Feature-switching: Dynamic feature selection for an i-vector based speaker verification system

Saranya M. S.[a], Padmanabhan R.[b], Hema A. Murthy[a]

[a]*Indian Institute of Technology Madras*
[b]*Indian Institute of Technology Mandi*

Abstract

Conventional speaker verification systems utilize information from different feature representations by means of fusion. In this paper, we propose an alternative technique which achieves a similar effect but utilizes a more effective feature selection technique.The underlying assumption of the method is that different speakers may be better represented, and hence better verified, in different feature spaces. This technique, which we term as feature-switching, performs verification using a feature representation most suitable to the speaker under consideration. Out of a possible set of candidate representations, the most optimal representation for a speaker is determined during enrollment. Then verification is performed using the optimal feature of the claimed speaker. Experimental evaluation of feature-switching is performed utilizing the classical GMM-UBM speaker verification system, as well as the i-vector-based verification system. Our results show that feature-switching achieves improved performance compared to conventional as well as fusion-based systems.

Keywords: feature selection, speaker verification, i-vector, GMM-UBM, total variability space

1. Introduction

Feature extraction is an important step in pattern recognition systems. For speech signals, feature extraction is a transformation from the acoustic space to a feature space. In text-independent speaker verification, the objective is to determine if two utterances (the enrollment utterance and the test utterance) are both spoken by a particular speaker. We expect that the transformation into the feature space effectively discriminates the utterances spoken by the speaker under consideration from those spoken by other speakers. Most speaker verification systems, however, apply the same transformation, no matter which speaker is being considered. In this paper, we explore a new paradigm which exploits the *diversity of information* present in different feature spaces for speaker verification. The underlying assumption is that different speakers may be better discriminated in different feature spaces. Hence, performance can be improved by utilizing the 'well-suited' feature space for each speaker. We term this technique *feature-switching* and the well-suited feature space as the *optimal feature space*.

Traditionally, the diversity of different feature transformations has been utilized by combining them. These include the so-called *early fusion*, which is a combination at the feature level, and *late fusion*, which is at the classifier (or decision) level. Combining the information from multiple feature transformations usually results in improved performance, albeit with an increase in system complexity. Feature-switching aims to utilize information from multiple feature representations in an unconventional manner. Early fusion

Email addresses: saranms@cse.iitm.ac.in (Saranya M. S.), padman@iitmandi.ac.in (Padmanabhan R.), hema@cse.iitm.ac.in (Hema A. Murthy)

systems typically work by concatenating feature vectors; hence, the resulting feature space is of higher dimensionality. This in turn, requires more data to effectively train statistical models. Late fusion requires individual systems to be developed and fused; in platforms with limited processing or storage space, this could be undesirable. A popular method for the late fusion of speaker verification systems is by using logistic linear regression [1, 2, 3]. The feature-switching technique attempts to get the benefit of multiple feature representations while reducing system complexity at the same time.

Most feature representations transform the speech signal into its spectral representation. The short-term Fourier transform is a complex quantity, with information present in both magnitude and phase spectra. It is known from linear system theory, that non-minimum-phase signals have different information in magnitude and phase spectra [4]. Several studies [5], [6], [7] have shown the complementary nature of magnitude and phase, and how combining feature vectors derived from each of them improves performance in various tasks. In this paper, we study the effectiveness of feature-switching for speaker verification using feature representations from magnitude-based and phase-based features. We perform feature-switching using the standard Mel-frequency cepstra (MFCC) [8], which is derived from short-term magnitude, and the modified group delay feature (MODGDF) [9], which is derived from the short-term phase. For each speaker, the better-suited of these two representations is determined beforehand. Then, feature-switching is applied for speaker verification by verifying some speakers using MFCC features, and others using MODGDF features.

We study feature-switching for speaker verification in the context of the classical Gaussian Mixture Model- Universal Background Model (GMM-UBM) system [10], and the more sophisticated i-vector based representation [11]. In both cases, our studies show that feature-switching improves verification accuracy, when compared to conventional systems which use only a single feature representation. In addition, feature-switching also shows an improvement over fusion systems.

The idea of feature switching for the GMM-UBM framework was initially proposed in [6], and was extended in [12] on older speech corpora. NTIMIT data was used in [6], and NIST 2003 speaker recognition evaluation (SRE) data was used in [12]. In this work, feature-switching is evaluated on the GMM-UBM framework, using the more challenging NIST SRE 2010 data [13]. Also, a new feature selection method is proposed to apply feature-switching on the i-vector framework, and experimental evaluation is performed on the same dataset.

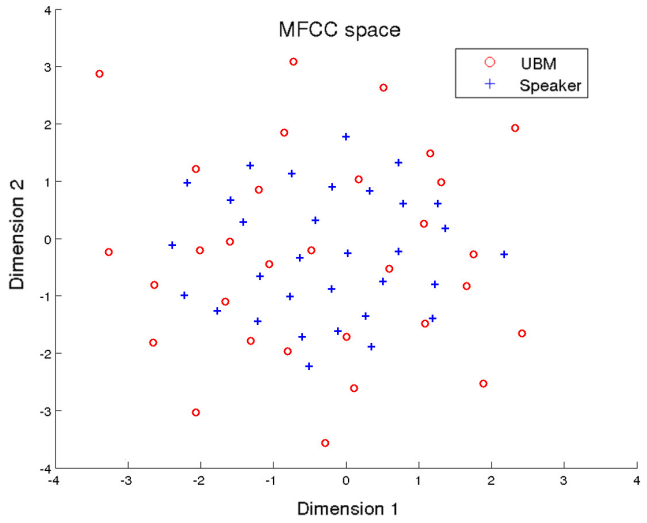
The rest of the paper is organized as follows: The effect of separability of features in different feature spaces is analysed in section 2. The process of selecting the optimal feature and feature-switching is explained in section 3. The candidate features used for feature selection is explained in section 4 followed by experimental evaluation in section 5. We conclude in section 6.

2. Separability analysis in different feature spaces

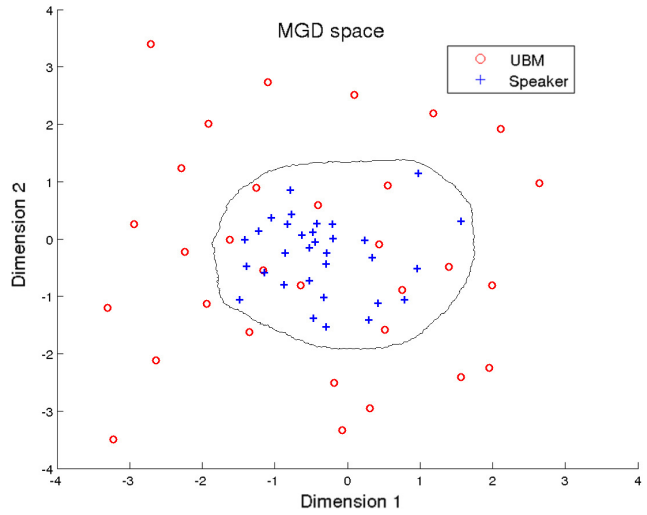
The underlying hypothesis for feature-switching is that representations of speakers are separated differently in different feature spaces. To study this, we perform separability studies in MFCC space and MODGDF space.

