Acoustic Feature Diversity and Speaker Verification

R.Padmanabhan, Hema A. Murthy

Department of Computer Science and Engineering,
Indian Institute of Technology Madras, Chennai, India
rpappu2@gmail.com, hema@lantana.tenet.res.in

Abstract

We present a new method for speaker verification that uses the diversity of information from multiple feature representations. The principle behind the method is that certain features are better at recognising certain speakers. Thus, rather than using the same feature representation for all speakers, we use different features for different speakers. During training, we determine the *optimal feature* for each speaker from candidate features, by measuring information-theoretic criteria. During evaluation, verification is performed using the optimal feature of the claimed speaker. Experimental results with four candidate features show that the proposed system outperforms conventional systems that use a single feature or a combination of features.
Index Terms: speaker verification, feature selection

1. Introduction

Feature extraction is a crucial step in all pattern recognition systems. A number of feature extraction algorithms have been proposed in the literature, each having some advantage over that of the other. We can view the process of feature extraction from speech signals as a projection from the acoustic space into the feature space. We wish to capture the most *relevant* information for the given task from the acoustic space into the feature space. To achieve this, two widely used techniques are: (a) intelligent feature selection and (b) feature combination.

In this paper, we explore a new paradigm which exploits the *diversity of information* present in different feature spaces to achieve better classification. The underlying assumption of this approach is that using a single feature (or a combination of features) for all classes may not result in optimal performance, as some features may identify certain classes better than others. For some classification problems, the classifier can take advantage of the availability of additional information for feature selection. In this paper, we explore this paradigm of ***feature diversity*** in the context of speaker verification.

We perform the study using four different features, namely: (a) the standard Mel-frequency cepstra (MFCC) [1], (b) linear predictive cepstral coefficients (LPCC) [1], (c) Fourier transform phase based modified group delay features (MODGDF) [2] and (d) Mel-frequency slope (fSlope) [3]. Earlier studies with MFCC and MODGDF have shown that they have different efficiencies at recognising sub-word units [4]. Our preliminary investigations into feature diversity for speaker recognition have been reported in [5] and [6].

In the proposed speaker verification system, we first determine which feature (MFCC, LPCC, MODGDF or fSlope) is most effective at identifying each speaker. This is done by measuring the amount of representative and discriminative information captured by each feature. We then develop a speaker verification framework which always uses the "good" feature by means of a simple lookup procedure during verification. This results in using different features for verifying different speaker claims: some claims are verified with MFCC features, some with LPCC features, some with MODGDF and yet others with fSlope features.

The rest of the paper is organised as follows. In Section 2, we review some of the methods for feature selection and combination. In Section 3, we describe the method used to determine the effective feature for a given speaker. Section 4 explains the proposed speaker verification framework, and experimental results are discussed in Section 5. Finally, we conclude in Section 6.

2. Feature selection and combination

Information theoretic measures like mutual information (MI) have been useful in performing feature selection and combination. The measure of MI between a given class and a feature stream is a popular method used to select features well suited for that class [7], [8]. Typically, this problem is viewed as choosing a feature representation S, such that the mutual information between the class C and the feature, $I(C, S)$ is maximised [7]. In [8], studies were made tying the expected classification error to the mutual information between speaker identity and different feature representations. It was determined that the probability of error is minimised when the MI is maximised.

The loss of information during feature extraction can be overcome by combining information from different feature representations as follows:

At the acoustic feature level: Combination at the acoustic level is typically achieved by simply concatenating individual feature vectors. This results in feature vectors of higher dimension. Modelling and classification is now performed on the combined feature representation. This is also referred to as *early fusion.*

At the classification level: Combination at the classification level is achieved by developing separate classification mechanisms and combining their scores. Thus this involves developing individual supervised classification subsystems, the scores of which are combined into the composite score, which is used for making the final classification. This method is also referred to as *late fusion.*

Careful combination of features usually results in improved classification performance when compared to using individual features separately.

