CSLT TECHNICAL REPORT-20150002 [JANUARY 10, 2015]

DNN-based Discriminative Scoring for Speaker Recognition Based on i-vector

Jun Wang[1], Dong Wang[1], Thomas Fang Zheng[2*] and Fanhu Bie[1]

[*]Correspondence:
fzheng@tsinghua.edu.cn
[2]Center for Speech and Language Technologies, Tsinghua University, ROOM 4-416, Information Sci & Tech Building, Tsinghua University, 100084 Beijing, China
Full list of author information is available at the end of the article

Abstract

One of the state-of-the-art approaches to speaker recognition is based on factor analysis, especially the i-vector model. By representing a speech segment as a vector in a low-dimensional vector space, the i-vector model can deal with the complex correlation among components of the Gaussian mixture model (GMM). On the other hand, it is well known that i-vectors contain both speaker and session variances, and therefore additional discriminative approaches are required to emphasize the speaker-dependent information in the 'total variance' space. Among various methods, the probabilistic linear discriminant analysis (PLDA) achieves the significant performance, partly due to its generative model framework that represents the speaker and session variances in a hierarchical way. A disadvantage of PLDA, however, lies in its Gaussian assumptions of the speaker and session variables, which is not necessarily true in most situations.

This paper presents a discriminative scoring approach for i-vector-based speaker recognition based on deep neural networks (DNN). This approach casts the recognition task to a binary classification problem and employs the DNN model to learn the complex decision boundary in the heterogeneous speaker space. Compare with the PLDA-based approach, the new approach does not rely on any artificial assumption on the distribution of data, and can optimize the model with respect to the recognition task directly. Our experiments on the NIST SRE08 core test demonstrate that the DNN-based approach outperforms the PLDA-based approach, and find that combining the DNN and PLDA scores leads to further gains. Finally, we compare the DNN model with a discriminative but shallow model, the support vector machine (SVM), and find that the DNN clearly outperforms the SVM, confirming the advantage of deep learning.

Keywords: DNN; i-vector; PLDA; speaker recognition

1 Introduction

Joint factor analysis (JFA) has gained much success in speaker recognition. This approach assumes that the speaker variance and session variance are derived from two independent random variables (factors) that follow the standard Gaussian distribution as a prior (usually in a low-dimensional subspace). The speaker representation of a speech segment is then derived by inferring the posterior probability of the speaker factor given the speech signal [1]. Recent research reveals that speaker and session variances may not be clearly separated by JFA, and the session factor inferred from JFA may still contain some speaker information. A better approach would be representing speaker and session variances as a single 'total variance' factor, so that more speaker information can be retained in the posterior inference. By this approach, a speech segment can be represented by an 'i-vector' which

corresponds to the mean vector of the inferred posterior distribution of the total variance factor. This is widely known as the total variance model or i-vector model [2]. PLDA [3] scoring method is widely used in i-vector and achieves the significant performance. Also, some discriminative scoring method are utilized in i-vector especially for speaker verification task. In the following sections, we'll briefly introduce PLDA scoring and discriminative scoring.

1.1 PLDA scoring

Involving both speaker and session variances is a particular advantage of the i-vector model as more speaker-related information is retained; however, at the same time, the disadvantage is also obvious: the 'mixed' representation leads to less discrimination among speakers. It is therefore important to employ some discriminative approaches to suppress the session variance and accentuate the speaker variance. For example, the with-in class covariance normalization (WCCN) technique employs a linear transform that is derived by optimizing a generalized linear kernel [4], and the nuisance attribute projection (NAP) seeks a projection that minimizes the discrepancy of signal pairs recorded in different channels [5]. These approaches, although originally proposed for the SVM-based approach, have been demonstrated to be effective in i-vector systems, as a post-processing to enhance speaker discrimination with i-vectors [6].

Another approach that remarkably improves the representative power of i-vectors is the probabilistic linear discriminant analysis (PLDA) [3]. On one hand, PLDA is a probabilistic version of LDA and so inherits LDA's discriminative nature; on the other hand, PLDA is a generative model which places a Gaussian prior on the underlying class variable, and so can model classes with very limited training data. This is a big advantage of PLDA in speaker recognition, since in most situations only very few utterances are available for enrollment and testing.