In the classical GMM-UBM framework [10], a speaker is represented by a GMM. Given feature vectors extracted from a speech utterance, the likelihood ratio of the speaker GMM and the UBM is computed. Better separation between the GMM and the UBM implies improved accuracy in verification.

Figures 1 and 2 illustrate the separability obtained in MFCC and MODGDF feature spaces for two speakers. In these figures, MFCC and MODGDF feature vectors are reduced to two dimensions using the Sammon mapping technique [14]. The mean vectors of a 32-component speaker GMM and UBM are plotted in two-dimensional space.

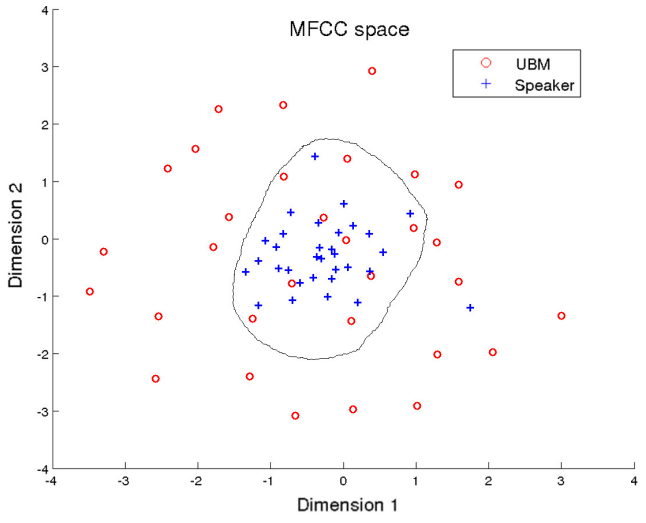


(a)

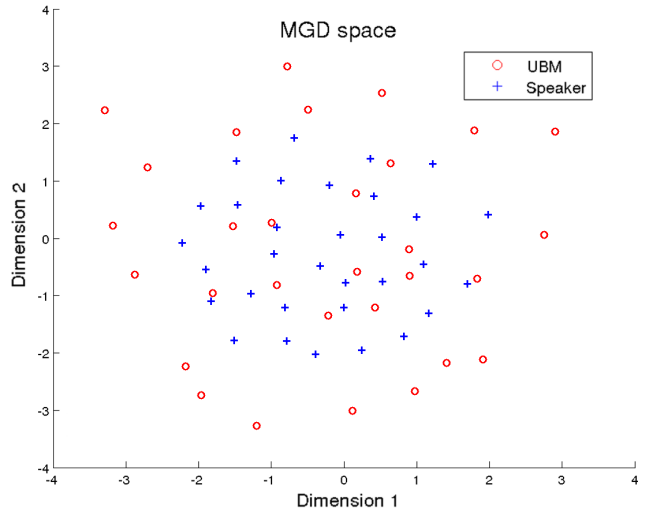


(b)

Figure 1: Sub-figures (a) and (b) show a speaker and background model centroids for a speaker in MFCC and MODGDF spaces. This speaker and the UBM are better separated in the MODGDF space.



(a)



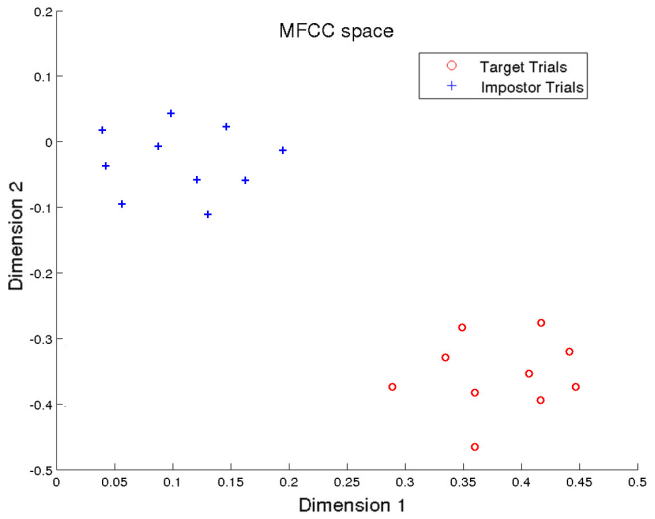
(b)

Figure 2: Sub-figures (a) and (b) show a speaker and background model centroids for a speaker in MFCC and MODGDF spaces. This speaker and the UBM are better separated in the MFCC space.

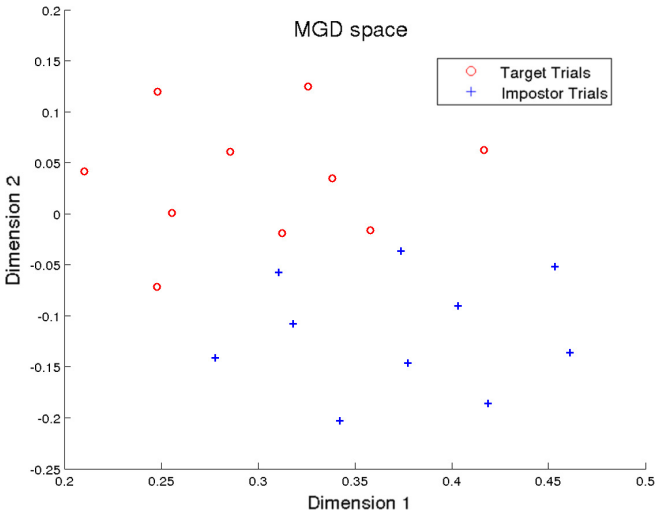
Sammon mapping represents high dimensional vectors in a lower dimensional space such that the geometric relations between the original data points are preserved as much as possible. The measure used by Sammon mapping is designed to minimize the differences between corresponding inter-point distances in the two spaces [15].

It can be seen that in Figure 1, there is better separation between the speaker GMM and the UBM in the MODGDF space when compared to the MFCC space. If this Sammon mapping reflects the properties of the higher dimensional space then, this speaker is better discriminated against the UBM in the MODGDF space. Similarly, for another speaker, the better separation is observed in the MFCC space (Figure 2).