3. Feature diversity

The process of feature extraction can be viewed as a channel coding process, where the information content of the signal has to be encoded efficiently for the specific task. It is well estab-

lished that formant centre frequencies can vary by 25% between speakers due to differences in vocal tract length. Other physiological speaker dependent properties include vital capacity, phonation coefficient and glottal air flow. Learned characteristics, including speaking rate, prosodic effects and dialect are also speaker dependent [9]. In the channel coding paradigm, this means there is considerable variation in the information source. It is thus reasonable to assume that use of the same channel matrix (ie. the feature extraction mechanism) for encoding all speakers is not necessarily optimal.

For classification tasks, one must consider two aspects of feature representation: (a) the ability to capture maximum information from the acoustic space into the feature space (representative property) and, (b) the ability to discriminate between different classes (discriminative property).

3.1. Using mutual information to measure representative property

We use mutual information (MI) to quantify the amount of information captured from the acoustic space to the feature space. All the features considered in this paper are derived from spectral representations, and the MI between them is a measure of how much information is captured among them. We thus measure the MI between the complex short-time Fourier spectrum (CFFT, which represents the signal in the acoustic space) and the individual feature streams (representing the signal in the respective feature space.)

Thus, the representation ability of a feature representation X is computed as:

$$\text{mi}(\text{CFFT}, X) \tag{1}$$

The feature representation that gives the maximum mutual information has the best representation ability.

3.2. Using KL-divergence to measure discriminative property

A feature that efficiently captures information need not be efficient in discriminating between them. The discriminative property of a feature representation is a measure of the inter-class separability. The Kulback-Leibler divergence (KL-divergence) or relative entropy is a measure of the distance between two probability distributions. For two Gaussians $\hat{f}$ and $\hat{g}$, the KL-divergence has the closed form expression

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

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

For Gaussian mixture models (GMMs), the KL-divergence has no closed form expression. Many contemporary speaker verification systems build Gaussian mixture speaker models by adaptation of a universal background model (UBM.) For adapted speaker models, there is a one-to-one correspondence between the component mixtures of the speaker model and the UBM. Moreover, when only the means are adapted (ie. the mixture weights and covariances of the speaker model and the UBM remain the same), we see from (2) that the KL-divergence between the speaker model λ_{spk} and the background model λ_{UBM} is proportional to the linear weighted squared Mahalanobis distance. In this case, the KL-divergence becomes

$$\text{kld}(\lambda_{\text{spk}}, \lambda_{\text{UBM}}) = \Sigma_i\, \pi_i\, \text{kld}(f_i, g_i) \tag{3}$$

where $\lambda_{\text{spk}} = \Sigma_i\, \pi_i\, f_i$ and $\lambda_{\text{UBM}} = \Sigma_i\, \pi_i\, g_i$ and $f = \mathcal{N}(\mu_{\text{spk}}, \Sigma)$ and $g = \mathcal{N}(\mu_{\text{UBM}}, \Sigma)$ and π are the mixture weights.

The KL-divergence between the speaker model and the background model is a measure of the discriminability between the target speaker and imposter. The feature representation that gives higher KL-divergence better separates the speaker model from imposters.

3.3. Application of feature diversity for speaker verification

Preliminary experiments with equations (1) and (3) reveal that different features have different efficiencies at representing and discriminating speakers. We now apply the above conjecture to speaker verification.

A feature representation that efficiently represents a given speaker, as well as discriminates against other speakers is termed an **optimal feature** for that speaker. The optimal feature for a speaker can be determined (from a list of candidate features) at enrolment time or by using development data.

In a speaker verification system, we have a claimed identity along with the test utterance and we compute the score only for the claimed model. We can hence use the optimal feature of the claimed speaker, as this is will result in better modelling and discrimination using the speaker GMM. This results in **dynamically using different features for different claims** or *feature-switching*. Thus we make use of the diversity of information in different feature spaces. Moreover, in the speaker verification context, we avoid the problem of mismatched numerical ranges as there is no need to compare scores from different feature streams.

