IDIAP RESEARCH REPORT

ANALYSIS OF POSTERIOR ESTIMATION APPROACHES TO I-VECTOR EXTRACTION FOR SPEAKER RECOGNITION

Srikanth Madikeri Petr Motlicek Marc Ferras
Subhadeep Dey

Idiap-RR-15-2018

OCTOBER 2018

Analysis of Posterior Estimation Approaches to I-vector Extraction for Speaker Recognition

Srikanth Madikeri, Petr Motlicek, Marc Ferras and Subhadeep Dey
Idiap Research Institute , CH-1920 Martigny, Switzerland
srikanth.madikeri, petr.motlicek, mferras, sdey@idiap.ch

***Abstract*—The i-vector approach to speaker recognition requires estimating Sufficient Statistics (SS) (i.e., zeroth- and first-order statistics) for a given utterance of speech with respect to a Universal Background Model (UBM) usually represented by Gaussian Mixture Models (GMM). To estimate SS, alternate approaches have also been experimented. Studies suggest that using acoustic phone posteriors estimated from a Deep Neural Network (DNN) based Automatic Speech Recognition (ASR) system can be useful in estimating accurate speaker representations with i-vectors. In this paper, we analyze and compare the UBM-GMM and several versions of DNN approaches together with subspace Gaussian Mixture Models to estimate i-vectors for a speaker. We show that better alignments of speech frames can lead to superior speaker verification performance. This is achieved through the use of the decoded output from the ASR system, whereas existing systems only use posteriors at the output of the DNN directly. The posteriors from the decoding lattices are rescaled suitably to deal with its sparse nature that can affect SS computation. We show that a direct correlation exists between senone recognition accuracy of the system generating the posterior and the performance of corresponding speaker recognition systems. The posterior estimation methods are compared on standard NIST 2010 SRE dataset. Significant improvements are obtained when using the ASR decoder, thereby confirming that with better frame-level alignments speaker verification performance improves. Equal Error Rate (EER) as low as 0.9% is achieved on the telephone condition of the evaluation set.**

Index Terms: speaker recognition, posterior estimation, i-vectors, GMM, DNN, SGMM

I. Introduction

Speaker recognition, concerned with the identification or verification of a person from his/her voice, has witnessed considerable progress in the last decade. With the introduction of techniques such as Joint Factor Analysis (JFA) [1] and Probabilistic Linear Discriminant Analysis (PLDA) [2], text-independent speaker verification systems are nowadays capable of achieving error rates under 1% for known acoustic conditions. State-of-the-art speaker recognition systems are built around the i-vector (*identity* vector) approach [3], modeling a speech recording by projecting its acoustic features onto a low-dimensional representation. The i-vector space, also referred to as the Total Variability Subspace (TVS), models many of the variabilities observed in the original recording, e.g. speaker, channel and language. Estimating an i-vector for a speech recording requires a sequence of short-term acoustic feature vectors to be aligned with the mixture components of a Universal Background Model-Gaussian Mixture Model (UBM-GMM). Sufficient Statistics (SS) are computed from this frame-to-mixture alignment. The statistics are used to project the utterance onto the TVS.

In [4], the UBM-GMM components were replaced by the output states of a DNN-based hybrid Automatic Speech Recognition (ASR) system (referred to as the DNN/HMM system in this paper). The posteriors at the output of the DNN were used to estimate SS. It demonstrated that a well-defined acoustic space significantly helps model speakers better, as opposed to unsupervised training of the UBM-GMM components. The improvements in modelling come from both well-defined nature of the DNN output states (typically senone units) and improved alignment accuracy from the discriminative classifier [5]. If defining the GMM components from the acoustic classes (more accurately, the states) of an ASR system was alone sufficient, a supervised GMM would perform just as good contrary to the findings in [6], thus highlighting the importance of better alignment. Extending this argument, it would be reasonable to assume that the alignment obtained after ASR decoding with a Language Model (LM) would further improve performance since it well known that frame-level alignment becomes even more accurate than the DNN output. However, in our experiments and in literature (where the DNN/HMM system provide acceptable baselines), no such results have been demonstrated. In this paper, we show that performance issues can arise due to the sparsity of the posteriors after ASR decoding as they affect the Gaussianity assumptions in LDA and PLDA. We propose two countermeasures: the first method modifies the decoder configuration such that the alignment is preserved but the posteriors are less sparse, and the second technique involves fusing the DNN output directly with the frame alignment obtained from the decoder.

The rest of the paper is organized as follows: Section III briefly describes the i-vector system used. State-of-the-art DNN/HMM ASR system and strategies to collect SS for speaker recognition are described in Section IV. The architecture of the proposed system is given in Section V. Experimental results are presented in Section VI. Finally, the conclusions are given in Section VIII.

II. Previous Work and Motivation

In this section, we describe previous work relevant to the techniques used in this paper. The i-vector approach forms the basis of state-of-the-art speaker recognition systems used today [3]. While many approaches attempted to combine the modelling power of DNN with i-vector, recently succesful approaches fall under two major categories: the use of posteriors from a DNN trained for ASR ([4], [7]) to compute

SS for i-vectors, and the use of Bottleneck Features (BNFs) obtained from a DNN (or a stacked DNN) trained for ASR [8]. The former will be referred to as the DNN/HMM system for brevity. In [8], stacked DNNs were trained for ASR. The DNNs had BN layers that were later used to train a UBM-GMM based speaker recognition system. Results showed that combining MFCCs with BNFs provided the best performance among several configurations of UBM-GMM systems. Both techniques still require training an ASR system, which in turn requires 1000s of hours of transcribed training data tying the speaker recognition system to one particular language.

The use of ASR systems to aid speaker recognition is not new. HMM/GMM based ASR systems have already been used as UBMs for speaker recognition systems [9], [6]. In [4], the authors comment that it did not fair better than their proposed system. It is also now well established that the discriminative power of DNN for acoustic state classification is superior to that of GMM modelling of HMM states. In [10], [11], phone recognizers were used to obtain speaker-discriminative features based on idiolectal uniqueness as motivated in [12]. In [13], Subspace GMM (SGMM) systems were used to train speaker vectors instead of using the conventional TVS training [14].

The performance gains obtained with the DNN/HMM setup was further improved by employing techniques originally developed to improve ASR systems. In [15], fMLLR transform was applied to MFCC features to train the DNN, while still using conventional MFCCs to train the i-vector system. Using fMLLR did not show improvements in the telephone-telephone condition of the NIST SRE 2010 dataset ([16]) when LDA-PLDA backend was used, but better results were obtained when Nearest-Neighbour Discriminant Analysis (NDA) was applied instead. Results on other conditions in the same dataset with the NDA backend showed improvements for conditions 3 and 4 with the minDCF10 metric, but not in terms of Equal Error Rate (EER). The results suggest that applying speaker transform can be helpful in mismatch conditions.