A similar analysis is performed in i-vector space by considering i-vectors derived from different feature representations. For a given speaker, target trials are those spoken by the speaker himself or herself, and are also called true-speaker trials. Non-target trials are spoken by other speakers, and are also called impostor trials. 500-

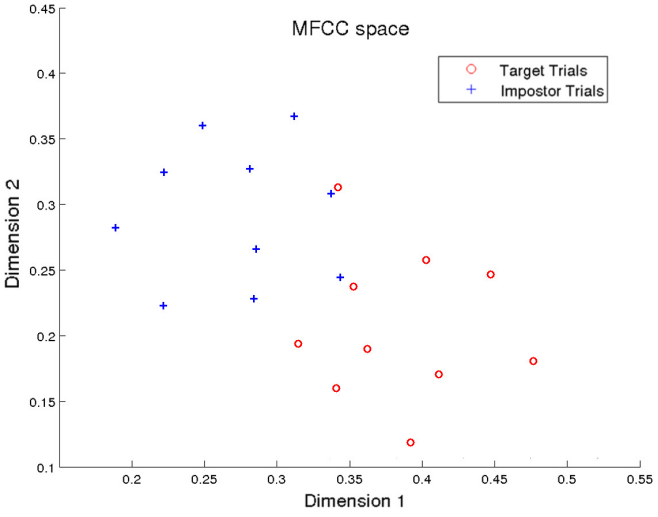


(a)

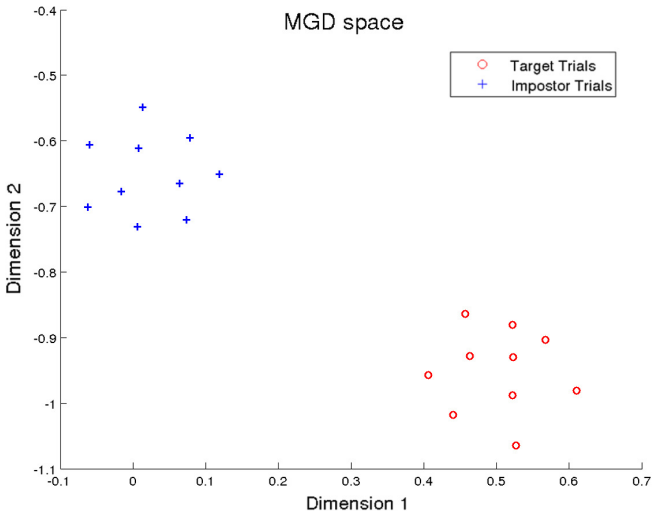


(b)

Figure 3: Sub-figures (a) and (b) show the i-vectors derived from MFCC and MODGDF for target and impostor trials. For this speaker, target and impostor i-vectors are better separated in the MFCC space.



(a)



(b)

Figure 4: Sub-figures (a) and (b) show the i-vectors derived from MFCC and MODGDF for target and impostor trials. For this speaker, target and impostor i-vectors are better separated in the MODGDF space.

dimensional i-vectors are reduced to two dimensions using Sammon mapping. The better separation of target and non-target i-vectors for two different speakers in MFCC and MODGDF space can be readily seen in Figures 3 and 4 respectively.

3. Optimal feature selection and feature-switching

Speaker verification is a two-class problem. A verification trial consists of a test utterance from an unknown speaker, and a speaker claim. Feature-switching can be naturally applied to the verification scenario by performing verification in the well-suited feature space of the claimed speaker. This well-suited feature representation is henceforth termed as the *optimal feature*. The optimal feature is determined for every speaker during enrollment and stored in a look-up table. During testing, the optimal feature of the claimed speaker is looked-up, and verification is performed in the optimal feature space. The overall architecture of the feature-switching system is shown

in Figure 5.

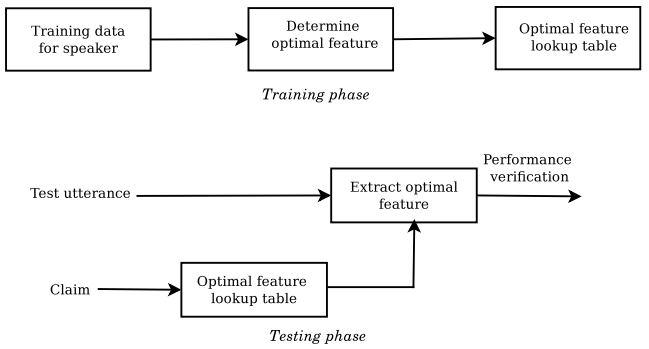


Figure 5: System architecture for training and testing phases in feature-switching.

3.1. Determining the optimal feature for the GMM-UBM framework

For the GMM-UBM framework, the method of determining the optimal feature for a particular speaker, given a set of candidate feature representations, was described in [12]. The optimal feature is determined by evaluating the representation ability and discrimination ability of each candidate feature representation. Given an enrollment utterance, the mutual information between extracted feature vectors and the complex Fourier transform (CFT) is used as an estimate of the information captured by the feature vectors. Thus, the representation ability of the feature representation is given as

$$\mathrm{mi}(\mathrm{C}, X), \tag{1}$$

where mi represents the mutual information, C is the complex short-time Fourier transform representation of an utterance, and X is a set of feature vectors computed from the utterance.

The mutual information (MI) between two random variables is a measure of how much information about one variable can be obtained, given knowledge of the other [16]. In [17] the connection between MI and classification accuracy for a speaker verification system is illustrated. In [16, 18] MI is used to choose appropriate feature representations from several possible options.

The CFT is a representation of the acoustic signal. The CFT representation is the basis for deriving information from the magnitude or phase spectra. MFCCs are derived from the magnitude spectrum, and the MODGDF features are derived from the phase spectrum. Thus, viewing the CFT, MFCC, and MODGDF as alternate representations of the acoustic signal, the MI between CFT and MFCC is an approximation of how much information is captured from the magnitude spectrum. Similarly, the MI between CFT and MODGDF is an approximation of the information captured from the phase spectrum. Thus, whenever spectral features are used, the MI between the CFT representation and a feature representation X is a metric to measure the information captured in the feature space from the acoustic space.

As in Equation (1), if C represents a collection of N-point CFT vectors and X represents a collection of feature vectors (MFCC or MODGDF), derived from the same utterance, the MI between C and X is estimated as in Algorithm 1 [19].

A speaker model is said to be more discriminative if it is well separated from other speaker models. In the context of GMM-UBM speaker verification, the "other" speaker model represents the alternate hypothesis (or in other words, the UBM.) Intuitively, if some of the components of a speaker model are well-separated from the UBM in a particular feature space, then that speaker model is better discriminated from most of the other speakers' models in that feature space. On the other hand, if the speaker model is close to the UBM, then the speaker model will not have much speaker-specific information.

The discrimination ability is determined by estimating the Kullback-Leibler divergence (KL-divergence) between the UBM (λ_{ubm}) and the speaker GMM (λ_{spk}) adapted from it. Because of the one-to-one correspondence between the mixture components of the background model and the speaker model, the KL-divergence can be approximated.

Algorithm 1 Mutual information calculation

Input: CFT of a speech signal: $C = \{\mathbf{c_i}\}$, and feature vectors of the speech signal : $X = \{\mathbf{x_i}\}$ where $i = 1, 2, \cdots, M$ (M = number of frames).

Output: Mutual information between C and X.

1: Vector quantize the set C to form a codebook A. Vector quantize the set X to form a codebook B. Let both codebooks have P centroids.