1.2 Discriminative scoring

In spite of the success of PLDA, there are still some limitations with this model. Particularly, the model assumes that the prior probability of the class variable and the conditional probability of i-vectors are both Gaussian. This is not necessarily true in practice. In addition, the speaker recognition task is essentially a classification task, i.e., distinguishing genuine speakers and imposters, for which a discriminative model is a more reasonable choice.

We therefore seek a discriminative model which relaxes the Gaussian assumption and scores a trial by predicting the posterior that the trial is from the genuine speaker. A possible approach of this category is to train a one-vs-all model for each speaker; in the test phase, the confidence score of a trial is derived from these speaker-dependent models. A popular discriminative model that is employed in this way is the support vector machine (SVM) [5, 7, 8], which has been widely used in the transform-based systems (e.g., based on MLLR spervectors [9, 10, 11]) and can be easily migrated to i-vector systems [2, 12]. An obvious problem of this one-vs-all strategy is that the positive samples are highly sparse and so the resulted model may be highly biased. In addition, keeping an SVM for each speaker is awkward for large-scale applications.

A more ideal approach is to train a universal classifier for all speakers. The input is a pair of i-vectors, with one i-vector representing the ascertained speaker and the other representing the tested utterance, and the output is the probability of the two i-vectors belonging to the same speaker. For example, an SVM can be used as such a classifier, and with the cosine kernel, the SVM falls back to the widely used cosine distance measure.

1.3 Motivation

Although complex kernels can be used to improve the discriminative power of the SVM, such as with the Kullback–Leibler divergence, it was found in experiments that the improvement is rather marginal. We argue that a possible reason is that the i-vector space is heterogeneous, i.e., the decision boundary varies in a complex way, so that it is difficult to be linearly separated even in the high-dimensional projection space (represented by the kernel function implicitly). This shortage of the SVM model in complex decision tasks is largely attributed to the fact that it is a shallow model and the feature mapping (the mapping from the i-vector space to the projection space, represented by the kernel function) is not trainable, so it is difficult to learn complex decision boundaries, even with a non-linear kernel. A better discriminative model would be capable of learning the feature mapping based on the training data, and hence learns the decision boundary in a more flexible way.

Recently, deep learning gains tremendous success in machine learning and signal processing. A representative model in the deep learning regime is the deep neural network (DNN), which is essentially a neural network that involves many hidden layers. A particular power of the DNN is that it can learn complex and high-order features from primary features. This capability has been used in speech recognition to learn more powerful features from the primary spectrum. The 'learned feature' has substituted for the conventional Mel frequency cepstral coefficient (MFCC) that was deliberately designed by researches and has dominated speech recognition for several decades. Interestingly, the DNN is particularly powerful in learning decision boundaries in heterogeneous conditions, e.g., in the acoustic space with multiple noises, as has been shown in [13].

In this paper, we present a DNN-based scoring approach for i-vector-based speaker recognition. The basic idea is to employ the power of the DNN in learning high-order features from primary inputs and its capability in learning heterogeneous conditions to learn the complex decision boundary in the heterogeneous i-vector space. Specifically, we first derive some primary discriminative features from the i-vector pairs, and then train a DNN model that learns the genuine speaker/imposter decision from the primary features. Our experiments demonstrated that this approach is highly effective and can achieve better performance than the state-of-the-art PLDA model in the NIST SRE08 core test.

Note that the idea of discriminative scoring by neural networks has been published in [14]. This paper is an extension of that work with more materials on deep models and a complete set of experiments. In addition, we notice a similar work was proposed by [15] at the time this study was conducted. The difference is that our work uses different discriminative features and focuses on comparison with PLDA.

The rest of the paper is organized as follows: Section 2 gives a brief introduction for the i-vector, PLDA and DNN techniques, for the sake of completeness. Section 3

presents the DNN-based scoring approach, and Section 4 presents the experiments. The paper is concluded in Section 5.

2 Theory background

For presentation convenience, we'll first go through the i-vector model and PLDA, then we'll briefly introduce deep neural network (DNN).

2.1 i-vector model

The conventional approach to speaker recognition is based on the universal background model-Gaussian mixture model (UBM-GMM) architecture. The i-vector approach is an extension to the UBM-GMM approach and assumes that both speaker and session variances of a speech segment concentrate in a low-dimensional subspace of the model supervector (the concatenation of the mean vectors of all the GMM components). This subspace is referred to as the total-variance space, and a speech segment can be represented by an identity vector (i-vector) in this space.