In [6], Time Delay DNN (TDNN) to estimate frame-level posteriors was introduced. An interesting comparison with a supervised GMM (sup-GMM) is also presented in their work. While the sup-GMM perform better than the conventional GMM, the TDNN based system outperforms all other systems.

While the use of ASR systems, partially or completely, have been throughly experimented, the variety of such systems do not tell us which one to use. And for each configuration how to use it. With the success of using posteriors from DNN to compute i-vectors, it is natural to extend the technique to use the output of the decoder to compute the same. However, performance issues can arise despite the increase in alignment accuracy owing to the sparse nature of the posteriors obtained after decoding. The sparsity can affect the Gaussianity assumptions of i-vectors and lead to sub-optimal results. Thus, in this paper we tackle this issue by rescoring the output of the decoded system with no weight on language model. That is, after decoding only the acoustic posteriors are used.

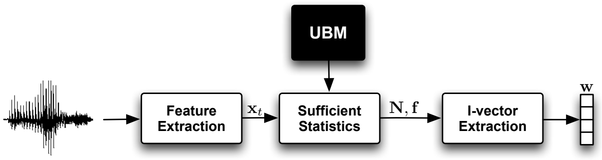

Fig. 1. Block diagram showing the i-vector extraction procedure from a speech recording.

III. I-VECTOR EXTRACTION

The i-vector extractor projects Gaussian mean supervectors on a low-dimensional subspace called *total variability space* (TVS) [3]. The i-vector extraction procedure is shown in Figure 1 and further described below. The underlying variability model used for i-vector extraction is

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{1}$$

where $\mathbf{s}$ is the supervector adapted with respect to a UBM-GMM from a speech recording. The vector $\mathbf{m}$ is the mean of the supervectors usually obtained from the UBM-GMM, $\mathbf{T}$ is the matrix with its columns spanning the total variability subspace and $\mathbf{w}$ is the low-dimensional i-vector representation. In the above model, the i-vector is assumed to have Gaussian distribution with zero mean and unit variance as prior distribution.

Given a sequence of MFCC feature vectors $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t\}$, the first-order statistics $(\mathbf{f})$ are estimated to obtain the i-vector representation. The subvector $\mathbf{f}_c$ of $\mathbf{f}$ is given by

$$\mathbf{f}_c = \Sigma_c^{-\frac{1}{2}} \left(\sum_n \gamma_{n,c} \mathbf{x}_n - \mu_c \right), \tag{2}$$

where $\mathbf{f} = [\mathbf{f}_1^t, \mathbf{f}_2^t, \ldots, \mathbf{f}_C^t]^t$, C is the number of mixtures in the UBM-GMM, $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ are the mean and covariance matrix of the c^{th} mixture. The posterior $\gamma_{n,c}$ for the n^{th} frame of speech with respect to the c^{th} mixture is given by

$$\gamma_{n,c} = \frac{\omega_c \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c\right)}{\sum_{k=1}^{C} \omega_k \mathcal{N}\left(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)}. \tag{3}$$

where ω_c is the weight of c^{th} mixture component. Given the first order statistics, the i-vector is estimated as follows

$$\mathbf{w} = \left(\mathbf{I} + \sum_{c=1}^{C} N_c \mathbf{T}_c^t \boldsymbol{\Sigma}_{\boldsymbol{c}}^{-\frac{1}{2}} \mathbf{T}_c \right)^{-1} \mathbf{T}^t \boldsymbol{\Sigma}^{-1} \mathbf{f}, \tag{4}$$

where $\mathbf{T}_c$ is the submatrix of $\mathbf{T}$ for the c^{th} mixture, $\boldsymbol{\Sigma}$ is a block diagonal matrix with each block given by $\boldsymbol{\Sigma}_c$ for $c = 1, 2, \ldots C$ and

$$N_c = \sum_n \gamma_{n,c} \tag{5}$$

is the effective number of feature vectors assigned to the cluster c. The i-vector estimation equation (Equation 4) is the MAP estimate of $\mathbf{w}$ assuming Gaussian distribution.

The matrix $\mathbf{T}$ is trained from a large development data set containing multiple speakers with multiple speech recordings for each speaker. An Expectation-Maximization (EM) algorithm is used to estimate the matrix [14], [17].

The i-vectors obtained from a speech utterance are further projected onto a discriminative space using techniques such as LDA, WCCN [18], [3] and PLDA [2], [19]. Prior to training and evaluating i-vector using PLDA, a length normalization is applied [20]. Using PLDA parameters two i-vectors can be compared as belonging to the same class or as belonging to two different classes, thus generating a simple log likelihood ratio to score a pair of speech utterances.

IV. DNN/HMM ASR for SS computation

Recent studies have leveraged the enormous successes of DNN-based modelling for speaker recognition [4], [7], [21], [8]. DNNs enable non-linear modelling of data, which is often required for real world data such as audios, videos, etc. The non-linear processing helps to model complex processes such as speech production. Thus, DNN-based modelling of speech units has found applications in ASR [22], Text to Speech Synthesis (TTS), prosodic modelling, etc.

A. Senone posteriors for speaker recognition

State-of-the-art ASR systems employ DNNs to model fundamental speech units such as senones, which are clustered, context-dependent sub-phonetic HMM states generated by a set of phonetic decision trees. State posterior probabilities are estimated for a frame of speech with context of typically 9 to 13 frames where each state represents a cluster of senones. The likelihoods from the posterior probabilities, along with word sequence likelihoods from the Language Model (LM), are then passed to the decoder to obtain the ASR output [23], [22]. In [4], it was shown that a DNN trained for ASR can replace the traditional UBM-GMM to estimate SS for i-vector extraction. The posteriors obtained at the output of the DNN forward pass process are directly used to compute SS. We term this system HMM/DNN forward pass system. This technique resulted in large performance gains for speaker verification systems as better alignments are obtained with respect to the UBM components. The results showed that replacing unsupervised training of the UBM components with well-defined acoustic classes can have a significant impact on verification performance. The high-scoring component in the posteriors for each speech frame is expected to come from linguistically related senones. Later, this technique was also extended to language identification [21], [24].

Although there has been sufficient evidence that phone-level classes possess speaker-discriminative information ([25]), successful integration into the state-of-the-art framework such as i-vector PLDA was not achieved until recently. The effectiveness of senone posteriors for i-vector extraction provides new research directions for speaker recognition. Particularly, we seek to investigate whether we can take advantage of accurate senone alignments obtained by using the LM and the ASR decoder [26]. The LM not only provides more accurate alignments but may also help capture speaker-dependent characteristics closely related to the speech contents, which is useful for better speaker discrimination [12], [27].

