DYNAMIC SELECTION OF MAGNITUDE AND PHASE BASED ACOUSTIC FEATURE STREAMS FOR SPEAKER VERIFICATION

R.Padmanabhan [1], *Rajesh M. Hegde* [2], *Hema A. Murthy* [1]

[1] Department of Computer Science and Engineering,
Indian Institute of Technology Madras, Chennai, India
Email:{padmanabhan, hema}@lantana.tenet.res.in

[2]Department of Electrical Engineering
Indian Institute of Technology Kanpur, India
Email: rhegde@iitk.ac.in

ABSTRACT

The use of joint acoustic features at the feature level leads to vectors of large dimensions and computational complexity. In this paper we propose a method for dynamic acoustic feature stream selection based on mutual information criterion for speaker verification. The method is based on the intuition that different acoustic features are better suited for recognising different speakers. An optimal selection of a particular feature stream for a particular speaker is necessary, assuming that the feature streams have complementary information. We use acoustic features derived from the magnitude and the phase spectrum for this method as they have empirically demonstrated diversity in previous work. An information theoretic measure based on mutual information is proposed to hypothesise the appropriate feature stream for the appropriate speaker. Separability analysis based on the Bhattacharya distance is also presented to verify this hypothesis. The proposed technique requires a claimed identity for dynamic feature selection during the test phase. We therefore apply it to the task of speaker verification. Reasonable improvements in verification performance are noted from the DET curves. The proposed method also significantly reduces the computational complexity compared to the use of joint feature streams.

1. INTRODUCTION

Several acoustic features in combination at the feature level have been widely used for improved recognition performance in speaker identification and verification. All these methods have additional computational complexity associated with their use when compared to use of individual feature streams. A workaround to this problem can be the use of specific feature streams for specific speakers since it is reasonable to assume that a single feature might not be the best for recognising different speakers.

A similar technique has been used in [8], for syllable recognition. Here, acoustic feature diversity was incorporated in the linguistic feature space. For each syllable in the vocabulary, the feature that gives better identification accuracy is determined from training data. During the testing phase, the language model gives a list of probable syllables. The likelihoods of these syllables are computed using a weighted combination of features, with a higher weight being given to the "better" feature. As a result, there was significant reduction in the word error rate.

A similar approach to speaker identification is not practically feasible as we do not have a claimed identity. Therefore in order to dynamically switch between two different acoustic feature streams we need a claimed identity as in a speaker verification system. A speaker verification problem is therefore selected for employing the proposed technique as we have a speaker claim, and based on the claim, one can use the appropriate feature for the claimed speaker.

Feature-speaker pairs are first determined from the training data ie. which feature is optimum for which speaker. In the testing phase, since we do not have a list of probable speakers, the claimed speaker identity can be used for the purpose of determining the optimum feature pair. The two feature streams considered in this work are the Mel frequency cepstral coefficients (MFCC) [2], and the modified group delay (MODGDF) [6]. These two features are derived from the magnitude and phase parts of the Fourier spectrum respectively. Speaker recognition systems built using these two features individually give almost comparable performance [6] and improve the performance when used jointly [5].

The rest of this paper is organised as follows. We start with a discussion on feature combination schemes, followed by the procedure to quantify mutual information between the Fourier magnitude spectrum and the two acoustic feature streams. This is followed by a description of the dynamic acoustic feature stream selection procedure based on maximising the mutual information to compute feature-speaker pairs. Separability analysis results are also illustrated next to substantiate this conjecture of computing optimal feature-speaker pairs. A speaker verification system based on this approach is described next, giving reasonable improvements in verification performance as illustrated by the DET curves.

2. CONVENTIONAL METHODS OF COMBINING ACOUSTIC FEATURES

Methods for combining information contained in multiple feature streams can be broadly classified into three types [5]:

- Early fusion: Multiple acoustic features are simply concatenated before training and testing.
- Late fusion: Probabilities or distortion values from multiple features are combined at various stages to make a decision.
- Hypothesis fusion: 1-best or N-best outputs of each

acoustic feature is used to generate a single hypothesis.

