. EXPERIMENTS ON KING DATA

Experiments speaker have been carried out using the 49 speakers of the King database [8] which have exactly 10 recording sessions per speaker for both the wideband WB (clean speech) and narrowband NB (telephone) speech portions. The first 5 recording sessions are recorded at approximately 1 week intervals and the others at approximately 1 month intervals. The first session for each speaker was used for GMM training and the other 9 sessions for testing.

Test speech for each session was divided into overlapping 5s test segments with a 10ms frame advance. Silence was removed after this division of the test speech. In total 1756234 test segments are created this way. After silence removal the average duration of the test segments was 3.5s and 2.2s for the WB and NB portions respectively. The accuracy of our SI systems is the percentage of test segments correctly identified.

4.1. Pure MFCC and DeltaMFCC Optimisation

To optimise the purely MFCC and DeltaMFCC based SI systems required for the output fused approach we trained SI systems using GMMs with model order M=1…60 inclusive. For each value of M, GMM were trained with one of 4 training conditions (I_{fixed}=25 and $\Delta_{ethresh}$=0.005, I_{fixed}=40 and $\Delta_{ethresh}$=0.0005, I_{fixed}=80 and $\Delta_{ethresh}$=0.00005, I_{fixed}=200 and $\Delta_{ethresh}$=0.000005). For the remainder of this paper we will simply refer to each training condition by its I_{fixed} value. So when we refer to I_{fixed}=25 we also mean $\Delta_{ethresh}$=0.005.

To determine the optimum SI system we chose the system where further increasing M or I_{fixed} , resulted in less than 0.1% performance improvement. Table 1. shows the optimum SI system parameters. The actual average number of iterations used to train each of the GMMs are also shown.

Data	Param. Type	M	I_{fixed}	Avg. I	Sys. Acc. (%)
WB	MFCC	30	80	77	65.0
	DeltaMFCC	20	25	12	45.9
NB	MFCC	50	80	78	30.9
	DeltaMFCC	10	40	26	24.0

Table 1: Optimum system parameters for King data.

During the optimisation process it was found that for the narrowband data a purely DeltaMFCC based system had comparable perfomance with a purely MFCC based system when the value of M was between 3 to 10 inclusive. With this exception, it appears that the purely MFCC based system always performs better than the purely DeltaMFCC based system for the same model order. These results are shown in Figure 1 for I_{fixed} =200 but they also hold for the other three values of I_{fixed}.

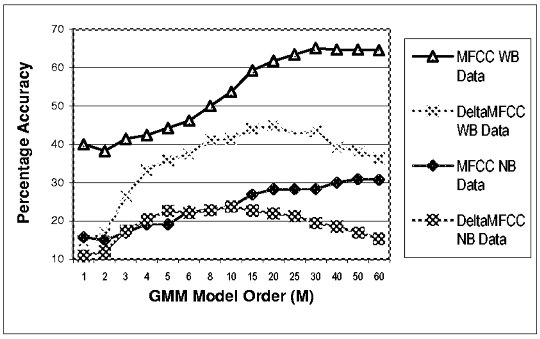


Figure 1: Accuracy vs the GMM model order (I_{fixed} =200) for the King data.

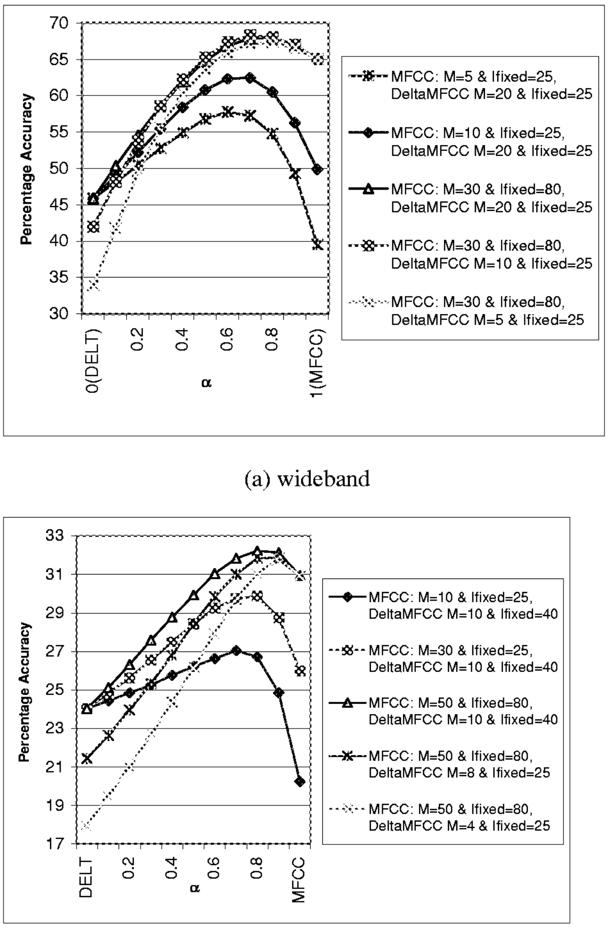
4.2. Optimising the Input Level Fusion System