We propose to study the estimation of SS from senone posteriors obtained at the output of ASR decoding to take advantage of better senone alignments. Posterior vectors to estimate SS are obtained from the: (i) word recognition lattices (i.e., word LM was used to generate word lattices), (ii) phone recognition lattices (i.e. phone LM was used to generated phone lattices). Eventually senone-level posteriors are extracted from these lattices similar to posterior vectors extracted with only acoustic models (e.g. DNN forward-pass). Even though the senone alignments are more accurate, they need not result in better speaker recognition performance because of their inherent sparsity. Such high sparsity arises as a result of smoothing the posterior vectors obtained from the DNN and smoothed by the ASR decoder based on word sequence probabilities from the LM. We show, through senone recognition rates, that this may not be favorable for speaker recognition systems given the nature of SS estimation as given in Equations 2, 5. The contribution of senones is directly determined by not only their presence in the lattice generated by the ASR decoder, but also by the posterior values themselves. Extremely low values contribute little to the SS and may prove detrimental to the speaker recognition performance as they tend to have an effect similar to missing the senones altogether.

In this paper, we propose to rescale the ASR lattices with only the acoustic likelihoods, thereby completely ignoring the language model probabilities for SS estimation. Once the lattice is generated for an utterance, the forward-backward algorithm is applied with an acoustic scale suitable to generating posterior values that can be suitably handled by the i-vector system. As a result, all the active states in the lattice contribute to the SS. In Section VI-G, it is shown that this method of post-processing the posteriors does not affect the senone recognition rate.

B. Integration into i-vector framework

To integrate an ASR system into the i-vector framework, the parameters of the UBM-GMM are estimated from frame-level posterior probabilities, computed by the ASR system, and their corresponding features as required by Equations 1, 2, and 3. This process replaces the M step of the EM algorithm used to estimate the parameters of a GMM [28], [29], [30]. The update equations are given as follows

$$\omega_c = \frac{\sum_n \gamma_{n,c}}{\sum_c \sum_n \gamma_{n,c}} \tag{6}$$

$$\boldsymbol{\mu}_c = \frac{\sum_n \gamma_{n,c} \mathbf{x}_n}{\sum_n \gamma_{n,c}} \tag{7}$$

$$\boldsymbol{\Sigma}_c = \frac{\sum_n \gamma_{n,c} \left(\mathbf{x}_n - \boldsymbol{\mu}_c\right) \left(\mathbf{x}_n - \boldsymbol{\mu}_c\right)'}{\sum_n \gamma_{n,c}}. \tag{8}$$

In our experiments, we also considered $\boldsymbol{\Sigma}_c$ to be diagonal as no major improvements were observed when a full covariance matrix was used.

C. Bottleneck features

Bottleneck Features (BNF) have provided an alternate method to leverage the modelling power of DNN [8], [31]. In [8], the BN extracted from a stacked DNN is used as a complementary feature to MFCCs in the UBM-GMM i-vector system. Significant performance improvements were obtained with respect to the conventional UBM-GMM i-vector system. In this paper, we also compare the approach to use BN with approaches that derive senone posteriors from DNN-based ASR systems. First, we show that the score-level fusion of individual systems trained on MFCC and BNFs perform as well fusing the features and training one single system. Next, we show that

V. Proposed Analysis

The different systems presented earlier give rise to multiple posterior extraction procedures. All such systems, apart from the baseline i-vector PLDA system with UBM-GMM, are developed for ASR. Posteriors can be obtained either using only the acoustic model (AM) or using both the AM and the LM. The HMM/DNN system produces posteriors at the end of the DNN forward pass and after the decoding stage. We analyse the influence of LM on the acoustic likelihoods with respect to the performance of i-vector systems trained using such posteriors.

As state-of-the-art ASR systems employ speaker adaptation techniques such as fMLLR, comparisons are made on two types of ASR systems: one that applies fMLLR and one that does not apply any speaker adaptation technique. The former system is termed HMM/SD-DNN while the latter is called HMM/SI-DNN, where the acronyms SD and SI signify that the systems are speaker dependent and speaker independent, respectively. The speaker dependency indicates the use of fMLLR transforms on the input MFCC features in the training and decoding stages. A HMM/GMM system is used to estimate the fMLLR transform over a single speech recording. This transform is applied on MFCC features before being passed to the HMM/SD-DNN system as input. The same procedure is followed to train the SD-DNN.

In general, to train HMM/DNN systems, alignments from HMM/GMM systems act as labels. Such GMM-based systems have considerably lesser model complexity (e.g. in terms of the number of parameters) and their performance is worse compared to the HMM/DNN system (as will be reported in Section VI). Using posteriors obtained from such systems can help demonstrate the importance of senone-alignment accuracy for SS estimation. Thus, as a result of the two DNN systems developed, we also propose the study of using alignments from the HMM/SI-GMM and HMM/SD-GMM systems. The speaker recognition performance of i-vector systems that use posteriors from these systems to compute SS are compared with each other and the baseline UBM-GMM system as shown in Figure 2.

Previous studies have shown that posteriors from a DNN trained for ASR can help improve speaker verification performance. This system only differs from the UBM/GMM i-vector system in the source of posteriors to estimate SS. Thus, we also conduct experiments to analyse speaker discriminability of these posteriors. In such an experiment, the zeroth-order statistics over a recording are collected and normalized to sum to 1.0. This posterior vector is used as the speaker i-vector and two i-vectors are compared using the Kullback-Leibler (KL) divergence measure.

In addition to the experiments mentioned above, we also propose to separately analyse the effect of using MFCCs with context - similar to the input to DNNs. Conventionally, DNNs are trained with a higher dimensional feature vector obtained by stacking multiple neighbouring MFCC frames as input. In this paper, the DNN uses a context of 9 frames ($\approx$ 190ms). However, other systems presented do not use as much context explicitly (although in the case of ASR systems the context is also taken care while decoding). Thus, we analyse the effect of using a longer context, in this case a context of 190 ms, to estimate the UBM-GMM parameters. The UBM-GMM is then used to compute SS while keeping the original MFCC feature vectors to train the i-vector extractor. This was done in order to reduce the model complexity of T-matrix as features with higher dimensionality would require more data. To reduce the dimensionality of the stacked MFCC, LDA is applied with senones as classes. The labels for each frame of the LDA are derived from the alignment obtained from the HMM/SI-GMM system.

