The IBM 2016 Speaker Recognition System

Seyed Omid Sadjadi[1], Sriram Ganapathy[2★], Jason W. Pelecanos[1]

[1]IBM Research, Yorktown Heights, NY, USA
[2]Dept. of Electrical Eng., Indian Institute of Science, Bangalore, India

sadjadi@us.ibm.com

Abstract

In this paper we describe the recent advancements made in the IBM i-vector speaker recognition system for conversational speech. In particular, we identify key techniques that contribute to significant improvements in performance of our system, and quantify their contributions. The techniques include: 1) a nearest-neighbor discriminant analysis (NDA) approach that is formulated to alleviate some of the limitations associated with the conventional linear discriminant analysis (LDA) that assumes Gaussian class-conditional distributions, 2) the application of speaker- and channel-adapted features, which are derived from an automatic speech recognition (ASR) system, for speaker recognition, and 3) the use of a deep neural network (DNN) acoustic model with a large number of output units ($\sim$ 10k senones) to compute the frame-level soft alignments required in the i-vector estimation process. We evaluate these techniques on the NIST 2010 speaker recognition evaluation (SRE) extended core conditions involving telephone and microphone trials. Experimental results indicate that: 1) the NDA is more effective (up to 35% relative improvement in terms of EER) than the traditional parametric LDA for speaker recognition, 2) when compared to raw acoustic features (e.g., MFCCs), the ASR speaker-adapted features provide gains in speaker recognition performance, and 3) increasing the number of output units in the DNN acoustic model (i.e., increasing the senone set size from 2k to 10k) provides consistent improvements in performance (for example from 37% to 57% relative EER gains over our baseline GMM i-vector system). To our knowledge, results reported in this paper represent the best performances published to date on the NIST SRE 2010 extended core tasks.

1. Introduction

There have been significant advancements in the speaker recognition field over the past few years. The research trend in this field has gradually evolved from joint factor analysis (JFA) based methods, which attempt to model the speaker and channel subspaces separately [1], towards the i-vector approach that models both speaker and channel variabilities in a single low-dimensional (e.g., a few hundred) space termed the total variability subspace [2]. State-of-the-art i-vector based speaker recognition systems employ universal background models (UBM) to generate frame-level soft alignments required in the i-vector estimation process. The i-vectors are typically post-processed through a linear discriminant analysis (LDA) [3] stage to generate dimensionality reduced and channel-compensated features which can then be efficiently

modeled and scored with various backends such as a probabilistic LDA (PLDA) [4, 5].

Until recently, Gaussian mixture models (GMM) trained in an unsupervised fashion (i.e., with no phonetic labels) were commonly used to represent the UBM in speaker recognition. However, inspired by the success of deep neural network (DNN) acoustic models in the automatic speech recognition (ASR) field, [6] proposed the use of DNN senone (context-dependent triphones) posteriors for computing the soft alignments, which resulted in remarkable reductions in speaker recognition error rates. The performance improvements reported in [6] are consistent with the observations made in our earlier effort [7] where a supervised GMM-HMM acoustic model (derived from an ASR system) was utilized to estimate the hyperparameters of a phonetically inspired UBM (PI-UBM) for speaker recognition. More recently, a supervised GMM-UBM (with full covariance matrices) based on DNN posteriors was also successfully evaluated for telephony speaker recognition [8]. These approaches are motivated by the fact that many of the speaker-dependent characteristics, which are conditioned on some phonetic units/classes, may be more effectively modeled using a UBM trained with explicit phonetic information.

In this paper, we report on the latest advancements made in the IBM i-vector speaker recognition system for conversational speech. Particularly, we first describe key components that contribute to significant improvements in performance of our system. These components include: 1) a nearest-neighbor based discriminant analysis (NDA) approach [9] for channel compensation in i-vector space, which, unlike the commonly used Fisher LDA, is non-parametric and typically of full rank, 2) speaker- and channel-adapted features derived from feature-space maximum likelihood linear regression (fMLLR) transforms [10, 11], which are used both to train/evaluate the DNN and to compute the sufficient Baum-Welch statistics for i-vector extraction, and 3) a DNN acoustic model with a large number of output units ($\sim$ 10k senones) to compute the soft alignments (i.e., the posteriors). To quantify the contribution of these components, we evaluate our system in the context of speaker verification experiments using speech material from the NIST 2010 speaker recognition evaluation (SRE) which includes 5 extended core conditions involving telephone and microphone trials.