2: Let $\hat{C}_j$and $\hat{X}_k$ denote centroids in A and B with $j = \{1, 2, \cdots, P\}$ and $k = \{1, 2, \cdots, P\}$. The relative frequency of each centroid is an approximate measure of the probability of occurrence of that centroid. The mapping of an element $\mathbf{c_i}$ to a codevector $\hat{C}_j$ is denoted by $Q(\mathbf{c_i}) = \hat{C}_j$.

$$P(\hat{C}_j) = \frac{|i \ : \ Q(\mathbf{c_i}) = \hat{C}_j|}{M} \tag{2}$$

$$P(\hat{X}_k) = \frac{|i \ : \ Q(\mathbf{x_i}) = \hat{X}_k|}{M} \tag{3}$$

where $|.|$ denotes the cardinality.

3: The joint probability of occurrences of the centroids $\hat{C}_j$ and $\hat{X}_k$ is given by the number of points belonging to the cluster pair $(\hat{C}_j, \hat{X}_k)$

$$P(\hat{C}_j, \hat{X}_k) = \frac{|i : Q(\mathbf{c_i}) = \hat{C}_j \text{ and } Q(\mathbf{x_i}) = \hat{X}_k|}{M} \tag{4}$$

4: Using Bayes rule, the conditional probability is obtained from the joint probability

$$P(\hat{X}_k|\hat{C}_j) = \frac{P(\hat{C}_j, \hat{X}_k)}{P(\hat{C}_j)} \tag{5}$$

5: From the probabilities, we can estimate the entropy of CFT : $H(C)$, entropy of X : $H(X)$ and average conditional entropy : $H(C|X)$ as follows

$$H(C) = -E\left[log_2 P(C = \mathbf{c_i})\right] \tag{6}$$

$$H(X) = -E\left[log_2 P(X = \mathbf{x_i})\right] \tag{7}$$

$$H(X|C) = \sum_{\mathbf{c}} P(C = \mathbf{c}) H(X|C = \mathbf{c}) \tag{8}$$

6: Compute *mi(C,X) = H(X) − H(X|C).*

As in [20], for two unimodal Gaussian distributions $\hat{f}$ and $\hat{g}$, the KL-divergence has the closed form expression

$$\begin{aligned}\text{kld}(\hat{f}, \hat{g}) = \frac{1}{2}\Bigg[\log\frac{|\Sigma_g|}{|\Sigma_f|} + \text{Tr}|\Sigma_g^{-1}\Sigma_f| - d + \\ (\mu_f - \mu_g)^T\Sigma_g^{-1}(\mu_f - \mu_g)\Bigg],\end{aligned} \tag{9}$$

where $\hat{f} = \mathcal{N}(\mu_f, \Sigma_f)$ and $\hat{g} = \mathcal{N}(\mu_g, \Sigma_g)$.

For multi-modal speaker models λ_{spk}, whose means $\mu_{\text{spk},i}$ are adapted from the means $\mu_{\text{ubm},i}$ of the UBM model λ_{ubm} (the covariances and mixture weights are same as that of the UBM), the KL-divergence is approximated by [21]

$$\text{kld}(\lambda_{\text{spk}}, \lambda_{\text{ubm}}) \approx \sum_{i=1}^{K} \pi_i \ \text{kld}(f_i, g_i), \tag{10}$$

where,

$$\lambda_{\text{spk}} = \sum_{i=1}^{K} \pi_i \, f_i,$$

$$\lambda_{\text{ubm}} = \sum_{i=1}^{K} \pi_i \, g_i,$$

$f_i = \mathcal{N}(\mu_{\text{spk},i}, \Sigma_i)$, and
$g_i = \mathcal{N}(\mu_{\text{ubm},i}, \Sigma_i)$.
π_i are the mixture weights and i varies from 1 to K, the number of mixture components. Here, f_i and g_i are the corresponding unimodal Gaussian distributions.

The optimal feature for a particular speaker is determined from the combined representative and discriminative measures of each of the P candidate features. For the p-th feature representation, we determine

$$\theta_p = \text{mi}(\text{C}, X_p),$$
$$\gamma_p = \text{kld}(\lambda_{\text{spk},p}, \lambda_{\text{ubm},p}),$$

where X_p are feature vectors, the speaker model $\lambda_{\text{spk,p}}$ and UBM $\lambda_{\text{ubm,p}}$ are in the p-th feature space, and p ranges from 1 to P.

A linear combination of these two measures is determined as

$$\phi_p = \alpha\theta_p + (1 - \alpha)\gamma_p, \tag{11}$$

where α is a weighting parameter determined experimentally. The optimal feature $\hat{p}$ for a given speaker is determined as

$$\hat{p} = \underset{p}{\text{argmax}}\{\phi_p\} \tag{12}$$

3.2. Determining the optimal feature in the i-vector framework

The i-vector representation [11] is a fixed-length representation of speech utterances. Given an $FM \times 1$ supervector of means, μ derived from a UBM, a speaker and recording specific supervector s is assumed to be of the form

$$s = \mu + Tw. \tag{13}$$

Here, the acoustic feature vector is F-dimensional, the UBM has M components, T is an $FM \times D$ low-rank matrix, and w is a $D \times 1$ latent vector with a standard normal distribution $w \sim \mathcal{N}(0, I)$. The i-vector is estimated as the mean of the posterior distribution of w, given the utterance. Procedures to estimate the hyperparameters μ and T, and estimate i-vectors from an utterance can be found in [11].

The i-vector representing an utterance includes information about the speaker and the channel. To compensate for unwanted channel effects, several preprocessing steps like length normalization [22], and within-class covariance normalization (WCCN) [23] are performed. A popular method to measure the similarity between two i-vectors is the cosine distance [11].

To apply feature switching in the i-vector framework, given an utterance, i-vectors are estimated from different acoustic feature vectors and their associated hyperparameters. The better-suited i-vector representation for an enrollment speaker is determined by comparing the speaker's i-vector with the i-vectors of N other speakers from a held-out dataset. In our experiments, NIST SRE 2008 [24] is used as the held-out dataset to choose the optimal features. 1270 male speakers and 1993 female speakers from the short2 training condition are chosen to estimate the optimal features for every enrolled speaker. The short2 training condition includes both telephone and microphone utterances. The optimal feature space $\hat{p}$ for the i-th speaker in the enrollment data is estimated as:

$$\hat{p} = \arg\min_{p}\{S_p\}, \tag{14}$$

where

$$S_p = \frac{\sum_{j=1}^{N} d(w_{p,i}, w_{p,j})}{N}. \tag{15}$$

Here, $w_{p,j}$ represents the i-vector for the j-th speaker from the held-out dataset extracted using the p-th feature representation. $w_{p,i}$ represents the i-vector for the i-th speaker from the enrollment dataset extracted using the p-th feature representation. d is a distance measure (for example, cosine similarity) between i-vectors. For the i-th speaker, the i-vector representation giving the minimum average similarity with the other speakers from the held-out dataset is used as the optimal feature representation. In summary, for feature-switching, different speakers are verified using i-vectors derived from different acoustic features.