Finally, the performances of all the above mentioned systems are compared with an alternative method exploiting DNNs for speaker recognition. In particular, BN features produced from a DNN trained for ASR are concatenated to the conventional MFCC features to extract i-vectors [8]. A GMM is estimated on such features on the same development set as other systems. A UBM-GMM based i-vector PLDA system is developed with these features. With the performance of the system using BN features we can directly compare the two mainly used approaches to exploit DNNs for speaker recognition.

VI. Experimental Setup

Speaker verification experiments were conducted on the female subset of the NIST 2010 SRE data, following the official protocol [16]. Following the setup of [8], system performance is evaluated on conditions 1 through 5 (labelled as cond1 through cond5 in this article) and results are given as EER (Equal Error Rates) and minDCF (minimum Decision Cost Function). Conditions 1 through 4 use interview style recordings for training and condition 5 uses telephone recordings. For testing, conditions 1 and 2 have interview style data, conditions 3 and 5 have telephone data and condition 4 has data recorded over microphones.

In the following sections, the setup of the explored speaker verification systems are detailed.

System configurations

The i-vector based systems explored in this work differ in how posteriors are computed to estimate the SS (i.e., from UBM-GMM, HMM/GMM, HMM/SGMM, and HMM/DNN models), whereas all other processing steps are kept the same across systems.

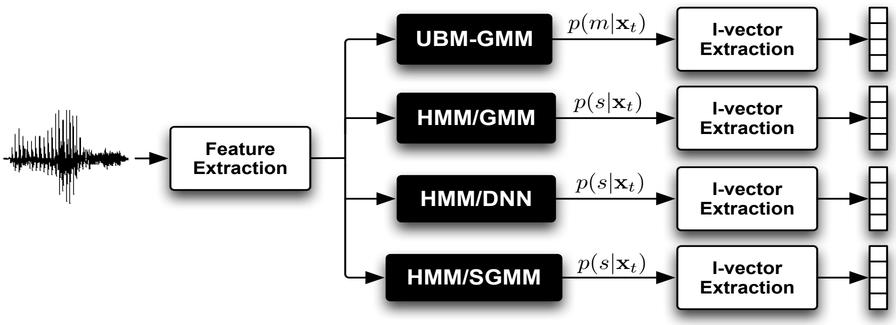

Fig. 2. Block diagram showing the architecture of the proposed system that uses posteriors obtained from the different ASR systems compared to using the conventional UBM-GMM system.

A. Feature configuration and training data

The front-end used 20 MFCC features along with delta and acceleration parameters, extracted every 10 ms using a window of 30 ms (as used by systems such as [8]). They were further processed through a short term Gaussianization module ([32]) with a context of 300 frames. All systems presented in this paper use the same feature configuration. Although it is common for ASR systems to use only 13 MFCC dimensions with delta and acceleration, we preserved the MFCC configuration for both ASR and i-vector systems as it was observed the Word Error Rate (WER) of ASR systems dropped by $\approx 2\%$ absolute with the increased number of coefficients.

The female part of Fisher English Part I and II data ($\approx$) 2000 hours) was used to train the UBM-GMM system, the T-matrix, the parameters in Eq. 7 and 8, and all the acoustic models used in this work. LDA and PLDA parameters were trained with the following datasets: The NIST datasets - SRE 2004, 2005, 2006, 2008 and 2008 extended, Switchboard Part II and Part III, and Switchboard Cellular Part I and II.

The ASR system employs a CMU dictionary with 42k words and a 3gram Language Model (LM) for decoding with word LM [13]. Additionally, a bigram phone LM were trained on the same Fisher training data for our language independent experiment with the HMM/SI-DNN system.

B. Baseline UBM-GMM I-vector system

A UBM-GMM with 2048 components and i-vector extractor of 400 dimensions were trained. The i-vector dimension was reduced to 350 after LDA, followed by length normalization before being scored using PLDA.

The Kaldi toolkit [33] was used for LDA and PLDA training. A standard i-vector extractor was implemented for Kaldi as well [34], based on the baseline system described in [17].

C. HMM/GMM system parameters

The HMM/SI-GMM system uses context-dependent triphone states with GMM observation probability density functions, and a total of 1'530 senones and 300k Gaussians [35]. The number of senone states was automatically derived by the tree-clustering algorithm that was constrained to have around 2k states in order to be comparable with the number of UBM-GMM components. We note that state-of-the-art ASR systems have higher number of states. Comparing the ASR performance of ASR systems with 2k and 7k states, we observed about 1% absolute drop in WER. The system is built on the speaker recognition front-end, as described in Section VI-A. We trained the HMM/SD-GMM system, in which fMLLR transforms estimated per-recording are used, as we will not have access to speaker labels during enrollment and testing of speaker recognition systems.

D. HMM/SGMM system parameters

The UBM with 2048 components is trained by clustering all the Gaussians in a HMM/GMM based ASR system. A system with 4300 states and 100k Gaussians (following state-of-the-art system parameters for HMM/SGMM as in [36]) and similar number of sub-states as the number of Gaussians were used [37]. Unlike ASR systems, the SGMM developed here is tuned to speaker recognition purposes. The phonetic subspace is constrained to a dimension of 40 (i.e. $S = 40$) while the speaker dimension is set to 400.

E. HMM/DNN system parameters

The input to the DNNs are 540 dimensional vectors obtained by stacking 9 MFCC feature vectors and the output classes are senone probabilities. As mentioned in Section IV-A, two DNN models, HMM/SI-DNN and HMM/SD-DNN, were trained with the same configuration with alignments from the HMM/SI-GMM and HMM/SD-GMM, respectively. We used the Kaldi toolkit to train a DNN with 6 hidden layers with 2'000 sigmoid units per layer and softmax units at the output. The DNN parameters were initialized with stacked Restricted Boltzmann Machine (RBM) that are pretrained in a greedy layer-wise fashion [38], [39]. The utterances and frames are presented in a randomized order while training both of these networks using stochastic gradient descent to minimize the cross-entropy between the labels and network output. The DNNs are trained on the same features as the GMM-HMM baselines, except that the features are globally normalized to have zero mean and unit variance. The fMLLR transforms are the same as those estimated for the GMM-HMM system during training and testing.

TABLE I

ASR results on Fisher development set in Word Error Rates (WER) [%]. The HMM/GMM systems have 1'530 states with 300k Gaussians and the HMM/SGMM system has 4300 states with 100k substates.

System	WER [%]
HMM/SI-GMM	42.3
HMM/SD-GMM	35.9
HMM/SI-DNN	26.0
HMM/SD-DNN	25.2
HMM/SI-GMM (with BN features)	27.8
HMM/SGMM	31.1

Senone posteriors from the SI- and SD-DNNs are obtained using two different methods: (i) a DNN forward pass, and (ii) full DNN/HMM decoding, i.e. using both AM and LM. To generate i-vectors, the means and the covariance matrices for each of the DNN outputs are computed from the first and second order moments of the feature vectors using the DNN posteriors. Only diagonal covariances are estimated.