2. System Overview

In the following subsections, we briefly describe the major components of our speaker recognition system. Specifically, we first provide an overview of GMM- versus DNN-based i-vector extraction, which is followed by algorithmic descriptions for the LDA and the NDA for channel compensation in the i-vector space. A schematic block diagram of the system is depicted in
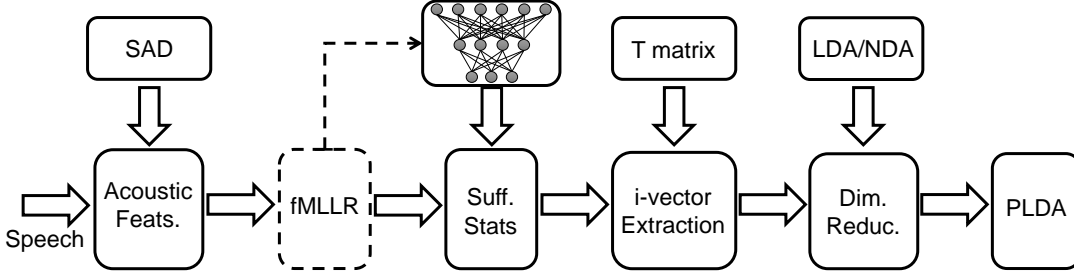

Figure 1: *Block diagram of the IBM speaker recognition system with fMLLR speaker- and channel-adapted features, DNN posterior based i-vectors, and NDA dimensionality reduction.*

Fig. 1.

2.1. I-vector extraction

The i-vector representation is based on the total variability modeling concept which assumes that speaker- and channel-dependent variabilities reside in the same low-dimensional subspace [2]. The key idea here is that variability within and across sessions can be described via a small set of parameters (a.k.a factors) in a low-dimensional subspace spanned by the columns of a low-rank rectangular matrix, $\mathbf{T}$, entitled the *total variability matrix*. Mathematically, the adapted mean supervector, $\mathbf{M}(s)$, for a given set of observations, s, can be modeled as,

$$\mathbf{M}(s) = \mathbf{m} + \mathbf{T}\,\mathbf{x}(s) + \boldsymbol{\epsilon}, \qquad (1)$$

where $\mathbf{m}$ is the prior mean supervector, $\mathbf{x}(s) \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a multivariate random variable termed an identity vector "i-vector", and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ is a residual noise term to account for the variability not captured via $\mathbf{T}$ ($\boldsymbol{\Sigma}$ is typically copied from the UBM). In other words, for the given observation set s, the i-vector represents the coordinates in the total variability subspace.

In order to learn the bases for the total variability subspace, one needs to compute the Baum-Welch statistics which are defined as,

$$N_g(s) \quad = \quad \sum_t \gamma_{tg}(s), \qquad (2)$$

$$\mathbf{F}_g(s) \quad = \quad \sum_t \gamma_{tg}(s)\,\mathbf{O}_t(s), \qquad (3)$$

where $N_g(s)$ and $\mathbf{F}_g(s)$ denote the zeroth- and first-order statistics for speech session s, respectively, with $\gamma_{tg}(s)$ being the posterior probability of the mixture component g given the observation vector $\mathbf{O}_t(s)$ at time frame t.

The observation vector $\mathbf{O}_t(s)$ can be either the conventional raw acoustic features such as mel-frequency cepstral coefficients (MFCC) or their speaker- and channel-adapted versions which is computed through a per recording fMLLR transform [11, 10] typically obtained with a GMM-HMM system. Note from Fig. 1 that the same fMLLR transformed features can be used to train/evaluate the DNN as well as compute the sufficient Baum-Welch statistics for i-vector extraction.