Mathematically, letting the UBM involve C Gaussian components, and the acoustic features of a speech segment associated with the c-th component follow a Gaussian distribution with mean $\mathbf{M}_c$ and covariance $\mathbf{\Sigma}_c$. The i-vector model assumes that $\mathbf{M}_c$ is generated from a low-dimensional variable $\mathbf{w} \in R^M$ which follows a Gaussian distribution, via a linear transformation $\mathbf{T}_c$:

$$\mathbf{M}_c = \mathbf{m}_c + \mathbf{T}_c\mathbf{w} \tag{1}$$

where $\mathbf{m}_c \in R^D$ is the mean of the c-th component of the UBM, and $\mathbf{T}_c \in R^{D\times M}$ is the loading matrix associated with the c-th component. The speaker factor $\mathbf{w}$ follows the standard normal distribution $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The loading matrices $\{\mathbf{T_c}\}$ can be trained by an EM procedure [16]. Once $\{\mathbf{T_c}\}$ has been obtained, a speech segment X can be represented by the posterior probability $p(\mathbf{w}|X)$ which can be inferred according to (1). Specifically, since the prior $p(\mathbf{w})$ is a Gaussian, the posterior $p(\mathbf{w}|X)$ is a Gaussian as well:

$$p(\mathbf{w}|\mathbf{X}) \sim \mathcal{N}(\bar{\mathbf{w}}, \mathbf{\Xi}) \tag{2}$$

where the mean vector $\bar{\mathbf{w}}$ and covariance matrix $\mathbf{\Xi}$ can be computed from the zero- and first- order statistics of X. Details of the derivation can be found in [2].

In speaker recognition, the mean vector $\bar{\mathbf{w}}$ is taken as the identity vector (i-vector) of the speech segment, and the true/imposter decision is conducted based on the distance (cosine distance is an often choice) between the i-vectors of the test speech and the enrollment speech. Note that an i-vector involves both speaker and non-speaker (e.g., channel, content, emotion) information. In order to improve the discriminative capability of i-vectors for speakers, transform approaches such as WCCN [4], NAP [5] and LDA [17] are usually applied before computing the distance.

2.2 PLDA

It is well known that the linear discriminant analysis (LDA) corresponds to a generative model given by:

$$\mathbf{w}_{i,j} = \tilde{\mathbf{w}}_i + \mathbf{A}\mathbf{u}$$

where $\mathbf{w}_{i,j}$ is an observation vector (i-vector in speaker recognition) of the i-th class (the i-th speaker in speaker recognition), $\tilde{\mathbf{w}}_i$ is the mean vector of the class, and $\mathbf{u}$ follows a Gaussian distribution: $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. This formulation implies that the LDA assumes that the class conditional distributions $p(\mathbf{w}_{i,j}|\tilde{\mathbf{w}}_i)$ are Gaussian and share the same covariance matrices among classes. Ioffe [3] extended this model by placing a Gaussian prior on $\tilde{\mathbf{w}}_i$, which leads to a hierarchical Bayesian model shown in Figure 1.

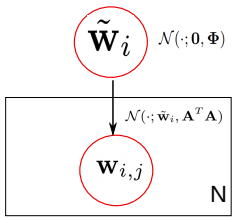

Figure 1 The graphical model of PLDA, where N is the number of samples of class i.

By this extension, the class mean $\tilde{\mathbf{w}}_i$ is treated as a continuous variable instead of a discrete parameter as in the traditional LDA. This significantly improves model generability and thus classes with very few samples can be well represented due to the prior [3]. This is particularly attractive for speaker recognition where in most cases only a few enrollment/test utterances (and hence i-vectors) are available for a speaker. A multitude of researches reported that the PLDA can significantly improve performance of i-vector systems and achieve the state-of-the-art performance [6].

2.3 DNN

Deep neural networks (DNN) have gained brilliant success in many research fields including speech recognition, computer vision (CV), and natural language processing (NLP) [18]. A DNN is a neural network (NN) that involves more than one hidden layers. NNs have been studied in the speech community for a decade. For example in speech recognition, the NN has been used to substitute for the GMM to produce frame likelihood [19], or to produce long-context features that are used to substitute for or augment to short-time features, e.g., MFCCs [20]. Although promising, the NN-based approach did not deliver overwhelming superiority over the conventional approaches. The revolution took place in the ASR community in 2010, after a close collaboration between academic and industrial research groups, including the University of Toronto, Microsoft, and IBM [18, 21, 22]. These researches found that very significant performance improvements can be accomplished with DNNs when appropriate initialization is conducted, e.g., by pre-training with restricted Boltzmann machines (RBMs).