In the final DNN setup, inspired by a stacked DNN [8], we trained another DNN on the same data as before to extract BN features. We did not train a stacked DNN as was done in [8] so that the performance of the system can be compared with the rest. A BN layer with 80 linear outputs is introduced after 4 hidden layers and another hidden layer follows the bottleneck layer. Finally, the output layer has the same number of states (1'530).

F. ASR results

The performance of the ASR systems, namely the HMM/SD-GMM system, the HMM/SI-GMM system, the SGMM system, the HMM/SI-DNN system and the HMM/SD-DNN system are compared in Table I. The systems are evaluated on a subset of the Fisher data set that was kept aside for evaluation. As expected, the Word Error Rate (WER) is lower for the DNN systems and speaker adaptation is always observed to be useful. The SGMM system outperforms the HMM/SD-GMM system. The HMM/SD-DNN system outperforms all other systems with a WER of 25.2%. This performance is 1.2% better in absolute terms than the HMM/SI-DNN system.

In addition to the conventional ASR approaches mentioned above, we explore an alternative decoding method in order to move away from the language dependence of the ASR systems used in this paper. The LM that was trained for a large vocabulary speech recognition is replaced by a simple phone bigram LM trained on the Fisher dataset. This system is referred to as “HMM/SI-DNN (phone bigram)”. Although such a rudimentary LM will certainly not improve the ASR performance we consider it as a step in the direction of language-independent ASR development for speaker verification. Although there are numerous ways in which language independence can be achieved, we considered an approach that would have minimal changes with the state-of-the-art ASR system while retaining the well-trained acoustic models. We consider other techniques as a part of our future work.

TABLE II

SENONE RECOGNITION RATES ON FISHER DEVELOPMENT SET OF ALL THE ASR SYSTEMS USED

System	SRR (%)
HMM/SI-GMM	55.2
HMM/SD-GMM	56.2
HMM/SI-DNN (forward pass)	53.5
HMM/SI-DNN	73.4
HMM/SD-DNN (forward pass)	52.4
HMM/SD-DNN	72.3

The Phone Error Rates (PER) of the HMM/SI-DNN system and HMM/SI-DNN (bigram) system were also compared. On the Fisher development set, a deterioration of approximately 2% was observed from using the word LM to phone LM.

G. Senone Recognition Rates

In this section, the Senone Recognition Rates (SRR) of the HMM based ASR systems are analysed. The performances are presented in Table II. The SRR is the percentage of senones correctly identified according to the groundtruth (which may be obtained by aligning the reference transcription using an ASR system). A speech frame is considered correctly identified if the maximum scoring senone matches with the groundtruth. As expected, the SRR improves with better acoustic modelling and is the best when an ASR decoder is used with the word LM. The SRR improves by 2.4% absolute after decoding the HMM/SI-DNN forward pass posteriors. The improvement is still significant considering the number of frames in the dataset ($\approx 779.5k$ frames).

Although it can be expected that the speaker recognition should improve with better alignments, the posterior values per frame obtained from the lattices with the optimal AM and LM likelihoods scaling parameters are extremely sparse. For instance, when the posteriors are thresholded, that is posteriors less than a certain value (e.g. 10^{-5}) are floored to 0.0, the speech frame is no longer aligned to the true senone in $\approx$ 17% of the frames. Thus, even though the alignment obtained after decoding of HMM/SI-DNN is more accurate compared to using only the posteriors after forward-pass, such low scoring posteriors do not contribute to the SS. Thus, the likelihoods are re-scaled prior to the forward-backward algorithm. The best scaling, in terms of SRR, was obtained when the AM was 0.01 and the LM scale was 0.0. Other values for LM scale were also explored, but it proved beneficial to ignore the LM likelihoods once the recognition lattices are generated. The LM contribution is still available in the refined alignments.

VII. RESULTS

In this section, we discuss the results of our experiments using all the systems described earlier. First, the speaker verification performances are presented, followed by the results of the experiments on the amount of speaker content available on the different types of posteriors used. The motivation for the latter is provided by results that acoustic model-based posteriors provided significantly better speaker recognition performance. Thus, to observe whether the posteriors provide

TABLE III
COMPARISON OF SPEAKER RECOGNITION PERFORMANCE IN TERMS OF EQUAL ERROR RATE (EER)/MINIMUM DECISION COST FUNCTION (MINDCF) WHEN USING DIFFERENT POSTERIOR EXTRACTION TECHNIQUES, NAMELY UBM-GMM, DNN AND SGMM

System	Decoding (Y/N)	Speaker Adaptation (Y/N)	Cond1	Cond2	Cond3	Cond4	Cond5
Baseline							
UBM-GMM	×	×	1.41/0.22	2.43/0.40	1.58/0.38	1.32/0.35	2.25/0.28
MFCCs with Context							
UBM-GMM from MFCCs with Context	×	×	1.03/0.20	1.78/0.27	1.28/0.20	0.98/0.24	1.97/0.24
HMM/GMM based ASR systems							
HMM/SI-GMM	✓	✓	0.83/0.18	1.69/0.25	1.33/0.14	0.74/0.19	1.32/0.16
HMM/SD-GMM	✓	✓	0.97/0.13	1.41/0.18	0.60/0.08	0.63/0.17	1.54/0.16
SD-DNN forward pass	×	×	0.7/0.15	1.09/0.15	0.77/0.08	0.55/0.15	1.18/0.13
HMM/SD-DNN	✓	✓	0.78/0.14	1.13/0.13	0.80/0.09	0.50/0.13	1.00/0.13
SI-DNN forward pass	×	×	0.65/0.16	1.12/0.17	0.66/0.09	0.55/0.14	1.02/0.16
HMM/SI-DNN	✓	✓	0.81/0.15	1.36/0.17	0.73/0.09	0.48/0.14	0.91/0.12
HMM/SI-DNN (bigram)	✓	✓	0.88/0.16	1.63/0.22	1.16/0.20	0.98/0.19	1.23/0.19
HMM/SGMM based ASR system							
HMM/SGMM	✓	✓	1.17/0.25	2.32/0.32	1.61/0.31	1.17/0.24	1.58/0.22
Bottleneck features							
UBM/GMM (MFCC+BN)	×	×	0.87/0.1	1.35/0.2	0.48/0.1	0.33/0.16	0.84/0.11

speaker discriminability by themselves or through the i-vector framework.