Traditionally, the frame-level soft alignments, $\gamma_{tg}(s)$, in (2) and (3) are computed with a GMM acoustic model trained in an unsupervised fashion (i.e., with no phonetic labels). However, in [7], a supervised GMM-HMM acoustic model (derived from a speech recognition system) was utilized to estimate the

GMM-UBM hyperparameters for speaker recognition, assuming that class-conditional distributions for the various phonetic classes are Gaussian. More recently, inspired by the success of DNN acoustic models in automatic speech recognition (ASR) field, [6] proposed the use of DNN senone (context-dependent triphones) posteriors for computing the soft alignments, $\gamma_{tg}(s)$, which resulted in remarkable reductions in speaker recognition error rates. Motivated by these results, in this paper, we explore the DNN senone posterior based i-vectors for speaker recognition, and compare their effectiveness against GMM i-vectors on this task. Furthermore, we also investigate the impact of the senone set size on speaker recognition performance. It is worth noting that increasing the number of components in the unsupervised GMM acoustic model (with diagonal covariance matrices) for speaker recognition did not seem to result in much performance gains, if at all, in the recent studies [6, 8].

2.2. Linear discriminant analysis (LDA)

As noted before, i-vectors model speaker- and channel-dependent information within the same total variability subspace. Therefore, in order to select the most relevant feature subset for the speaker recognition task, LDA can be applied to i-vectors to annihilate the directions not informative for speaker recognition. In addition, reducing the dimensionality of i-vectors via LDA can improve the computational efficiency of the subsequent backend components in the system.

LDA computes an optimum linear projection $\mathbf{A} : \mathbb{R}^d \mapsto \mathbb{R}^n$, by maximizing the ratio of the inter-class scatter to intra-class variance, where $\mathbf{A}$ is a rectangular matrix with n linearly independent columns. Here, the within- and between-class scatter matrices are used to formulate a class separability criterion which converts the matrices into a single statistic. This statistic takes on larger values when the between-class scatter is larger and the within-class variance is smaller. Several such class separability criteria are described in [3], of which the following is the most widely used,

$$\hat{\mathbf{A}} = \underset{\mathbf{A}^T \mathbf{S}_w \mathbf{A} = \mathbf{I}}{\arg\max} \left[\mathrm{tr}\left(\mathbf{A}^T \mathbf{S}_b \mathbf{A}\right) \right], \qquad (4)$$

where $\mathbf{S}_b$ and $\mathbf{S}_w$ denote the between- and within- class scatter matrices, respectively. The optimization problem in (4) has an analytical solution that is a matrix whose columns are the n eigenvectors corresponding to the largest eigenvalues of $\mathbf{S}_w^{-1} \mathbf{S}_b$.

There are three disadvantages associated with the parametric nature of the scatter matrices $\mathbf{S}_b$ and $\mathbf{S}_w$. First, the underlying distribution of classes is assumed to be Gaussian with a common covariance matrix for all classes. Therefore, one

cannot expect the parametric LDA to generalize well to non-Gaussian and multi-modal (as opposed to unimodal) distributions. It is well known in the speaker recognition community that the actual distribution of i-vectors may not necessarily be Gaussian [12]. This is in particular more problematic when speech recordings are collected in the presence of noise and channel distortions [9, 13]. In addition, for the NIST SRE type of scenarios, speech recordings come from various sources and collects (sometimes out-of-domain), therefore unimodality of the distributions cannot be guaranteed. Second, notice that the rank of $\mathbf{S}_b$ is $C-1$, which means the parametric LDA can provide at most $C-1$ discriminant features. However, this may not be sufficient in applications such as language recognition where the number of language classes is much smaller than the dimensionality of the i-vectors [13]. Nevertheless, this may not pose a challenge for speaker recognition tasks in which the number of training speakers exceeds the dimensionality of the total variability subspace. Finally, because only the class centroids are taken into account for computing $\mathbf{S}_b$, the parametric LDA cannot effectively capture the boundary structure between adjacent classes which is essential for classification [3].

To overcome the above noted limitations of LDA, an NDA technique was proposed in [14], that measures both the within- and between-class scatters on a local basis using a nearest neighbor rule. We have previously evaluated the NDA for both speaker and language recognition tasks on high-frequency (HF) radio channel degraded data [9, 13], where it compared favorably to the LDA. We provide a brief description of NDA in the next section.