Encouraged by the success in ASR, the DNN (and the unsupervised version, deep Boltzmann machine (DBM)) model has been investigated in a wide range of research fields of speech processing, including speech synthesis [23, 24], music pattern analysis [25, 26], speech enhancement [27, 28], voice activity detection [29], and music recommendation [30]. Particularly, a very recent study applies DNN to speaker recognition [31, 32]. The basic idea is to use a DNN model trained for speech recognition to substitute for the UBM, so that the rich information in phones can be employed to build an accurate conditional probability model than the Gaussian models of the GMM that are trained in an unsupervised way. In this paper, we employ the capability of the DNN model in learning complex discriminative functions to score the distance of i-vector pairs, i.e., the probability that a pair of i-vectors belong to the same speaker.

3 DNN-based discriminative scoring

3.1 Concern for PLDA

In spite of the success of PLDA in speaker recognition, this model possesses some limitations, particularly the underlying Gaussian assumption on the prior distribution of the mean of the speaker classes and the conditional distribution of the i-vectors of a speaker. There is little justification for this assumption, except the concern on computational tractability in model training and inference. We notice that this assumption can be relaxed to some extent by replacing Gaussians with Gaussian mixtures, as mentioned in [33], however this will greatly improve model complexity and the effectiveness has not yet been demonstrated in speaker recognition. Another concern for PLDA is the generative modeling itself: the optimization objective is to fit the data. Although the fitting takes class discrimination into account, it is still suboptimal with respect to the recognition task, i.e., the task of true/imposter decision. A desirable model should be discriminative in nature, and the optimization criterion should be related to the true speaker/imposter decision error rate.

A simple discriminative approach designs a one-vs-all classifier for each class, as has been used in conventional SVM-based systems and migrated to i-vector systems [34]. This approach, however, needs to build many classifiers and suffers from data sparsity and data imbalance, since the positive samples (i-vectors of genuine speakers) are often rare and much less than the negative samples (i-vectors of imposters). An ideal approach is to build a single classifier that can make decisions for all speakers, as the PLDA approach does. The most straightforward way is to collect a number of i-vector pairs and label them as the same or different speakers, and then train a discriminative model, e.g., SVM, to predict the posterior probability that a pair of i-vectors belong to the same speaker. This approach, according to our experiments, is promising on a small set of enrollment speakers; however, when the number of speakers increase, the performance decreases drastically.

We argue that the difficulty with the i-vector pair modeling is two fold: firstly, the i-vector pair is a purely raw feature, and it is hard to learn a reasonable discriminative model based on the raw feature with limited training data. This is analog to learning from raw speech signals in speech recognition where the performance is usually bad. To solve this problem, a feature extraction, even if very simple, is

necessary. Secondly, we argue that the SVM is a shallow model, so it can not deal with the speaker space that becomes highly complex and heterogeneous when the number of speakers becomes large. For such a problem, a deep model, such as DNN, is more appropriate, due to its capability of learning complex feature mappings with the low-level layers and complex decision boundaries with the high-level layers.

3.2 DNN-based scoring

We follow the argument in the previous section and design a new DNN-based discriminative scoring approach for i-vector pairs. The main process involves two steps: discriminative feature extraction and discriminative model training.

In the first step, the main task is to derive some features from the raw i-vector pairs, and the goal is to make it easy for the discriminative model to learn the decision boundary. The features should be simple and straightforward, and represent as much discriminative information as possible. There might be many choices but we choose the simple sub-vector Euler distance as presented in this section.

3.2.1 Primary discriminative feature extraction

First of all, a number of i-vector pairs $\{(v_{i,1}, v_{i,2})\}$ are collected and are labeled as positive (+1) and negative (-1) samples according to whether or not $v_{i,1}$ and $v_{i,2}$ belong to the same speaker, leading to a training set $\Delta = \{(v_{i,1}, v_{i,2}; l_i)\}$ where l_i is the label of the i-th pair. In order to obtain the most discriminative information while keeping the feature set compact, the LDA is applied to project the i-vector pairs to a low dimensional subspace, resulting in a projected training set $\Delta' = \{(v'_{i,1}, v'_{i,2}; l_i)\}$ where v' is the projection of v with the LDA.