A. Speaker verification results

The results on 5 conditions (cond1 through cond5) of the NIST SRE 2010 dataset are presented in Table III. The systems that use decoding and speaker adaptation are marked accordingly. Both Equal Error Rates (EER) and minDCF (minimum Decision Cost Function) values are reported. A target prior of 0.0001 was used with equal cost of 1.0 for false alarms and misses. The baseline system is the conventional i-vector PLDA system as described in Section III. For the matching microphone conditions the EER is already as low as 1.4%. For mismatched conditions that have a large number of trials, such as cond2, the EER was approximately 2.4%.

The speaker verification results of the HMM-based systems are given next. The results are presented in the order of model complexities of the systems - from HMM/GMM systems to hybrid HMM/DNN systems.

The senone posteriors obtained after decoding from the HMM/SI-GMM system already provided benefits to the speaker verification system. Although the framework for integrating acoustic class-based posteriors from ASR systems exist already, these results have seldom been reported. In this system, significant improvements are observed for all conditions. Absolute improvements in EER of up to ≈0.9% are obtained. In case of cond5, this translates into a relative improvement of ≈41%. Thus, even with a less sophisticated ASR system whose WER is worse by 10.7% relative than the state-of-the-art (see Table I) it is possible to obtain considerable improvements from the baseline system. The results clearly demonstrate the significance of constraining the acoustic space linguistically through ASR systems although the availability of large amounts of annotated data has its costs.

With the HMM/SD-GMM, further improvements were observed over the HMM/SI-GMM especially with respect to minDCF. In cond1, while there is clear loss in EER by 17% relative, the minDCF improved by 27.8% relative. The relative improvements in minDCF vary from none (cond5) to 57% (cond3). The system also provides the best performance among all systems for cond3 both in terms of EER and minDCF. Overall, speaker adaptation techniques, in particular fMLLR, does lead to better speaker verification especially in interview conditions. This also points us to a trend that better ASR performance could lead to better speaker recognition performance.

We now present results on the DNN-based systems. The SI-DNN forward pass system presented in Table III is, in principle, similar to the system presented in [4]. We also observed gains similar to that presented in literature. Compared to the baseline system, relative improvements in EER was observed to be at least ≈**54**%. The minDCF improved between 27% (cond1) and **60**% (cond4) relative. The forward pass system was then compared with the same DNN-based ASR system that used full ASR decoding.

The HMM/SI-DNN system can be compared to the baseline system, and the SI-DNN forward pass system. The system clearly improved over the baseline i-vector system with relative improvements in EER ranging between 42.5% for cond1 and 63.6% for cond3, and relative improvements in DCF varying between 31.8% and 76.3%.

While HMM/SI-DNN is better in the telephone based conditions, it performed worse for the interview conditions. This suggests that speaker adaptation may provide some help in mismatched conditions. Overall, for the family of systems that used HMM/SI-GMM mentioned till now, speaker verification systems were noticed to improve always. The system also provided the best performance in cond5 where the EER is as low as 0.91% and the minDCF is 0.12. This leads to the most interesting result from our studies. A comparison between the SI-DNN forward pass system and the HMM/SI-DNN system reveals that significant performance gains can be achieved by exploiting the decoder output in ASR systems. The EER and the minDCF improved by 10.7% and 25% relative, respectively. However, we found no performance benefits in any of the mismatch conditions (cond2 to cond4). We also note that these improvements vanish when using different acoustic scales, for example the AM likelihood scale regularly used for optimal ASR performance. Thus, direct application of the decoder output to speaker recognition does not necessarily help. In case of cond5, the EER reduced by 0.3% absolute while the minDCF increased to 0.19. Therefore, we believe that the highly sparse nature of the posteriors obtained after decoding has a detrimental effect on the performance even though decoding with word LM provides better frame-level alignments.

Similarly, the HMM/SD-DNN forward pass system and HMM/SD-DNN system were compared. Contrary to the results showed earlier with the SI-DNN systems, the system performance improves in two of the mismatched conditions (cond2 and cond4), while for the other mismatch condition (cond3), the minDCF does not change. This once again indicates that using speaker adaptation techniques can help negotiate channel variabilities. The HMM/SD-DNN forward pass system performs notably better than the HMM/SI-DNN in all but one condition (cond5). The minDCF reduces by only 0.01, from 0.12 to 0.13, even in the worst case.

As expected, the HMM/SD-DNN system further improves the performance over the HMM/SD-GMM system. Relative improvements in EER of up to 57% were achieved. Relative improvements in EER with respect to the baseline were ranging from ranging from 45% to as high as 63%.

In the results discussed so far, a strong correlation between the SRR as presented in Table II and the speaker verification metrics, especially for the telephone condition (cond5),can be seen. The EER decreases with the increase in SRR suggesting that better alignment of the features with mixture components could lead to better speaker modelling.

The use of phone LM over word LM in the HMM/SI-DNN (bigram) systems lead to performance deterioration. The minDCF doubled in cond3 while it increased from 0.12 to 0.19 in cond5. The results reiterate the importance of the choice of LM, but in another direction also show promising results towards the application of ASR techniques focused toward supporting multiple languages, which are certainly useful in practical scenarios.

Based on the results presented above it can be concluded that the discriminative modelling of acoustic classes influences speaker recognition performance in the i-vector framework. Another difference between the HMM/SI-GMM system and the HMM/SI-DNN system is the use of a temporal context of duration 200 ms. To fairly compare these two systems, the HMM/SI-GMM system was trained using MFCC involving the same context as the HMM/SI-DNN system to compute SS. These results are also presented in Table III under "MFCCs with context". Simply using context-based MFCC improves the recognition performance significantly. The best relative EER improvement achieved was $\approx 27\%$. Interestingly, in all conditions the performances are also better than the HMM/SI-GMM system (although there is only a marginal difference in minDCFs).

TABLE IV
SPEAKER DISCRIMINATIVE PROPERTIES OF POSTERIORS: THE STUDY INDICATES THE AMOUNT OF SPEAKER INFORMATION PRESENT IN THE POSTERIORS BY USING ZEROTH-ORDER STATISTICS AS I-VECTORS

System	Cond5 (EER in %)
UBM-GMM	24.5
HMM/SI-GMM	40.3
HMM/SD-GMM	40.3
HMM/SD-DNN (forward pass)	42.0
HMM/SI-DNN (forward pass)	41.7

Next, the performance of the HMM/SGMM system is also presented to compare alternative acoustic modelling techniques. The performance of the proposed approach using SGMM posteriors is consistently better compared to both the baselines in all but one of the cases (cond3), in which it worse by 0.1%. A best case absolute improvement of 0.7% is obtained on cond5. The HMM/SGMM system is worse than the HMM/SI-GMM in the interview conditions and has only marginally better minDCFs for telephone conditions. It performs consistently worse than the HMM/SD-GMM in all the conditions.