2.3. Nearest-neighbor discriminant analysis (NDA)

In order to alleviate some of the limitations identified for LDA, a nonparametric discriminant analysis techniques was proposed in [14]. In NDA, the expected values that represent the global information about each class are replaced with local sample averages computed based on the k-NN of individual samples. More specifically, in the NDA approach, the between-class scatter matrix is defined as,

$$\tilde{\mathbf{S}}_b = \sum_{i=1}^{C} \sum_{\substack{j=1 \\ j \neq i}}^{C} \sum_{l=1}^{N_i} w_l^{ij} \left(\mathbf{x}_l^i - \mathcal{M}_l^{ij}\right) \left(\mathbf{x}_l^i - \mathcal{M}_l^{ij}\right)^T, \quad (5)$$

where $\mathbf{x}_l^i$ denotes the lth sample from class i, and $\mathcal{M}_l^{ij}$ is the local mean of k-NN samples for $\mathbf{x}_l^i$ from class j which is computed as,

$$\mathcal{M}_l^{ij} = \frac{1}{K} \sum_{k=1}^{K} NN_k(\mathbf{x}_l^i, j), \quad (6)$$

where $NN_k(\mathbf{x}_l^i, j)$ is the kth nearest neighbor of $\mathbf{x}_l^i$ in class j. The weighting function w_l^{ij} in (5) is defined as,

$$w_l^{ij} = \frac{\min\left\{d^\alpha\big(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, i)\big), d^\alpha\big(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, j)\big)\right\}}{d^\alpha(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, i)) + d^\alpha(\mathbf{x}_l^i, NN_K(\mathbf{x}_l^i, j))}, \quad (7)$$

where $\alpha \in \mathbb{R}$ is a constant between zero and infinity, and $d(.)$ denotes the distance (e.g., cosine or Euclidean). The weighting function is introduced in (5) to deemphasize the local gradients that are large in magnitude to mitigate their influence on the scatter matrix. The weight parameters approach 0.5 for samples near the classification boundary (e.g., see $\{v_2, v_3, v_5, v_6\}$ shown in Figure 2), while dropping off to 0 for samples that are far from the boundary (e.g., see v_4 in Figure 2). The control
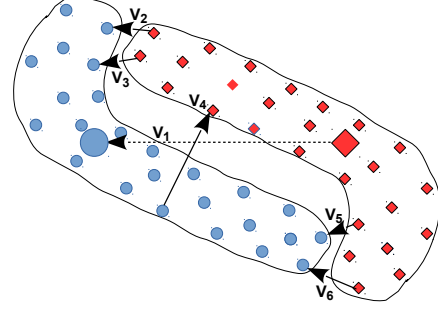

Figure 2: *Symbolic example illustrating the parametric versus nonparametric scatter between two classes.* v_1 *represents the global gradient of class centroids. The vectors* $\{v_2, \cdots, v_6\}$ *represent the local gradients.*

parameter α determines how rapidly such decay in the weights occurs.

The nonparametric within-class scatter matrix, $\tilde{\mathbf{S}}_w$, is computed in a similar fashion as in (5), except the weighting function is set to 1 and the local gradients are computed within each class. The NDA transform is then formed by calculating the eigenvectors of $\tilde{\mathbf{S}}_w^{-1}\tilde{\mathbf{S}}_b$.

Three important observations can be made from a careful examination of the nonparametric between-class scatter matrix in (5). First, notice that as the number of nearest neighbors, K, approaches N_j, the total number of samples in class j, the local mean vector, $\mathcal{M}_l^{ij}$, approaches the global mean of class j (i.e., $\boldsymbol{\mu}_j$). In this scenario, if we set the weight parameters to 1, the NDA transform essentially becomes the LDA projection, which means the LDA is a special case of the more general NDA.

Second, because all the samples are taken into account for the calculation of the nonparametric between-class scatter matrix (as opposed to only the class centroids), $\tilde{\mathbf{S}}_b$ is generally of full rank. This means that unlike the LDA that provides at most $C-1$ discriminant features, the NDA generally results in d-dimensional vectors (assuming a d-dimensional input space) for the classification. As we discussed before, this is of great importance for applications such as language recognition where the number of classes is much smaller than the dimensionality of the total subspace (or the input space in general).

Finally, compared to LDA, NDA is more effective in preserving the complex structure (i.e., local and boundary structure) within and across different classes. As seen from the example shown in Figure 2 (where k is set to 1 for simplicity), LDA only uses the global gradient obtained with the centroids of the two classes (i.e., v_1) to measure the between-class scatter. On the other hand, NDA uses the local gradients (i.e., $\{v_2, \cdots, v_6\}$) that are emphasized along the boundary through the weighting function, w_l^{ij}. Hence, the boundary information becomes embedded into the resulting transformation.