A number of simple discriminative features are then extracted, leading to a ready-to-use training set $\Delta'' = \{(f_i; l_i)\}$, where f_i is the feature set derived from the pair $(v'_{i,1}, v'_{i,2})$. The sub-vector Euler distance was used as the features in this work, computed on the first n dimensions of the two vectors in the pair, i.e., $\{f_i(j) = |v'_{i,1}(j) - v'_{i,2}(j)|^2; j = 0, ..., n-1\}$. Note that $v'_{i,1}$ and $v'_{i,2}$ are in the LDA projection space and hence their first n dimensions are assumed to retain the most discriminative information. In addition, considering the success of the cosine distance in i-vector systems, it is also taken as a feature. In summary, the feature set involves $n + 1$ elements:

$$[f_i(0), f_i(1), ...f_i(n-1), \frac{< v'_{i,1}, v'_{i,2} >}{||v'_{i,1}||||v'_{i,2}||}] \tag{3}$$

where

$$f_i(j) = |v'_{i,1}(j) - v'_{i,2}(j)|^2. \tag{4}$$

3.2.2 Discriminative modeling

With the training data Δ'', a discriminative model can be constructed and optimized with respect to the decision error rate for genuine speakers and imposters. We experimented with a shallow model, the SVM, and a deep model, the DNN. For the SVM model, the primary discriminative features obtained from the previous

section is projected to the high-dimensional space by the kernel function (implicitly) where the decision is made by a linear model; for the DNN model, the primary discriminative features are projected to a high-dimensional space by the low-level layers of the DNN and the decision is made by a log-linear model at the top-level layer (the soft-max layer). We argue that the feature projection implemented by the DNN is trainable and hierarchical, so it is task-specific, and is able to learn high-order features that are more invariable for heterogeneous speakers. The DNN structure is shown in Figure 2, where the output s_{nn} corresponds to the posterior probability that the input i-vector pair represents the same speaker.

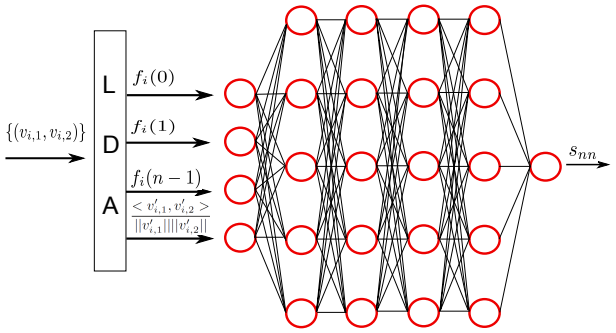

Figure 2 Architecture of the DNN-based scoring.

It should be noted that training the DNN model requires a balanced training set that involves the same positive and negative samples. With the balanced training set, the output of the model is an unbiased posterior, i.e., genuine speakers and imposters have equal priors. Therefore the model can not be used to make decisions directly. A threshold on the posterior s_{nn} needs to be determined on a development set to achieve the best performance in terms of the evaluation metric, which is the equal error rate (EER) in our work. From this perspective, the DNN-based approach is a scoring approach which extends the normally used cosine distance. In fact, if the feature set involves only the cosine distance, this DNN-based approach falls back to the cosine scoring.

We also highlight that the discriminative scoring architecture presented here is an instance of the well-known generative kernel approach [35, 36], where a discriminative model is based on a kernel machine, but the kernel function is derived from a generative model. Here the generative model is the i-vector model, and the kernel is the discriminative feature extraction process. The only difference is that we extended the architecture by using multiple kernels and use the kernels as blocks to build a new kernel machine (in the SVM-based scoring) or a neural model (in the DNN-based scoring).

3.3 PLDA-DNN combination

The advantage of the DNN-based approach, when compared with the PLDA approach, relies on the fact that it relaxes the Gaussian assumption of the later. This advantage, however, is evident only when the speaker space is rather complex. In many cases, the PLDA can model the speaker space well. In addition, the DNN

Table 1 Evaluation conditions reproduced from [37]

Trial condition	Number of trials	Description
c1	957	only interview speech in training and test
c2	17,941	interview speech from the same microphone type in training and test
c3	18,898	interview speech from different microphones types in training and test
c4	6,378	interview training speech and telephone test speech
c5	4,354	telephone training speech and noninterview microphone test speech
c6	22,152	only telephone speech in training and test
c7	10,607	English language telephone speech in training and test
c8	4,959	only English language telephone speech spoken by a native U.S. English speaker in training and test
Total	56,343	All trials in evaluation set

requires sufficient training data. In the condition where the training data is limited, the DNN approach is expected to be inferior to the PLDA approach, due to the lack of prior knowledge for distributions of i-vectors.