Finally, we compare the effect of using BN features along with MFCC in the UBM-GMM i-vector system. A direct comparison with the HMM/SI-DNN forward pass system shows significant difference in EERs and minDCFs for many conditions. Notably, the performances of the BN system in cond1 and cond5 are significantly better than the HMM/SI-DNN forward pass system. However, the performance for other conditions show very little difference in minDCF. The EER of the system for cond2 is worse by 17% relative. Note that cond2 has the most of number of trials. In general, the forward pass system works consistently better for mismatch conditions as evidenced by the differences in minDCF. This suggests that the two methods may be providing complimentary information that can be further exploited.

B. *Speaker verification from posteriors*

To further investigate the influence of the posteriors used to compute sufficient statistics we compared them for the amount of speaker information they contained. This would suggest if the posteriors indeed represent speaker-independent acoustic classes or speaker-dependent acoustic classes and also direct future research on how to train the posterior-extraction for optimal speaker recognition.

In this experiment, the zeroth-order statistics were used as i-vectors. That is, the i-vectors for an audio recording were obtained by accumulating posteriors over time and renormalizing them so that the elements of the vector sum to 1. The i-vector now is simply the zeroth-order statistics, which are a stack of N_c vectors as defined in Eq. 5. Two i-vectors were compared using the KL divergence measure. If $\mathbf{w}_1$ and $\mathbf{w}_2$ are two i-vectors, the KL divergence is given as follows

$$\mathrm{KL}\left(\mathbf{w}_1,\mathbf{w}_2\right)=\sum_{i=1}^{C}\left(w_{1,i}log\frac{w_{1,i}}{w_{2,j}}\right) \tag{9}$$

where C is the number of mixture components of a GMM, or the number of senone classes in the ASR system as the case may be, $w_{1,i}$ are the elements of the vector $\mathbf{w}_1$, $w_{2,i}$ are the elements of the vector $\mathbf{w}_2$.

Two i-vectors are similar if the KL divergence is closer to 0. Even though state-of-the-art speaker recognition performance is not expected, the error rates can suggest the amount of speaker content available in the posteriors. The results of the speaker verification experiments are presented in Table IV. Once again, the systems are only evaluated on the the core condition 5 (cond5) of the NIST SRE 2010 dataset. The trends on other conditions are the same. A major trend in these results point to the speaker dependence of the UBM-GMM posteriors and the lack thereof in the ASR posteriors. While the EERs of the UBM-GMM system are in the range of 15% to 25%, of the ASR based systems are between 40% and 45%. The clear contrast in the results and the corresponding speaker verification results with the i-vector PLDA systems (see Table III) suggest that the supervised training of the acoustic classes is beneficial for i-vector systems. In fact, the better the ASR performance the more speaker independence. However, comparing Tables III and IV indicate that more speaker independence can lead to better speaker verification.

C. System fusion results

In this section, the results of score fusion of a subset of the systems presented in Table III are presented. Only the DNN based systems are fused together. This is primarily done to exploit the difference in system performances across all five conditions, especially between the SI-DNN and SD-DNN systems.

The results of fusion are presented in Table V. In general, a combination improved system performance considerably when systems that used decoder and those did not were combined. The combination of systems that used only forward pass posteriors did not provide significant benefits. Only minor improvements in terms of the minDCF are observed when fusing the scores from the HMM/SI-GMM system and the HMM/SI-GMM forward pass system. The combination of HMM/SI-DNN with the forward pass systems showed significant improvements. In mismatch conditions, as hypothesized earlier, with the fusion of HMM/SI-DNN and HMM/SD-DNN (forward pass) the minDCFs were observed to reduce for all conditions. The best results for cond2 and cond5 were thus obtained, with minDCF values of 0.13 and 0.10, respectively. The results while fusing the HMM/SI-GMM with the HMM/SI-GMM forward pass systems are quite similar. Once again, the EERs and minDCFs reduced in all conditions showing favorable indications of combining systems using posteriors from different stages of the ASR system.

VIII. Conclusions

This paper presented a comparison of different approaches to SS estimation using different ASR systems for i-vector based speaker recognition. We improved on the existing technique that uses posteriors obtained from a DNN trained for ASR by passing them to the decoder. The lattices generated by the ASR decoder were used for SS estimation with the motivation to use better alignment to compute these statistics. ASR systems employing different acoustic modelling methods were studied. Our results reveal that the proposed method of using rescaled AM posteriors from ASR lattice can indeed improve speaker verification performance. We also showed that the performance gains are positively correlated to the senone recognition accuracy of the models. The EER of the HMM/SI-DNN forward pass system improves by 10.7% relative corresponding to a relative gain of 25% in minDCF.

The use of ASR systems that employed the fMLLR speaker adaptation technique provided added benefits in channel mismatch conditions for speaker verification. Thus, further improvements were observed from the score-level fusion of the HMM/SD-DNN and HMM/SI-DNN based systems. Along with the systems presented, we also experimented with the HMM/SGMM system.

An independent experiment was conducted to isolate the effect of using features with a long context. Our results showed that the presence of a context significantly improves the standard i-vector system performance. A best case relative improvement of 27% was obtained.

IX. Acknowledgements

This work was supported by the EC FP7 funding under “Speaker Identification Integrated Project (SIIP)”.

References

[1] P. Kenny, P. Ouellet, N. Dehak, V. Gupta, and P. Dumouchel, “A study of inter-speaker variability in speaker verification,” *IEEE Tran. on Audio, Speech and Language Processing*, vol. 16, no. 5, pp. 980–988, 2008.

[2] S. J. Prince and J. H. Elder, “Probabilistic linear discriminant analysis for inferences about identity,” in *IEEE 11th International Conference on Computer Vision (ICCV)*. IEEE, 2007, pp. 1–8.

[3] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, “Front-end factor analysis for speaker verification,” *IEEE Tran. on Audio, Speech and Language Processing*, pp. 788–798, 2011.

[4] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, “A novel scheme for speaker recognition using a phonetically-aware deep neural network,” in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 1695–1699.

[5] S. Madikeri, S. Dey, and P. Motlicek, “Analysis of language dependent front-end for speaker recognition,” *Proc. Interspeech 2018*, pp. 1101–1105, 2018.

[6] D. Snyder, D. Garcia-Romero, and D. Povey, “Time delay deep neural network-based universal background models for speaker recognition.” ASRU, 2015.

[7] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, “Deep neural networks for extracting baum-welch statistics for speaker recognition,” in *Proc. Odyssey*, 2014.