3. Experiments

This section provides a description of our experimental setup including speech data, the ASR system configuration, and the speaker recognition (SR) system configuration used in our evaluations.

Table 1: *Description of the 5 core enrollment/test conditions in the NIST 2010 SRE.*

Condition	Enroll	Test	Mismatch	#Target Trials	#Impostor Trials
C1	Interview microphone	Interview microphone (same type)	No	4,034	795,995
C2	Interview microphone	Interview microphone (different type)	Yes	15,084	2,789,534
C3	Interview microphone	Telephony	Yes	3,989	637,850
C4	Interview microphone	Room microphone	Yes	3,637	756,775
C5	Telephony	Telephony (different type)	Yes	7,169	408,950

3.1. Data

We conduct the core of our speaker recognition experiments using conversational telephone and microphone (phone call and interview) speech material extracted from datasets released through the linguistic data consortium (LDC) for the NIST 2004-2010 SRE [15, 16], as well as Switchboard Cellular (SWBCELL) Parts I and II and Switchboard2 (SWB2) Phase II and Phase III corpora. These datasets contain speech spoken in U.S. English (the non English portion was filtered out) from a large number of male and female speakers with multiple sessions per speaker. The NIST SRE 2010 data is held out for evaluations, while the remaining data are used to train the system hyper-parameters (i.e., the i-vector extractor, LDA/NDA, and PLDA). There are a total of 5 extended core tasks in the NIST SRE 2010 that involve telephone and microphone trials from both male and female speakers [17]. A more detailed description of the 5 tasks is presented in Table 1.

3.2. DNN system configuration

The architecture of the DNN acoustic model used to generate the soft alignments for i-vector extraction is shown in Fig. 3. The model, which has 7 fully connected hidden layers with 2048 units per layer except for the bottleneck layer that has 512 units, is discriminatively trained using the standard error back-propagation and cross-entropy objective function to estimate posterior probabilities of 10,000 senones (HMM triphone states). The training is accomplished using the IBM Attila toolkit [18] on 600 hours of conversational telephone speech (CTS) data from the Fisher corpus [19] with a 9-frame context of 40-dimensional speaker-adapted feature vectors obtained
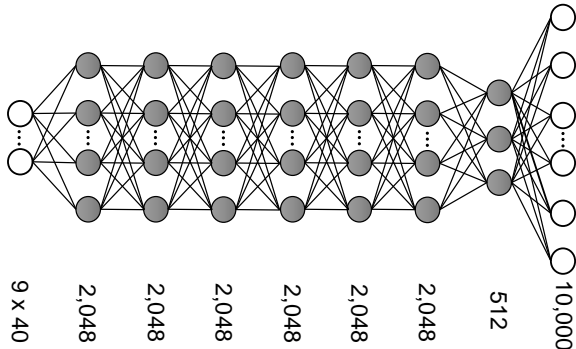
through per recording fMLLR transforms [10, 11]. The fMLLR transforms are generated for each recording with decoding alignments obtained from a GMM-HMM acoustic model. The GMM models are trained with 40-dimensional features which are derived from 13-dimensional MFCCs as follow; the base cepstral features from 9 consecutive frames are first spliced after cepstral mean-variance and vocal tract length normalizations (VTLN). An LDA transform is then applied to reduce the final feature vector dimensionality to 40. The range of the LDA transformation is diagonalized by means of a global semi-tied covariance transform (see [20, 21] for more details). In addition to running experiments with all the 10k senones, we also explore smaller senone set sizes of 2k and 4k which are obtained by merging the 10k HMM states using a phonetic decision tree with maximum-likelihood (ML) criterion [22].

3.3. SR system configuration

For speech parameterization (other than the fMLLR based features), we extract 13-dimensional MFCCs (including c_0) from 25 ms frames every 10 ms using a 24-channel mel filterbank spanning the frequency range 200-3500 Hz. The first and second temporal cepstral derivatives are also computed over a 5-frame window and appended to the static features to capture the dynamic pattern of speech over time. This results in 39-dimensional feature vectors. For non-speech frame dropping, we employ an unsupervised speech activity detector (SAD) based on voicing energy features [23]. After dropping the non-speech frames, global (recording level) cepstral mean and variance normalization (CMVN) is applied to suppress the short term linear channel effects.