Early fusion mechanisms using magnitude and phase based features have been successfully used for phoneme, syllable, speaker and language recognition using the MFCC and MODGDF [5]. Although the method shows good improvement in performance, the computational requirement for working on a high dimensional (84 dimensions in [5]) feature streams is high. Late fusion methods combine weighted probabilities from multiple feature streams to give a final classification decision. Some form of normalisation is required to combine the probabilities, as their ranges are usually different. The weights are usually determined empirically.

3. DYNAMIC ACOUSTIC FEATURE STREAM SELECTION BASED ON MUTUAL INFORMATION

Information-theoretic approaches for the study of feature selection [3] have given interesting results for speaker identification. They can be used for determining the *optimal feature* for a particular speaker. Optimality is defined in terms of the recognition or verification performance in this work and is determined from the training data.

3.1 Entropy and mutual information

The entropy of a discrete random variable $X = \{x_0, x_1 \ldots x_{N-1}\}$ is defined as

$$H(X) = -E[\log_2 p_i]$$

where $p_i = \Pr(X = x_i)$, $E[.]$ is the expectation operator, and Pr denotes probability. The conditional entropy

$$H(X|Y = y_k) = -E[\log_2 p_i]$$

where now $p_i = \Pr(X = x_i|Y = y_k)$. The average conditional entropy $H(X|Y)$ is defined as

$$H(X|Y) = \sum_y \Pr(Y = y)H(X|Y = y)$$

The mutual information (MI) between two random variables X and Y is defined as

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

The mutual information is a measure of how much information X and Y have in common. It can also be viewed as how much uncertainty exists in one if we know the other. In [3], the minimum classification error probability was shown to be related to the mutual information between speakers and features. It was also shown here that the probability of error is minimised when the MI is maximised.

3.2 Computation of mutual information

Dynamic feature stream selection requires computation of mutual information between the individual feature streams and a reference spectrum. We use the short term Fourier transform magnitude spectrum as a reference spectrum for computation of the mutual information. Algorithm 1, outlines the computation of the mutual information using the reference magnitude spectrum and the acoustic features in the cepstral domain. Note that the two specific features under consideration are the MFCC and the MODGDF.

Algorithm 1 mi($\mathcal{X}$,$\mathcal{Y}$)

```
{Given: Training data for one speaker, consisting of $N$ frames}
From the $N$ training frames, extract the set of short-time spectra $\mathcal{X} = \{X_i\}$ and set of feature vectors (MFCC or MODGDF) $\mathcal{Y} = \{Y_i\}$ where $i = 1 \ldots N$.
Perform vector quantisation on the set $\mathcal{X}$ to form a codebook $C$. Similarly perform vector quantisation on the set $\mathcal{Y}$ to form a codebook $D$. Let both codebooks have $P$ centroids.
The relative frequency of each centroid is an approximate measure of the probability of occurrence of that centroid. Let $\hat{X}_j$ and $\hat{Y}_j$ denote centroids and $C_j$ and $D_j$ denote clusters in $C$ and $D$ respectively, with $j = 1 \ldots P$. The the probabilities can be estimated as:
$\Pr(\hat{X}_j) \approx \frac{|C_j|}{N}$
$\Pr(\hat{Y}_j) \approx \frac{|D_j|}{N}$
$\Pr(\hat{Y}_j|\hat{X}_j) \approx \frac{|D_j, C_j|}{|C_j|}$
From the probabilities, we can estimate $H(\mathcal{X})$, $H(\mathcal{Y})$ and $H(\mathcal{X}|\mathcal{Y})$.
Compute $mi(\mathcal{X}, \mathcal{Y}) = H(\mathcal{X}) - H(\mathcal{X}|\mathcal{Y})$.
```

3.3 Dynamic feature selection based on mutual information

Having computed the mutual information as outlined in section 3.2, we now turn to the proposed approach for dynamic acoustic feature stream selection for speaker verification. Let mutual information between the reference spectrum and an acoustic feature stream be

$$\theta_i = mi(\mathcal{X}, \mathcal{Y}_i)$$

where

$$i \in \{\text{MGD}, \text{MFC}\}$$

and MGD represents the acoustic feature stream MODGDF and MFC corresponds to the acoustic feature stream MFCC. The optimum feature stream $\hat{i}$ is now selected as

$$\hat{i} = \arg\max_i \{mi(\mathcal{X}, \mathcal{Y}_i)\}$$

4. ANALYSIS OF ACOUSTIC FEATURE DIVERSITY

All our experiments are performed on a subset of 200 speakers from the NTIMIT database [4]. To verify our conjecture on optimal feature selection with mutual information, we investigate the separability of different speakers with different features before we apply them to a speaker verification task. The following two sections discuss results of acoustic diversity analysis using the Bhattacharya distance and some speaker identification experiments.