To optimise the input level fusion system, the same was used with *M* in the range 30 to 90. The results are shown in Table 2.

Data	M	I_{fixed}	Avg. I	Sys. Acc. (%)
WB	50	200	200	66.2
NB	50	200	173	27.1

Table 2: Optimum Input level fusion system parameters for King data

It is interesting to note that the optimum input level fusion system (shown in Table 2) was unable to outperform the optimum MFCC based system for the narrowband data under the available training conditions. This appears to be contrary to results obtained by others using a 16 speaker subset of narrowband speech from the King database [6].



(a) wideband

(b) narrowband

Figure 2: Output level fusion perfomance as a function of α

4.3. Optimising the Output Level Fusion System

Choosing an output level fusion system using the optimum MFCC and DeltaMFCC SI systems from Section 4.2 resulted in peak accuracies of 68.0% and 32.2% for wideband and narrowband data respectively with the optimum value of α=0.8. These systems are chosen as our "optimal" output fusion systems. Output level fusion performance as a function of α for a number of systems from Section 4. is shown in Figure 2. From the narrowband results of Figure 2 it can be seen that a GMM based system needs α somewhat greater than the 0.5 required for similar Vector Quantisation based SI systems [2].

It appears that if the model order is sufficiently close to the optimum, output fusion with DeltaMFCC does not improve the performance much. However, for smaller model orders (which are more attractive in terms of computational and storage requirements) a fused system gives better accuracy and there is an optimal linear combination for each order.

4.4. Computational Complexity

Let the complexity of training an SI system with *D*=1,*M*=1 and *I*=1 be equal to 1 unit/frame and the complexity of testing an SI system with *D*=1,*M*=1 and *I*=1 be 1 te_unit/frame (where prefix te denotes testing). Table 3 shows that the training complexity of the output level fusion approach is approximately 4 times more complex at our chosen optimums for the wideband data and about 2 times more complex for the narrowband data. Testing complexities are the same as shown in Table 3 except that they are in terms of te_units/sample.

	D*M*I (optimal)	Complexity units/frame
WB input level fusion	15*50*200	150000
WB output level fusion	15*(30*77+20*12)	38250
NB input level fusion	15*50*173	129750
NB output level fusion	15*(50*78+10*26)	62400

Table 3: Training complexity for the "optimum" fusion systems.

K	*M*	I_{fixed}	Avg. I	Sys. Acc. (%)
2	25	40	34	46.4
3	20	80	68	47.9
4	20	40	33	46.5
5	20	25	12	45.9

Table 4: Optimal DeltaMFCC based SI systems configuration and performance for K from 2 to 5 inclusive.

4.5. Window Size

The accuracy may also depend on the "window" size, *K*, used to compute the DeltaMFCC from MFCC. A study similar to the one in [2] was proposed for the wideband data. The optimum system for each *K* was found in the same fashion as for *K*=5 in section 4.2. The results are shown in Table 4. The optimal value of *K* is 3.

5. EXPERIMENTS ON NIST'96 SPEAKER RECOGNITION SPEECH DATA

As a final study the two fusion approaches are compared on NIST's 1996 Speaker Recognition Evaluation database under various telephone handset conditions. In the our study the development data consisting of 43 male and 45 female speakers is used. Training is performed using approximately the same amount of speech (on average 30s after silence removal, which was approximately 45s prior to silence removal) as was used for the King narrowband experiments. The 30s test sessions present were used for test data. The test data was divided up into 5s test segments in the same fashion as for the prior experiments on King (939966 test segments in total). SI input and output level fusion systems with identical configuration as in the NB data sections of Tables 1&2 are used.