In this paper, a 500-dimensional total variability subspace is learned and used to extract i-vectors from the recordings. To learn the i-vector extractor, out of a total of 60,178 recordings available from 1884 male and 2601 female speakers, we select 48,325 recordings from the NIST SRE 2004-2008, SWB-CELL, and SWB2 corpora. The zeroth and first order Baum-Welch statistics are computed for each recording using soft alignments obtained from either a gender-independent 2048-component GMM-UBM with diagonal covariance matrices, or the DNN acoustic model with 2k, 4k, and 10k senones. The GMM-UBM is trained using 21,207 recordings selected from the NIST SRE 2004-2006, SWBCELL, and SWB2 corpora.

After extracting 500-dimensional i-vectors, we either use LDA or NDA for inter-session variability compensation by reducing the dimensionality to 250. In order to train the NDA, we employ a one-versus-rest strategy to compute the inter-speaker scatter matrix in (5). This provides flexibility on the number of nearest neighbors used for computing the local means. A cosine similarity metric (as opposed to Euclidean) is used to find

Figure 3: *Architecture of the DNN acoustic model with 7 hidden layers used in our speaker recognition experiments.*

Table 2: *Performance comparison of IBM speaker recognition systems with various configurations on extended core condition 5 in the NIST SRE 2010. A DNN with 10k senones is used.*

System	EER [%]	minDCF08	minDCF10
GMM-MFCC-LDA	2.40	0.120	0.439
GMM-MFCC-NDA	1.55	0.076	0.286
DNN-MFCC-LDA	1.02	0.045	0.168
DNN-MFCC-NDA	0.76	0.036	0.147
DNN-fMLLR-LDA	0.82	0.032	0.120
DNN-fMLLR-NDA	**0.67**	**0.028**	**0.092**

Table 3: *Performance comparison of IBM speaker recognition systems with fMLLR features for 2k, 4k, and 10k DNN senones on extended core condition 5 in the NIST SRE 2010.*

System	#Senones	EER [%]	minDCF08	minDCF10
DNN-LDA	2k	1.19	0.054	0.212
DNN-NDA	2k	0.95	0.043	0.166
DNN-LDA	4k	0.98	0.041	0.169
DNN-NDA	4k	0.86	0.033	0.116
DNN-LDA	10k	0.82	0.032	0.120
DNN-NDA	**10k**	**0.67**	**0.028**	**0.092**

the k-nearest neighbors for each sample, and the exponent α in (7) is set to 1. The dimensionality reduced i-vectors are then centered (the mean is removed), whitened, and unit-length normalized. For scoring, a Gaussian PLDA model with a full covariance residual noise term [4, 5] is learned using the i-vectors extracted from all 60,178 speech segments (1884 male and 2601 female speakers) as noted previously. The Eigenvoice subspace in the PLDA model is assumed full-rank.