4.1 Separability analysis using the Bhattacharya distance

The Bhattacharya distance is used as a measure of separability between two classes. For normal distributions, the squared Bhattacharya distance d_B^2 between two classes with means μ_i and covariance matrix M_i is defined as

$$d_B^2 = \frac{1}{2}\ln\frac{|\frac{M_1+M_2}{2}|}{|M_1|^{1/2}|M_2|^{1/2}} + \frac{1}{8}(\mu_1-\mu_2)^T \left(\frac{M_1+M_2}{2}\right)^{-1}(\mu_1-\mu_2)$$

Figures 1 and 2 show the Bhattacharya distance between a speaker and the corresponding background model against dimension of the feature vector. The separability is plotted for models built with MFCC and models built with MODGDF (both speaker model and background model). Since speaker verification is a two class problem, greater the separability between the speaker model and the background model, the more accurate the verification.

From the measure of mutual information, we determine that the speaker in Figure 1 (denoted as speaker A) has MFCC as the optimal feature whereas the speaker in Figure 2 (denoted as speaker B) has MODGDF as the optimal feature. The mutual information values are tabulated in Table 1. From the plots, it can be seen that the speaker in Figure 1 has better separation with MFCC features, whereas the speaker in Figure 2 has better separation with MODGDF, thus substantiating the conjecture on optimum feature selection based on mutual information.

Table 1: Mutual information and optimal feature selection for speakers A and B.

Speaker	$mi(\mathcal{X},\mathcal{Y}_{\text{MFC}})$	$mi(\mathcal{X},\mathcal{Y}_{\text{MGD}})$	Optimal feature
Spk A	**2.867**	2.675	MFCC
Spk B	2.651	**2.664**	MODGDF

4.2 Acoustic diversity analysis using speaker identification experiments

To further explore the acoustic diversity of MFCC and MODGDF features, a conventional closed-set speaker identification system using each feature was evaluated using the same dataset. The identification result is summarised as the fraction of the number of test cases in Figure 3. Four tests were performed on each speaker, and the identification accuracy ranges from zero (none of the test cases were identified correctly) to four (all four test cases were identified correctly).

From the results of actual speaker identification it is reasonable to assume that each acoustic feature recognises different fractions of the number of test cases confirming the need for optimal feature selection for different speakers.

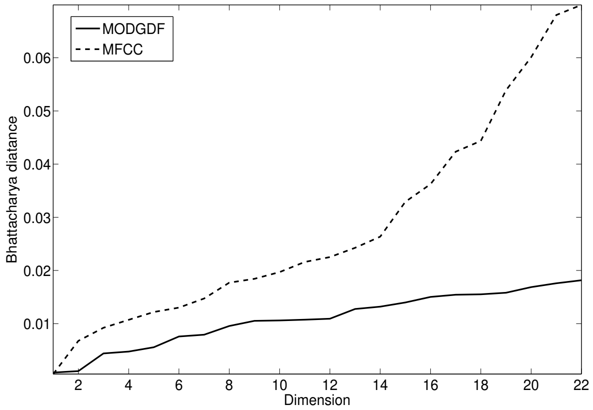

Figure 1: Bhattacharya distance versus feature dimension for speaker A and corresponding background model. MFCC shows higher separability for this speaker.

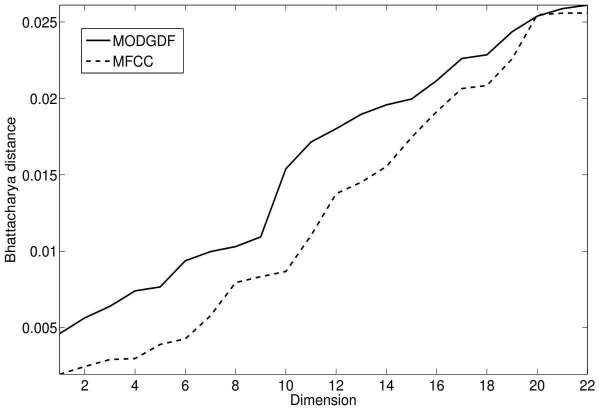

Figure 2: Bhattacharya distance versus feature dimension for speaker B and corresponding background model. MODGDF shows higher separability for this speaker.