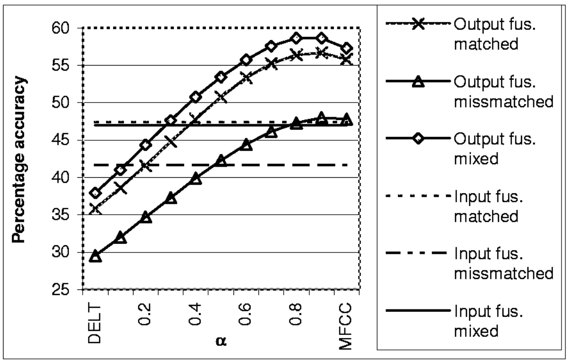


Figure 3: 88 speaker NIST 1996 development data SI accuracy.

The following three experimental conditions are evaluated: 1) missmatched handsets, 2 matched handsets and 3) mixed handsets. For the missmatched handset case training data is taken from handset 1 (session 1a) and tested with handset 2 and vice versa. In the matched handset case training and testing data from the same handset type are used. For the mixed handset experiments 50% of the training data was taken from each handset condition and the testing data from both handsets was used. The results of the experiments are shown in Figure 3.

Under all conditions it was possible to find an α for which output fusion outperformed the purely MFCC based system. The input level fused system was unable to outperform the purely MFCC based system for this dataset. It should be noted that if α=0.8 is chosen from our previous study then the output level fused system performed slightly worse than the purely MFCC based system for the missmatched condition case.

6. CONCLUSIONS

In all experiments carried output level fusion system always out performed the purely MFCC or DeltaMFCC based systems. In our experiments it was not possible to construct an input level fusion based system which could outperform the best purely MFCC based system for the telephone data used. This is contrary to results reported in the literature [6][5] and the cause of this may need be investigated in the future. Theoretically an input level fusion system should be able to classify better than each individual system if there is extra information. In practise, classifiers with input level fusion may not converge to "true" optimal states with limited training. Perhaps the higher dimensional input level fusion systems were better able to converge in these other studies due to the larger amounts of training data used by those researchers.

It may be necessary to recompute optimal α values for when migrating to different experimental conditions. Investigations into the information content and separability properties of the two parameter sets and their relationship to classifier performance will yield more insight into the experimental results presented.

7. REFERENCES

1. Furui, S., "Cepstrum Analysis Technique for Automatic Speaker Verification", *IEEE Trans. Acoust,, Speech, Signal Processing*, Vol. ASSP-29, No. 2, 254-272, 1981.
2. Soong, F.K., and Rosenberg, A.E., "On the use of Instantaneous and Transitional Spectral Information in Speaker Recognition", *IEEE Trans.. Acoust., Speech and Signal Processing*, Vol. 36, No. 6, 871-879, 1988.
3. Reynolds, D.A., "The Effects of Handset Variability on Speaker Recognition Performance: Experiments on the Switchboard Corpus", *Proc. ICASSP*, 113-116, 1996.
4. A.P. Dempster, N.M. Laird and D.B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm", *J. Royal Statistical Society*, Vol. 39, 1-39, 1977.
5. Li W.-Y., and O'Shaugnessy, D., "Hybrid Network Based on RBFN and GMM for Speaker Recognition", *Proc. Eurospeech*, 955-858, 1997.
6. Reynolds D.A. and Rose R.C., "Robust Text-Independent Speaker Identification Using Gaussian Mixture Speaker Models", *IEEE Transactions on Speech an Audio Processing*, Vol. 3, No. 1, 72-83, 1995.
7. Reynolds, D.A., "Speaker Identification and Verification using Gaussian Mixture Speaker Models", *Speech Communication*, v17, 91-108, 1995.
8. Godfrey, J., Graff, D., and Martin, A., "Public Databases for Speaker Recognition and Verification", *Proc. ESCA Workshop Automat. Speaker Recognition, Identification, Verification*, 39-42, 1994.