4. Results and Discussion

In this section, we summarize our results obtained with the experimental setup presented in Section 3. In the first experiment, we evaluated the effectiveness of the NDA versus the LDA for inter-session variability compensation and dimensionality reduction in the i-vector space. The outcome of this experiment is presented in Table 2 for the NIST SRE 2010 extended "tel-tel" trials (condition 5), in terms of the equal error rate (EER), minimum detection cost function with the NIST SRE 2008 [24] and 2010 [17] definitions (minDCF08 and minDCF10). It can be seen from the table that the systems with the NDA consistently
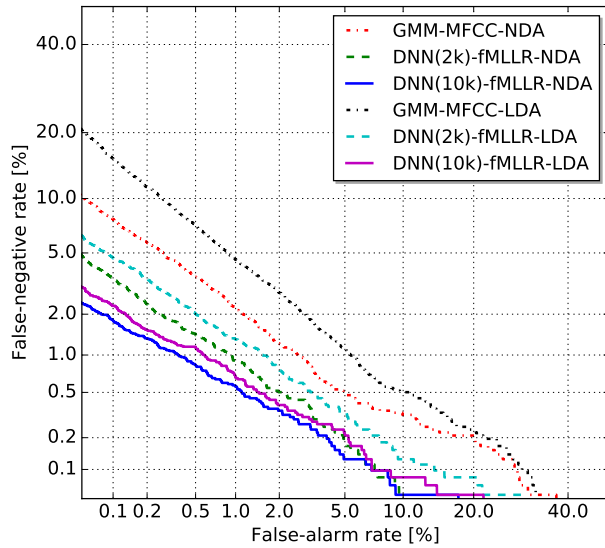
provide better speaker recognition performance across all three metrics. For the GMM based system, a relative improvement of 35% in EER is achieved with the NDA over the LDA, while for the DNN based systems with MFCCs and fMLLR features relative improvements of 26% and 18% are obtained, respectively. As we discussed before, this is due to the nonparametric representations for the scatter matrices in NDA that makes no assumptions regarding the underlying class-conditional distributions. In addition, NDA is more effective in capturing the local structure (as opposed to global bulk structure) and boundary information within and across different speakers. Another important observation that can be made from Table 2 is that, irrespective of the dimensionality reduction algorithm used, the systems with fMLLR features outperform the MFCC based systems. This is attributed to the ability of the fMLLR transforms in reducing the speaker and channel variabilities in the acoustic feature space.

In the next set of experiments, we investigated the impact of the number of senones on speaker recognition performance. Table 3 shows speaker recognition results on the NIST SRE 2010 "tel-tel" condition which are obtained with i-vectors computed using 2k, 4k, and 10k DNN senones and fMLLR features. Two important observations can be made from this table. First, the larger the number of senones, the better the performance. This is due to the discriminative nature of the DNN where increasing the granularity in the output layer improves the model ability in distinguishing among the various phonetic events. It is worth noting that increasing the number of components in the unsupervised GMM acoustic model (with diagonal covariance matrices) for speaker recognition did not result in much performance improvements in the recent studies [6, 8]. Second, irrespective of the number of senones used to calculate the sufficient statistics, the NDA based systems consistently perform better than the LDA based systems. We note that, in our experiments, increasing the number of senones beyond 10k did not yield much gains in performance.

Fig. 4 shows the detection error trade-off (DET) curves for the NDA and LDA based systems on the extended core condition 5 in the NIST SRE 2010. Consistent with our previous observations, it is seen that the NDA based systems achieve the best performance across a wide range of operating points on the DET curves. The performance gap between the NDA and LDA based systems is, however, reduced when DNN senone posteriors are used to compute the i-vectors, and increasing the senone set size from 2k to 10k further narrows this gap.

For completeness, we also evaluated the performance of our speaker recognition system on extended microphone and

Figure 4: *DET plot comparison of IBM speaker recognition systems with various configurations on extended core condition 5 in the NIST SRE 2010.*

Table 4: *Performance comparison of IBM speaker recognition systems with various configurations on extended microphone and telephone conditions (C1–C4) in the NIST SRE 2010. A DNN model with 10k senones is used.*

System	EER [%]				minDCF08				minDCF10			
	C1	C2	C3	C4	C1	C2	C3	C4	C1	C2	C3	C4
GMM-MFCC-NDA	1.39	1.89	1.80	1.46	0.053	0.084	0.081	0.061	0.215	0.313	0.315	0.251
DNN-MFCC-NDA	**0.84**	**1.41**	**0.83**	**0.63**	**0.027**	**0.046**	0.036	**0.022**	**0.104**	**0.157**	0.127	0.103
DNN-fMLLR-NDA	1.02	1.44	0.90	0.77	0.033	0.049	**0.034**	0.025	0.112	0.158	**0.119**	**0.096**

telephone conditions (C1–C4) in the NIST SRE 2010. The results are provided in Table 4 for both the GMM and DNN based systems. It is clear that the DNN based systems, with either MFCCs or fMLLR features, perform significantly better than the GMM based system. Additionally, the DNN based system trained with raw MFCCs tend to perform better than the fMLLR based system, at least in terms of EER. We speculate that this is because the fMLLR transforms, which are obtained using GMM-HMMs trained only on telephony data, are unable to effectively reduce the variability due to channel mismatch on microphone recordings.