When compared to the conventional methods of feature combination like early fusion and late fusion, feature switching is more efficient in terms of computation and storage. There is neither a need to work in a high dimensional feature space as in early fusion, nor a need to develop separate classification subsystems as in late fusion. Thus feature switching can bring about performance comparable to conventional feature combination systems.

3.4. Determining the optimal feature for a speaker

To determine the optimal feature for a speaker, we measure the representative property as well as the discriminative property of the candidate features under consideration (MFCC, LPCC, MODGDF and fSlope in this case.) To give more weightage to one property over the other, the weighting factor α is introduced. Thus, the optimal feature is a threshold based function of the linear combination of the representative property and the discriminative property.

The optimal feature of a given speaker can be determined from the training/development data of the speaker. For a given speaker, we define the representative function θ_i and the discriminative function γ_i for feature representation i:

$$\begin{aligned}\theta_i &= \text{mi}(\mathcal{X}, \mathcal{Y}_i)\\ \gamma_i &= \text{kld}(\lambda_{\text{spk},i}, \lambda_{\text{UBM},i})\end{aligned}$$

where $i \in \{\text{MFCC}, \text{LPCC}, \text{MODGDF}, \text{fSlope}\}$, $\mathcal{X}$ represents the complex Fourier spectrum, $\mathcal{Y}_i$ represents the i th feature rep-

resentation, $\lambda_{\text{spk},i}$ is the speaker model and $\lambda_{\text{UBM},i}$ is the background model, using the i th feature representation.

The optimal feature function ϕ_i is defined as a linear combination of θ_i and γ_i:

$$\phi_i = \alpha\theta_i + (1-\alpha)\gamma_i \tag{4}$$

where the weighting factor α is used to emphasise the representative or discriminative measure.

The optimum feature stream $\hat{i}$ is now selected as

$$\hat{i} = \arg\max_i \{\phi_i\} \tag{5}$$

4. Speaker verification framework

Figure 1 shows the architecture of the proposed speaker verification framework. In the training phase, the optimal feature is determined for each speaker using the optimal feature function (5). The ⟨speaker, optimal feature⟩ pair is stored in a lookup table (LUT), which is indexed by speaker identity. The LUT contains an entry for each of the registered speakers in the system. Different parameters of (5) result in different LUTs for the same set of registered speakers.

In the evaluation phase, the optimal feature of the claimed speaker is determined from the lookup table. The optimal features are extracted from the input speech waveform. The TNorm score [12] is computed against the corresponding models and the verification decision is made. This results in the verification system performing feature switching, by extracting different features for different claims.

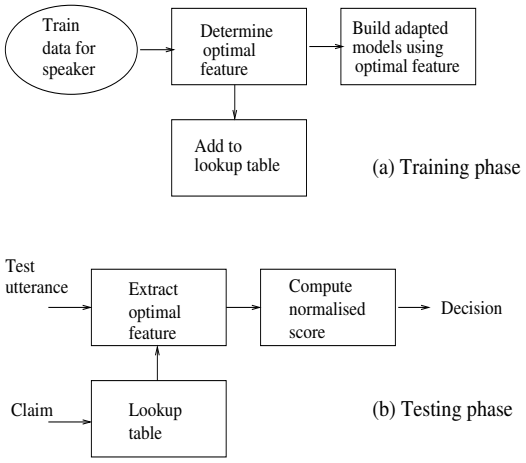

Figure 1: The proposed speaker verification system incorporating feature switching. (a) Training phase and (b) testing phase.

5. Experiments

We evaluate the proposed speaker verification system using feature switching and compare the performance to conventional systems that use only a single feature representation, as well as joint feature representation (early fusion.)

5.1. Database used for the study

The database used in this study is the one-speaker detection task of the NIST 2003 speaker recognition evaluation [10]. There are 149 male speakers and 207 female speakers, with about 2 minutes of training data for each speaker. Each test utterance is about 30 seconds long. More details about the database can be found in [10].