It is a natural idea to combine the two approaches and leverage their respective advantages. There are two possible ways to conducting the combination. In the feature-domain combination, the PLDA score, denoted by s_{plda}, is augmented to the original DNN inputs. The feature set hence is extended to:

$$[f_i(0), f_i(1), ...f_i(n-1), \frac{< v'_{i,1}, v'_{i,2} >}{||v'_{i,1}||||v'_{i,2}||}, s_{plda}]. \tag{5}$$

The second combination approach is in the score-domain, which combines the PLDA score and the DNN score by linearly interpolation, given by:

$$s_{cmb} = \alpha s_{nn} + (1-\alpha) s_{plda} \tag{6}$$

where α is a tunable parameter that can be determined on a development set.

4 Experiments and Analysis

4.1 Databases

In the experiments, the Fisher telephone speech database is used for training the systems, including the UBM model, the T matrix in the i-vector system, the LDA transform matrix, the PLDA model, the SVM and the DNN model. The training dataset involves $7,196$ female speech recordings, $12,837$ utterances in total.

The performance is evaluated on the core test of the NIST 2008 speaker recognition evaluation (SRE08) task [37]. All the experiments is conducted under speaker verification task. All the evaluation data are recordings of females; each enrollment segment and test speech segment consists a speech signal at least of 2 minutes. There are 8 test conditions in total, as have been reproduced in Table 1 from the NIST SRE2008 evaluation plan [37]. Note that part of the evaluation data have been selected as the cross-validation (CV) set to choose the hyperparameters in the DNN model, particularly the number of sub-vector distance features. The CV set involves 100 speakers and about $3,000$ trials in total. The number of trials of each condition in the evaluation has been shown in Table 1.

4.2 Experimental setup and baseline

All the speech data used in this study (training, development, evaluation) are sampled at 8 kHz and the sample precise is 16 bits. The acoustic feature used is the 19-dimensional Mel frequency cepstral coefficients (MFCC) together with the log energy. The first and second order derivatives are augmented to the static features, resulting in 60-dimensional feature vectors.

The UBM involves $2,048$ Gaussian components and was trained with about $4,000$ female utterances which were randomly selected from the training data (the Fisher database). The T matrix of the i-vector system was trained with all the female utterances in the training database, and the dimension of the i-vector is 400. The LDA and PLDA models were trained with utterances of $7,196$ female speakers, again randomly selected from the training database. The dimension of the LDA projection is set to 150.

Table 2 presents the performance of three baseline systems, based on i-vector, i-vector plus LDA and i-vector plus PLDA, respectively. The results shown in the table are the EERs on the entire evaluation set. Fig 3 presents the performance on all the 8 conditions. It is clear that both the LDA and PLDA systems outperform the i-vector system, and the PLDA system obtains the best overall performance, confirming the power of the PLDA model.

Table 2 Performance of three baseline systems on the entire evaluation set.

System	EER%
i-vector	21.89
LDA	18.46
PLDA	17.22

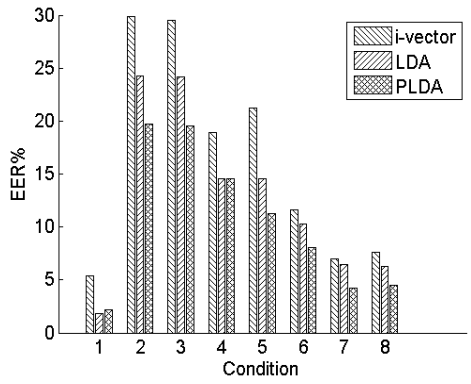

Figure 3 Performance of the baseline systems on different conditions.

4.3 Discriminative feature selection

We start to experiment with discriminative scoring models (SVM and DNN). In order to train the model, we select $32,500$ pairs of i-vectors (denoted by IP-TR) that are extracted using speech segments that are randomly selected from the Fisher database. As mentioned, the discriminative features are selected based on the first n dimensions of the LDA-projected i-vectors. We build a DNN model with features

selected with different n, then evaluate the performance of the DNN on the CV set and select the best n. For simplicity, the DNN structure is fixed, by setting the hidden layer to be 2 and the number of hidden units at each hidden layer to be 400.