4. Features from magnitude and phase spectra

The underlying assumption in feature-switching is that information in different feature representations can be utilized dynamically. Earlier studies [5], [6], [7] have demonstrated the complementary nature of the information in magnitude and phase spectra. Standard MFCC features are derived from the magnitude spectrum. A popular method for utilizing information from the phase spectrum is via group delay functions [25]. The modified group delay feature (MODGDF) [26], which is derived from the modified group delay function [27], has been explored as complementary features to MFCCs. The procedure for extracting MODGDF features is given in Algorithm 2 as in [9]. More details regarding the theory of MODGDF can be found in [26] and [27]. In our experiments the MODGDF parameters a, b and lifter_ω are set to 0.4, 0.9 and

Algorithm 2 MODGDF feature extraction

Input: A frame of speech $x(n)$

Output: MODGDF features $c(n)$

1: Compute the DFT of the speech frame $x(n)$ as $X(k)$.

2: Next, the DFT of the signal $n\,x(n)$ is computed as $\hat{X}(k)$.

3: Compute the cepstrally smoothed spectra of $X(k)$ and denote it as $S(k)$. The parameter lifter_ω is used to control the length of the window in the cepstral domain.

4: Compute the MODGD as:

$$\tau_m(k) = \left(\frac{\tau(k)}{|\tau(k)|}\right)(|\tau(k)|)^a$$

where

$$\tau(k) = \frac{X_R(k)\hat{X}_R(k) + X_I(k)\hat{X}_I(k)}{|S(k)|^{2b}}$$

and the parameters a and b are used to control the dynamic range of the MODGD.

5: Compute the MODGDF features by taking the DCT:

$$c(n) = \sum_{k=0}^{N_f-1} \tau_m(k)\cos(n(2k+1)\pi/N_f), \quad 0 \le n < N_c$$

where N_f is the DFT size and N_c are the number of cepstral coefficients.

8 respectively as mentioned in [9] (see Algorithm 2).

5. Experimental evaluation

This section details the experimental evaluation of speaker verification in the feature-switching framework. We give details about the dataset used, the development of the feature-switching system, and comparisons with baseline and fusion systems.

5.1. Development and evaluation data

Speaker verification experiments are performed on a subset of the NIST 2010 SRE dataset. The data contains telephone and microphone utterances under varying vocal effort, as detailed in [13] are used for enrollment and evaluation. These are summarised in Table 1. Gender-specific hyperparameters for the speaker recognition systems including the UBM and the T-matrix are developed using data from SRE99, SRE03, SRE04, SRE05, SRE06, SRE08, and SRE08-extended data.

5.2. Baseline verification systems

Feature-switching is performed on two speaker verification frameworks: the GMM-UBM system, and the i-vector system. The evaluation metric used is the equal error rate (EER), and is evaluated separately for male and female genders.

Voice activity detection (VAD): VAD is an important component in speech processing systems. In our systems, speech frames of 25 ms size, with a frame shift of 10 ms are utilized. Since the utterances are fairly clean, a simple VAD using a threshold on average short-term energy is utilized. This typically discards about 20-25% of the input frames.

GMM-UBM system: GMM-UBM systems are developed separately for MFCC and MODGDF feature representations. Conventional short-time feature vectors are extracted in each feature domain. Gender-dependent 1024-component UBMs are built from development data. Speaker-dependent GMMs are generated for the enrollment data by adapting the means of the top 10 maximum contributing mixture components of the UBM. For each test utterance, similarity scores are computed as the ratio of the log-likelihood of the extracted features with the speaker model and the UBM.

The baseline systems are denoted as follows. The names of the various systems compared are self-descriptive and individual systems are built for each gender.

1. UBM-MFC: Baseline GMM-UBM verification system with MFCC features
2. UBM-MGD: Baseline GMM-UBM verification system with MODGD features

Table 1: NIST 2010 conditions used in the evaluation

Condition	Channel	Training vocal effort	Testing vocal effort
C5	Telephone	Normal	Normal
C6	Telephone	Normal	High
C7	Microphone	Normal	High
C8	Telephone	Normal	Low
C9	Microphone	Normal	Low

i-vector system: Conventional short-time feature vectors are extracted and first and second order supervector statistics are computed using a 1024-component UBM. A total variability matrix of size 38912 $\times$ 500 is randomly initialized and estimated using development data, as detailed in [11, 28]. 500-dimensional i-vectors are estimated for each enrollment utterance and test utterance. Preprocessing steps to reduce channel variability including i-vector length normalization [29], linear discriminant analysis (LDA) and within-class covariance normalization (WCCN) are applied to the i-vectors. The LDA projection matrix is learned by utilizing speaker-specific recordings from the development data. Cosine similarity between enrollment and test i-vectors is utilized to determine the similarity score between them. The different baseline systems developed in this framework are referred to as ivec-MFC and ivec-MGD. As before, these are gender-specific.

In both the GMM-UBM and the i-vector frameworks, similarity scores calculated between the test and enrollment utterances are subjected to T-normalization (T-Norm) [30]. T-Norm also known as *test normalization* is performed during test phase. The log-likelihood ratio or the cosine similarity score (S) of a test utterance (X) is computed. 200 speakers from the NIST 2008 SRE having no overlap with the enrollment speakers in NIST 2010 are used as possible cohort speakers. For every test trial, out of these 200 speakers, 50 speakers with the highest likelihood scores on the same test utterance are chosen as final set of cohort speakers. The mean (μ_X) and standard deviation (σ_X) of these 50 scores are estimated and used to normalize the score (S) as

$$S_{tnorm} = \frac{S - \mu_X}{\sigma_X} \tag{16}$$

The EERs of these baseline systems are listed in Table 2.

Score-level fusion: Score-level fusion, also called late fusion (LF) is achieved by fusing the scores of individual feature-based baseline systems [31]. Our preliminary experiments had shown that score fusion outperforms the feature-level fusion (early fusion). Hence further experiments are done only on late fusion. The fusion of scores are a linear combination of the MFCC and MODGDF scores, given as,

$$\mathrm{S_{lf}} = \beta \mathrm{S_{MFC}} + (1-\beta)\mathrm{S_{MGD}} \tag{17}$$

where $\mathrm{S_{lf}}$ is the late fusion score, $\mathrm{S_{MFC}}$ is the MFCC score and $\mathrm{S_{MGD}}$ is the MODGDF score. Since our aim here is to compare the performance of feature-switching with the best performing late fusion, the optimal weighting parameter β was estimated using the evaluation data. A search over the range of β values from 0 to 1 was done to obtain the best verification performance. These late fusion systems are denoted as LF-UBM and LF-ivec. The performance of the fusion systems is given in Table 2.

5.3. Feature-switching

In the proposed feature-switching framework, different speaker claims are verified using different feature representations. Experimental evaluation of feature-switching is performed in both GMM-UBM and i-vector frameworks. The details of these systems are given below.