5. SPEAKER VERIFICATION FRAMEWORK

Figure 4 shows the architecture of the proposed speaker verification framework. In the training phase, the optimal feature (ie MFCC or MODGDF) is determined for each speaker using the mutual information between the feature and the Fourier spectrum. A lookup table is built which outputs the optimal feature for a given speaker. Gaussian mixture models, as described in [9] are built for each speaker using the optimal feature. Background models are also built for each speaker from training data of the other 199 speakers using the optimal feature.

In the testing phase, the optimal feature is determined for the claimed speaker from the lookup table. The corresponding features are extracted from the the input speech waveform. The verification decision is made from the likelihood ratio scores of the speaker model to the background model compared to a thresh-

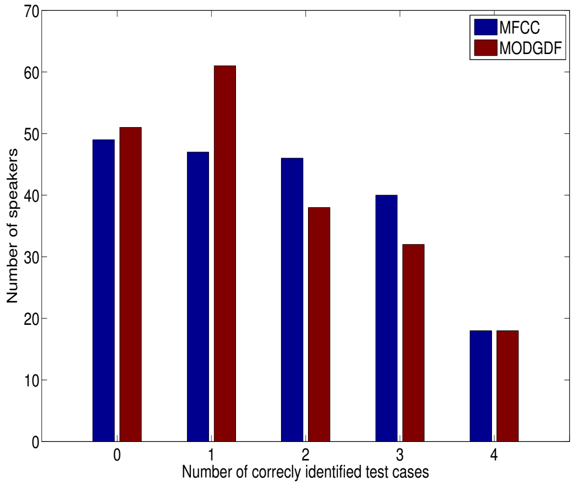

Figure 3: Histogram of identification accuracy (out of four test files for each speaker) for MFCC and MODGDF features.

old, as in conventional speaker verification systems [1].

The advantage of this approach is that we always choose a feature which shows better separability between a speaker and the corresponding background model. This results in lesser number of incorrect classifications (false alarms and misses.)

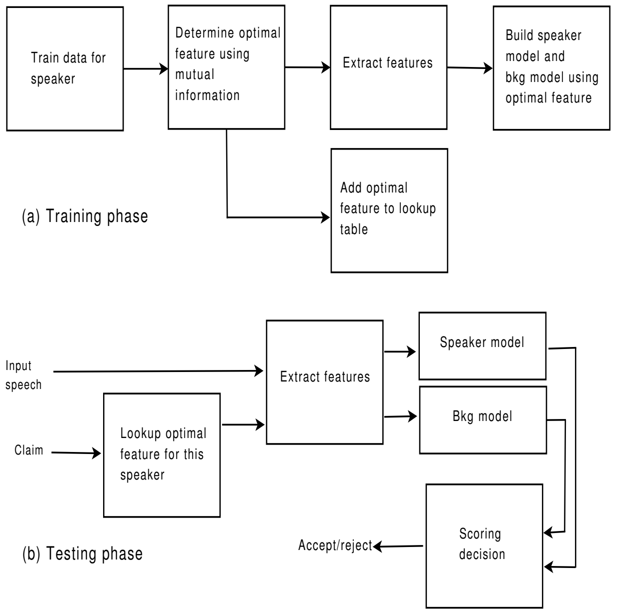
Train data for speaker
Determine optimal feature using mutual information
Extract features
Build speaker model and bkg model using optimal feature
Add optimal feature to lookup table
(a) Training phase
Input speech
Claim
Lookup optimal feature for this speaker
Extract features
Speaker model
Bkg model
Scoring decision
Accept/reject
(b) Testing phase

Figure 4: The proposed speaker verification system incorporating dynamic feature selection (a) training phase and (b) testing phase

6. PERFORMANCE EVALUATION

In this section, we describe the experimental setup and performance evaluation of the proposed speaker verification system incorporating dynamic feature selection.