5.2. Speaker verification system

The baseline speaker verification systems are developed as described in [11]. Four systems, each using MFCC, LPCC, MODGDF and fSlope respectively, were built. Systems utilising the early fusion of the above features were also built. In all systems, Gaussian mixture models, the means of which are adapted from a 1024-mixture background model, is used to represent each speaker. Separate background models were built for male and female speakers. Cepstral mean subtraction was done for channel compensation. More sophisticated channel compensation methods like RASTA processing and feature mapping were not done. During evaluation, TNorm scores [12], using a cohort-set of 50 speakers were computed on the test utterance, against the claimed model. Detection error tradeoff (DET) curves for some of the baseline systems are shown in Figure 2 and the equal error rates (EER) are tabulated in Table 1.

The proposed speaker verification system incorporating feature switching is developed as follows. The optimal feature for each speaker is determined using (5). The parameters of feature extraction for the optimal feature is same as that of the respective baseline system. The weighting factor α represents a tradeoff between representative features and discriminative features. Speaker-dependent values of α were determined empirically and used to determine the optimal feature.

For representative measures, the mutual information between a 512-point complex DFT spectrum and each feature representation is computed as given in [5]. For discriminative measures, the KL-divergence between the speaker model and the background model, as given in (3), is computed for each feature representation. Using (5), the ⟨speaker, optimal-feature⟩ LUT is populated from training data.

During evaluation, the optimal feature for a claimed speaker is looked up from the LUT and extracted from the utterance. T-Norm scores are computed using the corresponding models and cohorts. In addition to reducing the variability of speaker scores, the score normalisation also brings scores from the two different feature streams into comparable range. Thus, target scores and imposter scores from the different feature streams can be pooled together. Feature-switched speaker verification systems were built with various subsets of the candidate features. The DET curves and EERs of these are given in Figure 2 and Table 1 respectively.

5.3. Results and discussions

From table 1, we observe that the baseline MODGDF based system shows better verification performance (EER of 11.92%) than the baseline MFCC system (EER of 13.58%.) This indicates the robustness of phase-based features, as described in [6]. Early fusion of feature representations generally improves verification performance. Various feature-switched systems, in general, give performance better than the early fusion systems. Feature switching between LPCC, MODGDF and fSlope gives the best EER of 11.33%. These experiments demonstrate the effectiveness of feature switching.

From the joint feature results in table 1, we find that arbitrary early fusion of features (MFCC-MODGDF and MFCC-fSlope) actually degrades system performance when compared to the corresponding baseline systems. Feature switching between MFCC/LPCC and MFCC/fSlope does not improve per-

System	EER (%)
Baseline	
Baseline MFCC	13.58
Baseline LPCC	12.27
Baseline MODGDF	11.92
Baseline fSlope	11.99
Joint (early fusion)	
MFCC-MODGDF	13.21
MFCC-fSlope	12.04
MODGDF-fSlope	11.57
Feature-switching	
MFCC/LPCC	12.43
MFCC/fSlope	12.11
MFCC/LPCC/MODGDF	11.73
MFCC/MODGDF/fSlope	11.66
MFCC/MODGDF	11.63
MODGDF/fSlope	11.56
LPCC/MODGDF	11.56
MFCC/LPCC/MODGDF/fSlope	11.40
LPCC/MODGDF/fSlope	11.33

Table 1: Equal error rates of various speaker verification systems.

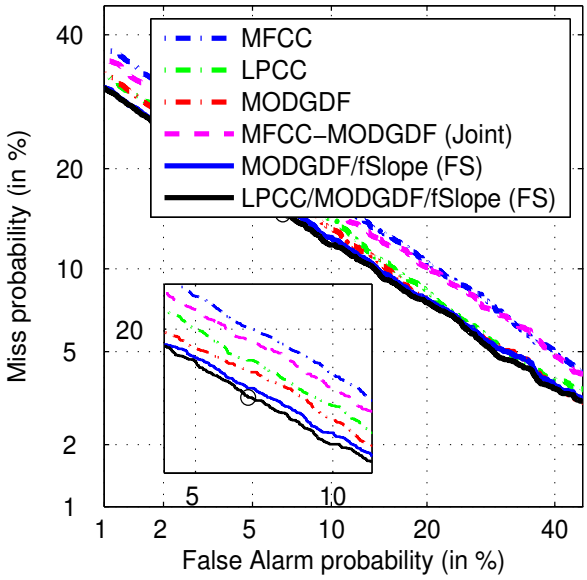

Figure 2: DET curves for baseline systems (MFCC, LPCC, MODGDF), early fusion system (MFCC-MODGDF) and feature switched systems (MODGDF/fSlope and LPCC/MODGDF/fSlope). The inset shows a zoomed-in version with the minimum decision cost function point.