Algorithm 3 Optimal feature selection for a speaker in GMM-UBM framework

Input: Mutual Information : θ_p, KL-Divergence : γ_p

Output: Optimal feature index $\hat{p}$.

procedure DETERMINEOPTIMALFEATURE(θ_p, γ_p) ▷ p - index of candidate feature spaces
  for each p in P **do** ▷ P ={MFCC, MODGDF}
    for each α in [0.0 to 1.0] **do**
      $A_\alpha = (\alpha * \theta_p) + (1 - \alpha) * \gamma_p$ ▷ A_α is the linear combination of θ_p and γ_p with weight α
      $\alpha = \alpha + 0.1$ ▷ α increases with a step size of 0.1
    end for
    $\phi_p = max\{A_\alpha\}$
  end for
  $\hat{p} = \underset{p}{\text{argmax}}\{\phi_p\}$
  return $\hat{p}$ ▷ $\hat{p}$ is the optimal feature space for this speaker
end procedure

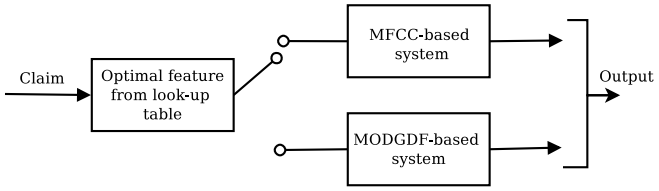


Figure 6: Testing phase of feature-switching system

GMM-UBM feature-switching system: The baseline systems (UBM-MFC and UBM-MGD) described in section 5.2 form the constituent systems for feature-switching. This is shown in Figure 6. For each enrollment speaker, the optimal feature is determined from the enrollment utterance, as described in Section 3.1. Here, the number of candidate features P is two, with $p = 1$ meaning MFCC features and $p = 2$ meaning MODGDF features (Equation 11).

The weighting parameter α is used as a trade-off between mutual information and KL-divergence, for determining the optimal feature of a given speaker (Equation 11). Different speakers can have different α values. Since there is no theoretical insight to determining the correct weight parameter, the following procedure is adopted for each of the N enrollment speakers. For a given speaker, the value of ϕ_p is determined across various values of α as,

$$\phi_p = \max\ (\alpha\theta_p + (1 - \alpha)\gamma_p), \tag{18}$$

where the values of α are varied from 0 to 1 in steps of 0.1. θ_p and γ_p are defined in Equation 11. Once ϕ_p is determined for $p = 1$ (MFCC) and $p = 2$ (MODGDF), the optimal feature is computed as in Equation 12. This procedure is summarised in Algorithm 3. The feature-switching system developed in this framework is denoted as FS-UBM.

i-vector feature-switching system: As in the case of the GMM-UBM system, the MFCC and MODGDF i-vector systems described in Section 5.2 are the constituent systems for the feature-switching system. For each speaker, the optimal feature representation is computed as described in Equation 14. The total number of target speakers (N) in male and female speaker verification systems are 2100 and 2651 respectively. The feature-switching systems developed in this framework are denoted as FS-ivec. The resulting EERs for these systems are also described in Table 2. Note that, for both GMM-UBM and i-vector-based systems, the optimal feature is determined for a speaker based on the enrollment data alone and is independent of the testing condition.

5.4. Result analysis

The various speaker verification systems described in Section 5 include four baseline systems (UBM-MFC, UBM-MGD, ivec-MFC, ivec-MGD), two late fusion systems (LF-UBM, LF-ivec) and two proposed feature-switching systems

Table 2: EERs (in %)for NIST 2010 male and female trials, conditions C5-C9

(a) Male trials using UBM-GMM

System	C5	C6	C7	C8	C9
UBM-MFC	11.1	12.1	10.6	8.0	4.1
UBM-MGD	14.0	14.4	7.4	9.7	5.6
LF-UBM	8.0	11.5	6.9	6.6	4.1
FS-UBM	**3.1**	**5.4**	**3.4**	**3.0**	**2.1**

(b) Female trials using UBM-GMM

System	C5	C6	C7	C8	C9
UBM-MFC	11.0	14.3	18.8	7.7	2.6
UBM-MGD	16.2	15.5	10.1	8.5	5.1
LF-UBM	11.0	11.5	10.1	7.6	2.5
FS-UBM	**8.9**	**6.8**	**8.5**	**5.3**	**2.0**

(c) Male trials using i-vector

System	C5	C6	C7	C8	C9
ivec-MFC	5.1	5.6	6.4	1.9	3.3
ivec-MGD	5.6	5.9	5.7	2.9	3.3
LF-ivec	4.5	5.4	2.9	1.9	2.6
FS-ivec	**3.5**	**3.6**	**2.7**	**1.6**	**2.2**

(d) Female trials using i-vector

System	C5	C6	C7	C8	C9
ivec-MFC	7.9	8.4	10.2	3.4	3.2
ivec-MGD	4.0	4.2	9.2	2.2	3.0
LF-ivec	3.7	4.2	9.2	2.2	2.4
FS-ivec	**3.2**	**3.5**	**6.0**	**1.6**	**1.9**

Table 3: Distribution of the speakers to MFCC and MODGDF optimal feature spaces.

Gender	MFCC	MODGDF	Total
FS-UBM system			
Male	668 (48%)	734 (52%)	1402
Female	1117 (64%)	641 (36%)	1758
FS-ivec system			
Male	1331 (94.9%)	71 (5.1%)	1402
Female	114 (6.5%)	1644 (93.5%)	1758

(FS-UBM, FS-ivec). Each of these has corresponding systems for male trials and female trials. The performance metric is the equal error rate (EER) [32]. The performance is tabulated in Table 2.

Compared to the best baseline systems, the score fusion systems provide an average relative improvement of 9.7% for male trials and 3.8% for female trials across all the conditions, in the GMM-UBM case. In the i-vector case, the average relative improvement of score fusion systems over the best baseline systems is 17.1% for male trials and 5.5% for the female trials across all the conditions.

The proposed feature-switching system outperforms the score fusion systems and the baseline systems in all the conditions. Compared to the best baseline systems, the average relative improvement across all the conditions is 58.6% for male trials and 28.3% for female trials in the GMM-UBM case and 33.8% and 27.1% in the i-vector case.

Comparison with other i-vector systems: The i-vector based speaker verification systems described in [33, 34] uses Probability Linear Discriminant Analysis (PLDA) and reports an EER of 3.57% and 3.59% for female speakers on test condition C5 respectively. For the same C5 test condition, [33] reports an EER of 2.86% for male data. In [35] EERs of 3.08% and 3.41% are observed for male and female speakers in C5 test condition with Gaussian PLDA (G-PLDA).

Comparison of the results in these systems must take into account the fact that the development data used by the systems and other configurations like number of UBM components, i-vector dimension, and final scoring methods are not the same.

Although the feature switching systems proposed in this work have performance similar to that of the above-cited systems, it has a significant improvement compared to that of the baseline systems as in Table 2.

To understand why feature-switching brings about improvements, we analyse the trials in the various evaluation conditions. In the baseline sys-

Table 4: The distribution of speaker trials in the feature spaces using feature-switching in GMM-UBM framework. EERs of the best baseline system and feature-switching systems are compared and the lower EER is in bold.