5. Conclusions

In this paper, we presented the recent improvements made in our state-of-the-art i-vector speaker recognition system. We investigated the impact of several key components of the system on performance using extended core tasks in the NIST 2010 SRE that involved both microphone and telephone trials. Some important observations made from our experiments were as follows: 1) the NDA was found to be consistently more effective than the LDA for inter-session variability compensation in i-vector based speaker recognition, 2) the fMLLR based features provided better representation than raw MFCCs for matched data conditions (i.e., telephony trials), and 3) the DNN based UBM with large number of components (i.e., 10k senones) resulted in remarkable improvements in the performance of our system. To the best of our knowledge, the results presented in this paper represent the best performances reported to date on the extended core tasks in the NIST 2010 SRE.

6. References

[1] P. Kenny, G. Boulianne, P. Ouellet, and P. Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Trans. Audio Speech Lang. Process.*, vol. 15, no. 4, pp. 1435–1447, 2007.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 4, pp. 788–798, 2011.

[3] K. Fukunaga, *Introduction to Statistical Pattern Recognition*, 2nd ed. New York: Academic press, 1990.

[4] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *Proc. IEEE ICCV*, Rio De Janeiro, October 2007, pp. 1–8.

[5] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Proc. INTERSPEECH*, Florence, Italy, August 2011, pp. 249–252.

[6] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *Proc. IEEE ICASSP*, Florence, Italy, May 2014, pp. 1695–1699.

[7] M. K. Omar and J. Pelecanos, "Training universal background models for speaker recognition," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2010)*, Brno, Czech, June 2010, pp. 52–57.

[8] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," in *Proc. IEEE ASRU*, Scottsdale, AZ, December 2015, pp. 92–97.

[9] S. O. Sadjadi, J. W. Pelecanos, and W. Zhu, "Nearest neighbor discriminant analysis for robust speaker recognition," in *Proc. INTERSPEECH*, Singapore, Singapore, September 2014, pp. 1860–1864.

[10] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Trans. Speech Audio Process.*, vol. 3, no. 5, pp. 357–366, September 1995.

[11] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," *Comput. Speech Lang.*, vol. 12, no. 2, pp. 75–98, 1998.

[12] P. Kenny, "Bayesian speaker verification with heavy tailed priors," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2010)*, Brno, Czech, June 2010.

[13] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, "Nearest neighbor discriminant analysis for language recognition," in *Proc. IEEE ICASSP*, Brisbane, Australia, April 2015, pp. 4205–4209.

[14] K. Fukunaga and J. Mantock, "Nonparametric discriminant analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 5, no. 6, pp. 671–678, 1983.

[15] C. Cieri, L. Corson, D. Graff, and K. Walker, "Resources for new research directions in speaker recognition: The Mixer 3, 4 and 5 corpora," in *Proc.INTERSPEECH*, Antwerp, Belgium, August 2007, pp. 950–953.

[16] L. Brandschain, D. Graff, C. Cieri, K. Walker, C. Caruso, and A. Neely, "Mixer 6," in *Proc. LREC*, Valletta, Malta, May 2010.

[17] NIST, "The NIST Year 2010 Speaker Recognition Evaluation Plan," http://www.nist.gov/itl/iad/mig/upload/NIST_SRE10_evalplan-r6.pdf, 2010.

[18] H. Soltau, G. Saon, and B. Kingsbury, "The IBM Attila speech recognition toolkit," in *Proc. IEEE SLT*, Berkeley, CA, December 2010, pp. 97–102.

[19] C. Cieri, D. Miller, and K. Walker, "The Fisher corpus: A resource for the next generations of speech-to-text," in *Proc. LREC*, Lisbon, Portugal, May 2004.

[20] S. Ganapathy, S. Thomas, D. Dimitriadis, and S. Rennie, "Investigating factor analysis features for deep neural networks in noisy speech recognition," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 1898–1902.

[21] G. Saon, H. K. Kuo, S. Rennie, and M. Picheny, "The IBM 2015 English conversational telephone speech recognition system," in *Proc. INTERSPEECH*, Dresden, Germany, September 2015, pp. 3140–3144.

[22] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proc. Workshop Human Lang. Tech.*, March 1994, pp. 307–312.

[23] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 197–200, 2013.

[24] NIST, "The NIST Year 2008 Speaker Recognition Evaluation Plan," http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf, 2008.