formance, since both these features are derived from the short-time magnitude spectrum. Whereas feature switching with MFCC/MODGDF, MODGDF/fSlope gives better performance, as they combine information from both short-time magnitude and phase.

Using speaker dependent values of the weighting factor α results in better performance than using the same value of alpha for all speakers. This implies that for different speakers, the optimal feature has different representative/discriminative requirements. We also note that for feature switching to give effective results, the correct determination of the optimal feature is crucial. Finally, we note that if the LUT is made up of entirely one feature for all speakers, then the performance of the feature switched system falls back to that of the corresponding baseline system.

6. Conclusions and further work

This paper demonstrated the use of feature-switching for speaker verification. Experimental results reveal that feature-switching is an effective method for utilising information from multiple feature representations, without using conventional feature-level fusion methods. The experiments also show that feature-level fusion of classifiers improve performance only for certain combinations of features.

Different parametric forms of the same feature representation can be used as optimal features for different speakers (for example the centre frequencies and bandwidths of the Mel filters or the window scale factors for MODGDF features can be different for different speakers.) The present form of the optimal feature function (5) uses a tradeoff between the representative property and the discriminative property This need not always be the case. Investigations into these issues are currently being pursued.

7. References

[1] L. Rabiner and B. Juang, "Fundamentals of speech recognition", Pearson Education, 2003.

[2] 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, Lang. Process., vol. 15, pp. 190-202, Jan. 2007.

[3] H. A. Murthy, F. Beaufays, L. P. Heck, and M. Weintraub, "Robust text-independent speaker identification over telephone channels", IEEE Trans. Speech Audio Process. vol. 7, no. 5, pp. 554-568, 1999.

[4] R. Rasipuram, R. M. Hegde, and H. A. Murthy, "Significance of group delay based acoustic features in the linguistic search space for robust speech recognition," in Proc. European Signal Process. Conf., Laussane. Switzerland, 2008.

[5] P. Rajan, R. M. Hegde, and H. A. Murthy, "Dynamic selection of magnitude and phase based acoustic feature streams for speaker verification," in Proc. European Signal Process. Conf., Glasgow, Scotland, 2009, pp. 1244-1247.

[6] P. Rajan, S. H. K. Parthasarathi, and H. A. Murthy, "Robustness of phase based features for speaker recognition," in Proc. Interspeech, Brighton, England, 2009, pp. 2355-2358.

[7] R. Battiti, "Using mutual information for selecting features in supervised neural net learning," IEEE Trans. Neural Netw., vol. 5, pp. 537-550, Jul. 2002.

[8] T. Eriksson, S. Kim, H.-G. Kang, and C. Lee, "An information theoretic perspective on feature selection in speaker recognition", IEEE Signal Proc. Lett., vol. 12, pp. 500-503, Jul. 2005.

[9] J. P. Campbell, "Speaker recognition: a tutorial", Proc. IEEE, vol. 85, pp. 1437-1462, Sep. 1997.

[10] "The NIST year 2003 speaker recognition evaluation plan", Online: http://www.itl.nist.gov/iad/mig/tests/sre/2003/index.html, 2003.

[11] F. Bimbot, et. al. "A tutorial on text independent speaker verification," EURASIP Jnl. Applied Signal Process., vol. 4, pp. 430-451, 2004.

[12] R. Auckenthaler, M. Carey, and H. Lloyd-Thomas, "Score normalization for text-independent speaker verification systems", Digital Signal Process., vol. 10, pp. 42-54, 2000.