		GMM-UBM Framework			
Gender	Condition	No. of trials	Best Baseline EER and Feature Space	Feature Switching EER	No. of trials evaluated in MGD/MFCC (MGD%/MFCC%)
Male	C5	14065	11.1 (MFC)	**3.1**	3934/10131 (28.0/72.0)
	C6	12975	12.1 (MFC)	**5.4**	2904/10071 (22.4/77.6)
	C7	12938	7.4 (MGD)	**3.4**	3915/9023 (30.3/69.7)
	C8	11116	8.0 (MFC)	**3.0**	2193/8923 (19.7/80.3)
	C9	10815	4.1 (MFC)	**2.1**	3281/7534 (30.3/69.7)
Female	C5	16317	11.0 (MFC)	**8.9**	2501/13817 (15.3/84.7)
	C6	15673	14.3 (MFC)	**6.8**	2400/13273 (15.3/84.7)
	C7	15398	10.1 (MGD)	**8.5**	4542/10856 (29.5/70.5)
	C8	17495	7.7 (MFC)	**5.3**	2747/14748 (15.7/84.3)
	C9	16716	2.6 (MFC)	**2.0**	4880/11836 (29.2/70.8)

Table 5: The distribution of speaker trials in the feature spaces using feature-switching in i-vector framework. EERs of the best baseline system and feature-switching systems are compared and the lower EER is in bold.

		i-vector Framework			
Gender	Condition	No. of trials	Best Baseline EER and Feature Space	Feature Switching EER	No. of trials evaluated in MGD/MFCC(MGD%/MFCC%)
Male	C5	14065	5.1 (MFC)	**3.5**	585/13480 (4.2/95.8)
	C6	12975	5.6(MFC)	**3.6**	739/12236 (5.7/94.3)
	C7	12938	5.7(MGD)	**2.7**	218/12720 (1.7/98.3)
	C8	11116	1.9 (MFC)	**1.6**	683/10433 (6.1/93.9)
	C9	10815	3.3 (MFC)	**2.2**	343/10472 (3.2/96.8)
Female	C5	16317	4.0 (MGD)	**3.2**	15720/597 (96.3/3.7)
	C6	15673	4.2 (MGD)	**3.5**	15176/497 (96.8/3.2)
	C7	15398	9.2 (MGD)	**6.0**	14686/712 (95.4/4.6)
	C8	17495	2.2 (MGD)	**1.6**	16846/649 (96.3/3.7)
	C9	16716	3.0 (MGD)	**1.9**	15945/771 (95.4/4.6)

tems, each evaluation trial gets verified in the same feature space (be it the GMM-UBM case or the i-vector case). Whereas in feature-switching, the trials get evaluated in the optimal feature space of the claimed speaker. This is summarised in Tables 4 and 5. In Table 4, the first entry states that the best baseline EER of 11.1% is achieved when all the trials are evaluated in MFCC space. But in feature-switching, based on the optimal feature of the claimed speaker, 10131 trials were evaluated in the MFCC space and 3934 trials were evaluated in the MODGDF space. This differential evaluation resulted in a higher number of correct verifications, hence bringing down the EER.

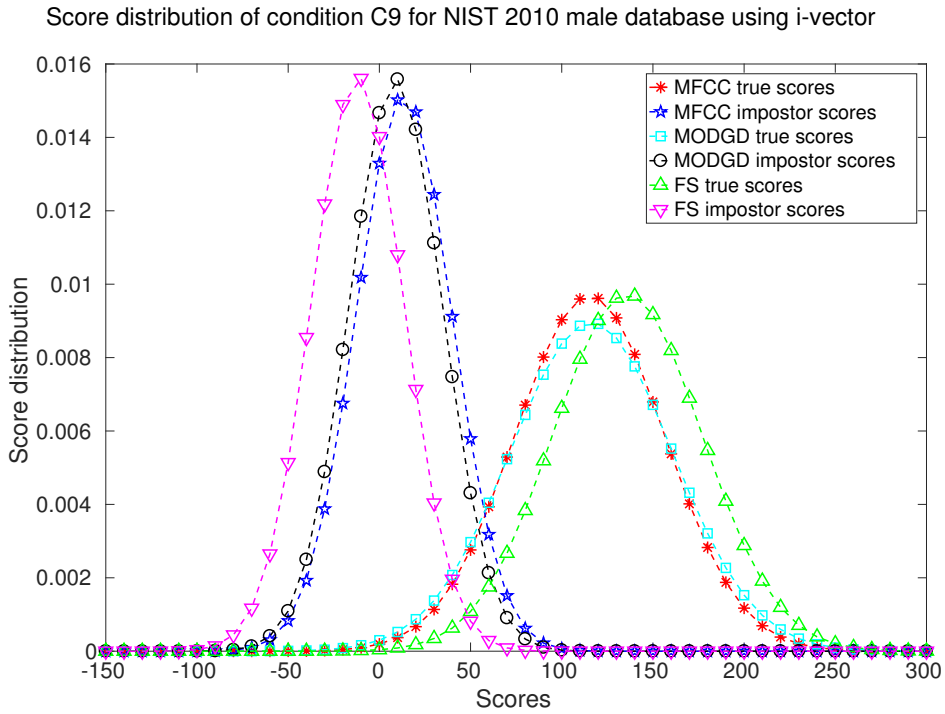


Figure 7: Score distribution of baseline systems (MFCC and MODGD) and feature-switching (FS) system for male database in test condition C7 using i-vector case.

The other entries in Tables 4 and 5 give details of the other cases. Figure 7 shows the score distribution comparison of true and impostor trials for condition C7, for the baseline systems and feature-switching system for male trials for the i-vector case.

The distribution of enrollment speakers among the optimal feature spaces are given in Table 3. The optimal feature of a given speaker does not need be the same for the GMM-UBM feature-switching system and the i-vector feature-switching system. In the GMM-UBM feature-switching system, male speakers and female speakers are more or less evenly split in the different feature spaces. Whereas, in the i-vector case, most of the male speakers get MFCC as their optimal feature and females get MODGDF. These two verification systems are based on different principles, and hence, it is difficult to compare them directly. Steps like LDA and WCCN may provide more discriminative abilities to the i-vector framework. However, the results give strong evidence that the differential evaluations of different trials can improve the performance.

6. Conclusion

In this paper, we developed the paradigm of feature-switching to perform text-independent speaker verification. By performing verification in a feature space that is well-suited to the speaker under consideration, improvements in accuracy are obtained. The method is evaluated using the classical GMM-UBM speaker verification framework, as well as the i-vector framework on NIST SRE 2010 data. Once the well-suited feature representation of a given speaker is determined, verification of that speaker can be performed in that feature space. Our experimental evaluation demonstrates that feature-switching provides benefits above that of conventional system fusion. On the NIST 2010 SRE dataset, an average improvement of 43.5% and 30.4% is attained in the GMM-UBM and i-vector cases respectively.

In principle, the method of feature-switching can be applied to any verification task; for example, to face verification. In this paper, we have applied feature-switching between two feature representations, which are derived respectively from magnitude and phase, and are known to be com-

plementary. The number of feature representations can be larger. Future research directions can include a more robust procedure to determine the optimal feature for a given speaker. An extended version of this might be to have customized feature representations for every class under consideration, and feature-switching between them during verification.