6.1 The database

All verification experiments were performed with a subset of 200 speakers from the NTIMIT database [4]. There are four test cases for each speaker. For each test case, a verification test was done with a claim stating to be each of the 200 speakers, one after the other. This resulted in a total of 1,60,000 verification tests, out of which 800 are actually true (ie. claim should be accepted) and the rest false (ie. claim should be rejected.)

6.2 Experimental results

As a baseline, conventional speaker verification using MFCC and MODGDF individually is done on the dataset. The detection-error tradeoff (DET) [7] curves for this is shown in figure 5. The DET curves for the proposed speaker verification system is also given in Figure 5. From the DET curves, we see that the phase-based MODGDF gives better verification performance than MFCC. For the most part, the proposed method gives performance comparable to the MODGDF-based method. But at some operating points (towards the top left in the graph) the proposed method is distinctly better. It should also be noted in this context that the proposed method has advantages of lesser computational requirement when compared to acoustic feature fusion methods like early fusion.

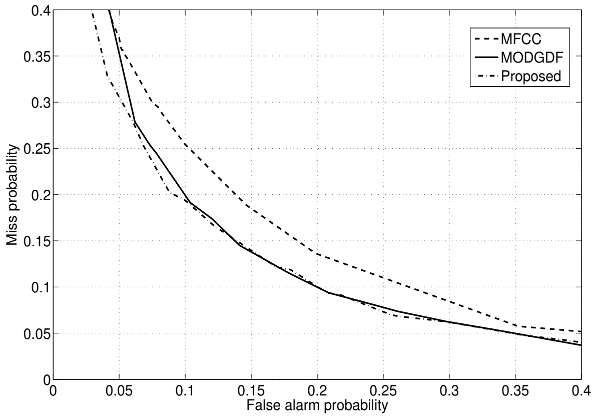

Figure 5: DET curves for speaker verification system.

7. CONCLUSIONS

In this paper, a speaker verification system which dynamically performs feature-switching is described. Mutual information between an acoustic feature stream and the Fourier magnitude spectrum is proposed to to determine the optimal feature-speaker combination. We have investigated this approach with two complimentary acoustic feature streams, namely the magnitude spectrum based MFCC and the phase spectrum based MODGDF. The proposed system implements feature switching using the feature-speaker combination obtained from training data and uses the optimal feature based on the claimed speaker identity during the testing phase. The system shows better performance when compared to conventional verification systems using a single feature.

The advantages of the proposed method is its lesser computational requirement, when compared to methods like early fusion, which operate in a much higher dimensional feature space. Also the problems of empirical weight determination for likelihood fusion are not present. Currently we are exploring methods by which a larger or varied set of acoustic features can be utilised to tap the diversity present in more than two feature streams. Other information theoretic measures are also being investigated to obtain the optimal feature for each speaker.

REFERENCES

[1] F. Bimbot et al. A tutorial on text independent speaker verification. *EURASIP Journal of Applied Signal Processing*, 4:430–451, 2004.

[2] J. P. Campbell. Speaker recognition: A tutorial. *Proceedings of the IEEE*, 85:1437–1462, 1997.

[3] T. Eriksson, S. Kim, H.-G. Kang, and C. Lee. An information theoretic perspective on feature selection in speaker recognition. *IEEE Signal Processing Letters*, 12:500–503, 2005.

[4] W. M. Fisher. NTIMIT. Linguistic Data Consortium, Philadelphia, 1993.

[5] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde. Significance of joint features derived from the modified group delay function in speech processing. *EURASIP Journal of Applied Signal Processing*, 2007.

[6] R. M. Hegde, H. A. Murthy, and V. R. R. Gadde. Significance of the modified group delay feature in speech recognition. *IEEE Trans. Audio, Speech and Language Processing*, 15:190–202, 2007.

[7] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assesment of detection task performance. In *Proceedings of Eurospeech*, volume 4, pages 1895–1898, 1997.

[8] R. Rasipuram, R. M. Hegde, and H. A. Murthy. Incorporating acoustic feature diversity into the linguistic search space for syllable based speech recognition. In *Proceedings of EUSIPCO*, 2008.

[9] D. Reynolds and R. Rose. Robust text independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3:72–82, 1995.