[8] P. Matejka, O. Glembek, O. Novotny, O. Plchot, F. Grezl, L. Burget, and J. Cernocky, “Analysis of DNN approaches to speaker identification,” *Proc. IEEE ICASSP, Shanghai, China*, pp. 5100–5104, 2016.

[9] D. Klusácek, J. Navratil, D. A. Reynolds, and J. P. Campbell, “Conditional pronunciation modeling in speaker detection,” in *Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP’03). 2003 IEEE International Conference on*, vol. 4. IEEE, 2003, pp. IV–804.

[10] W. D. Andrews, M. A. Kohler, J. P. Campbell, J. J. Godfrey, and J. Hernández-Cordero, “Gender-dependent phonetic refraction for speaker recognition,” in *Acoustics, Speech, and Signal Processing (ICASSP), 2002 IEEE International Conference on*, vol. 1. IEEE, 2002, pp. I–149.

TABLE V
COMPARISON OF SPEAKER RECOGNITION PERFORMANCE IN TERMS OF EQUAL ERROR RATE (EER)/MINIMUM DECISION COST FUNCTION (MINDCF) WHEN FUSING THE SI-DNN FORWARD PASS SYSTEM WITH OTHER ASR SYSTEMS THAT USE A DECODING COMPONENT

System	Cond1	Cond2	Cond3	Cond4	Cond5
HMM/SI-DNN + HMM/SI-DNN (forward pass)	0.68/0.13	1.15/0.14	0.57/0.08	0.43/0.14	0.93/0.10
HMM/SI-DNN + HMM/SD-DNN	0.78/0.15	1.17/0.14	0.73/0.09	0.43/0.14	0.95/0.12
HMM/SI-DNN + HMM/SD-DNN (forward pass)	0.71/0.13	1.09/0.13	0.64/0.06	0.44/0.14	0.95/0.10
HMM/SI-DNN (forward pass) + HMM/SD-DNN (forward pass)	0.68/0.15	1.08/0.15	0.71/0.07	0.50/0.14	1.13/0.15

[11] W. M. Campbell, J. P. Campbell, D. A. Reynolds, D. A. Jones, and T. R. Leek, “Phonetic speaker recognition with support vector machines,” in *Advances in neural information processing systems*, 2004, pp. 1377–1384.

[12] G. R. Doddington *et al.*, “Speaker recognition based on idiolectal differences between speakers.” in *INTERSPEECH*, 2001, pp. 2521–2524.

[13] P. Motlicek, S. Dey, S. Madikeri, and L. Burget, “Employment of subspace gaussian mixture models in speaker recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on.* IEEE, 2015, pp. 4445–4449.

[14] P. Kenny, G. Boulianne, and P. Dumouchel, “Eigenvoice modeling with sparse training data,” *IEEE Trans. on Speech and Audio Processing*, vol. 13, no. 3, pp. 335–354, May 2005.

[15] S. O. Sadjadi, S. Ganapathy, and J. W. Pelecanos, “The ibm 2016 speaker recognition system,” *arXiv preprint arXiv:1602.07291*, 2016.

[16] “The NIST Year 2010 Speaker Recognition Evaluation Plan, http://www.itl.nist.gov/iad/mig/tests/sre/2010/index.html.”

[17] O. Glembek *et al.*, “Simplification and optimization of i-vector extraction.” In Proc. of ICASSP, 2011, pp. 4516–4519.

[18] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* John Wiley & Sons, 2012.

[19] S. Ioffe, “Probabilistic linear discriminant analysis,” in *Computer Vision–ECCV 2006.* Springer, 2006, pp. 531–542.

[20] D. Garcia-Romero and C. Y. Espy-Wilson, “Analysis of i-vector length normalization in speaker recognition systems.” In Proc. of Interspeech, August 2011, pp. 249–252.

[21] F. Richardson, D. Reynolds, and N. Dehak, “Deep neural network approaches to speaker and language recognition,” *Signal Processing Letters, IEEE*, vol. 22, no. 10, pp. 1671–1675, 2015.

[22] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[23] F. Seide, G. Li, and D. Yu, “Conversational speech transcription using context-dependent deep neural networks.” in *Interspeech*, 2011, pp. 437–440.

[24] I. Lopez-Moreno, J. Gonzalez-Dominguez, D. Martinez, O. Plchot, J. Gonzalez-Rodriguez, and P. J. Moreno, “On the use of deep feedforward neural networks for automatic language identification,” *Computer Speech & Language*, vol. 40, pp. 46–59, 2016.

[25] P. Rose, *Forensic speaker identification.* CRC Press, 2003.

[26] S. M. Siniscalchi, P. Schwarz, and C.-H. Lee, “High-accuracy phone recognition by combining high-performance lattice generation and knowledge based rescoring,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 4, 2007, pp. IV–869.

[27] A. Johnson and D. Wright, “Identifying idiolect in forensic authorship attribution: an n-gram textbite approach,” *Language and Law/Linguagem e Direito*, vol. 1, no. 1, pp. 37–69, 2014.

[28] A. P. Dempster, N. M. Laird, and D. B. Rubin, “Maximum likelihood from incomplete data via the em algorithm,” *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38, 1977.

[29] C. M. Bishop, “Pattern recognition and machine learning,” 2006.

[30] J. A. Bilmes *et al.*, “A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models,” *International Computer Science Institute*, vol. 4, no. 510, p. 126, 1998.

[31] M. McLaren, L. Ferrer, and A. Lawson, “Exploring the role of phonetic bottleneck features for speaker and language recognition,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2016, pp. 5575–5579.

[32] J. Pelecanos and S. Sridharan, “Feature warping for robust speaker verification.” In Proc. of Speaker Odyssey, 2001, pp. 213–218.

[33] D. Povey, A. Ghoshal *et al.*, “The kaldi speech recognition toolkit,” in *In Proc. of ASRU 2011*, December 2011.

[34] S. Madikeri, S. Dey, P. Motlicek, and M. Ferras, “Implementation of the standard i-vector system for the kaldi speech recognition toolkit,” Idiap, Tech. Rep., 2016.

[35] M. Gales and S. Young, “The application of hidden markov models in speech recognition,” *Foundations and trends in signal processing*, vol. 1, no. 3, pp. 195–304, 2008.

[36] P. Motlicek, D. Povey, and M. Karafiat, “Feature and score level combination of subspace gaussinas in lvcsr task,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 7604–7608.

[37] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow *et al.*, “Subspace gaussian mixture models for speech recognition,” in *IEEE Intl. Conf. on Acoustics Speech and Signal Processing (ICASSP)*, 2010, pp. 4330–4333.

[38] G. E. Dahl, D. Yu, L. Deng, and A. Acero, “Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition,” *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 30–42, 2012.

[39] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, “Why does unsupervised pre-training help deep learning?” *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.