7. Acknowledgment

The authors would like to acknowledge the Defence Research and Development Organization, India for funding the research under the project CSE1314142DRDOHEMA.

References

[1] L. Burget, M. Fapso, V. Hubeika, O. Glembek, M. Karafiát, M. Kockmann, P. Matejka, P. Schwarz, J. Cernockỳ, BUT system for NIST 2008 speaker recognition evaluation, in: 10th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2009, pp. 2335–2338.

[2] L. Ferrer, M. Graciarena, A. Zymnis, E. Shriberg, System combination using auxiliary information for speaker verification, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008, pp. 4853–4856.

[3] O. Plchot, S. Matsoukas, P. Matejka, N. Dehak, J. Z. Ma, S. Cumani, O. Glembek, H. Hermansky, S. H. R. Mallidi, N. Mesgarani, et al., Developing a speaker identification system for the darpa rats project., in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2013, pp. 6768–6772.

[4] A. V. Oppenheim, R. W. Schafer, J. R. Buck, Discrete-time Signal Processing, Prentice-Hall, 2000.

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

[6] R. Padmanabhan, R. Hegde, H. Murthy, Dynamic selection of magnitude and phase based acoustic feature streams for speaker verification, in: 17th European Signal Processing Conference (EUSIPCO), 2009, pp. 1244–1248.

[7] P. K. Md. Jahangir Alam, T. Stafylakis, Combining amplitude and phase-based features for speaker verification with short duration utterances, in: 16th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2015.

[8] P. Mermelstein, S. B. Davis, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Transactions on Audio, Speech and Language Processing 28 (1980) 357–366.

[9] R. Hegde, H. Murthy, V. Gadde, Significance of the modified group delay feature in speech recognition, IEEE Transactions on Audio Speech and Language Processing 15 (1) (2007) 190–202.

[10] D. A. Reynolds, T. F. Quatieri, R. B. Dunn, Speaker verification using adapted Gaussian mixture models, Digital Signal Processing 10 (2000) 19–41.

[11] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, P. Ouellet, Front-end factor analysis for speaker verification, IEEE Transactions on Audio, Speech and Language Processing 19 (4) (2011) 788–798.

[12] P. Rajan, H. A. Murthy, Acoustic feature diversity and speaker verification, in: 11th Annual Conference of the International Speech Communication Association (INTERSPEECH), Mukahari, Japan, 2010.

[13] The NIST year 2010 speaker recognition evaluation plan, `http://www.itl.nist.gov/iad/mig/tests/sre/2010/` (2010).

[14] J. W. Sammon, A nonlinear mapping for data structure analysis, IEEE Transaction on Computers 18 (1969) 401–409.

[15] P. Henderson, Sammon mapping, Pattern Recognition Letters 18 (11-13) (1997) 1307–1316.

[16] D. P. Ellis, J. A. Bilmes, Using mutual information to design feature combinations., in: 1st Annual Conference of the International Speech Communication Association (INTERSPEECH), 2000, pp. 79–82.

[17] T. Eriksson, S. Kim, H.-G. Kang, C. Lee, An information-theoretic perspective on feature selection in speaker recognition, IEEE Signal Processing Letters 12 (7) (2005) 500–503.

[18] R. Battiti, Using mutual information for selecting features in supervised neural net learning, IEEE Transactions on Neural Networks 5 (4) (1994) 537–550.

[19] R. Padmanabhan, Studies on voice activity detection and feature diversity for speaker recognition, Ph.D. thesis, Indian Institute of Technology Madras (Aug 2012).
URL `http://lantana.tenet.res.in/website_files/thesis/Phd/pad_thesis.pdf`

[20] J. R. Hershey, P. A. Olsen, Approximating the Kullback Leibler Divergence Between Gaussian Mixture Models, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 4, IEEE, 2007, pp. IV–317–IV–320. doi:10.1109/icassp.2007.366913.
URL `http://dx.doi.org/10.1109/icassp.2007.366913`

[21] W. M. Campbell, D. E. Sturim, D. A. Reynolds, A. Solomonoff, Svm based speaker verification using a gmm supervector kernel and nap variability compen-

sation, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Vol. 1, 2006.

[22] D. Garcia-Romero, C. Espy-Wilson, Analysis of i-vector length normalization in speaker recognition systems, in: 12th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2011, pp. 249–252.

[23] A. O. Hatch, S. S. Kajarekar, A. Stolcke, Within-class covariance normalization for svm-based speaker recognition., in: 7th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2006.

[24] The nist year 2008 speaker recognition evaluation plan, http://www.itl.nist.gov/iad/mig/tests/sre/2008/ (November 2008 cited on June 2015).

[25] H. Banno, J. Lu, S. Nakamura, K. Shikano, H. Kawahara, Efficient representation of short-time phase based on group delay, in: IEEE International Conference on Acoustics, Speech and Signal Processing, Vol. 2, 1998, pp. 861–864 vol.2. doi:10.1109/ICASSP.1998.675401.

[26] H. A. Murthy, R. M. Hegde, V. R. R. Gadde, The modified group delay feature: a new spectral representation of speech, in: 8th International Conference on Spoken Language Processing (ICSLP), 2004.

[27] H. Murthy, V. Gadde, The modified group delay function and its application to phoneme recognition, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Vol. 1, 2003, pp. I–68–71 vol.1.

[28] P. Kenny, G. Boulianne, P. Ouellet, P. Dumouchel, Joint factor analysis versus eigenchannels in speaker recognition, IEEE Transactions on Audio, Speech, and Language Processing 15 (4) (2007) 1435–1447.

[29] D. G. Romero, C. Y. E. Wilson, Analysis of i-vector length normalization in speaker recognition systems, in: 12th Annual Conference of the International Speech Communication Association (INTERSPEECH), 2011, pp. 249–252.

[30] C. Barras, J. Gauvain, Feature and score normalization for speaker verification of cellular data, in: IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) , Vol. 2, 2003, pp. II–49–52 vol.2. doi:10.1109/ICASSP.2003.1202291.

[31] T. Kinnunen, V. Hautamäki, P. Fränti, Fusion of spectral feature sets for accurate speaker identification, in: 9th International Conference on Speech and Computer (SPECOM), 2004.

[32] S. T. Corporation, Technical document about far, frr and eer, http://ftp.syris.com/SYRIS_ACS_DVD-ROM/UserGuideManual/Reader/SYRDF5-S2MS%20&%20SYRDF6-PMS/About%20FAR_FRR_EER.pdf (2004).

[33] L. Burget, O. Plchot, S. Cumani, O. Glembek, P. Matjka, N. Brmmer, Discriminatively trained probabilistic linear discriminant analysis for speaker verification, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 4832–4835.

[34] P. Matejka, O. Glembek, et.al, Full-covariance UBM and heavy-tailed PLDA in i-vector speaker verification, in: IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP, 2011, pp. 4828–4831.

[35] D. Garcia-Romero, C. Y. Espy-Wilson, Analysis of i-vector length normalization in speaker recognition systems, in: 12th Annual Conference of the International Speech Communication Association INTERSPEECH, 2011, pp. 249–252.