Figure 4 shows the EER results on the CV set with different values of n. It can be seen that $n = 10$ is a good trade-off: a smaller n is worse, perhaps because of the lost of speaker information with the over compact LDA, and a larger n is also worse, due to the over-fitting towards non-speaker variance.

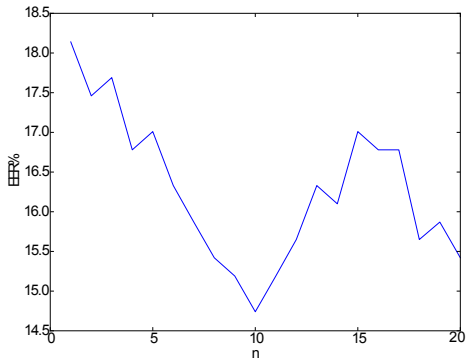

Figure 4 Performance of the DNN-based scoring as a function of the number of discriminative features. Results are reported in terms of EER on the CV set.

To investigate the generability of the CV-based feature selection, the DNNs built with different n are tested on three conditions (c1,c2,c3) of the evaluation task, leading to the results illustrated in Figure 5. It is observed that the curves on the evaluation set show similar patterns as on the CV set, although the optimal choices of n are not exactly the same. This suggests that the feature selection based on the CV set is well generalizable.

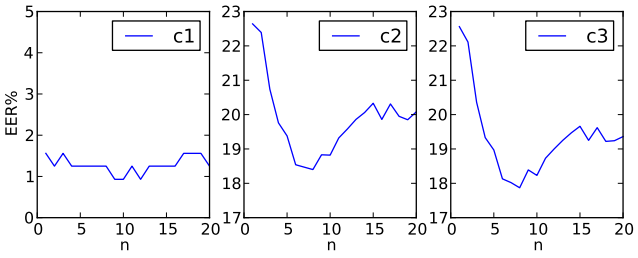

Figure 5 Performance of the DNN-based scoring with various number of discriminative features. Results are reported in terms of EER on three evaluation conditions, c_1-c_3.

4.4 DNN-based scoring

Based on the optimal number of discriminative features, i.e., $n = 10$, the DNN-based system is constructed. To investigate the impact of the deep structure, we build DNN systems with different number of hidden layers, and evaluate the performance

on the CV set and the evaluate set (the entire evaluation data) respectively. The EER results are presented in Figure 6, where L denotes the number of the hidden layers of the DNN. We can see that with more hidden layers the performance is indeed improved; however, L should not be larger than 4. This is reasonable since the training data is limited, and an over-complicated network tends to cause over-fitting.

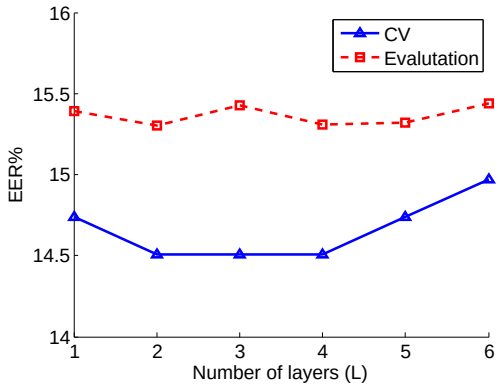

Figure 6 The results in EER with DNN-based scoring, with the number of hidden layers (L) varies from 1 to 6.

For comparison with the baseline systems, the EER result of the DNN system with L=2 is presented in Table 3, in the row denoted by 'DNN(n=2)'. It can be seen that the DNN-based scoring significantly improves the performance.

Table 3 The results in EER of baseline systems on the entire evaluation set.

System	EER%
i-vector	21.89
LDA	18.46
PLDA	17.22
DNN (L=2)	15.30
DNN(L=2) +PLDA: feature	15.18
DNN(L=2) +PLDA: score	15.24
SVM(linear)	17.92
SVM(polynomial)	17.59
SVM(RBF)	16.30

Further check the performance of the DNN-based scoring on different conditions. Figure 7 shows the results. It can be seen that the DNN-based scoring generally outperforms the i-vector and the LDA baselines, demonstrating that the DNN-based scoring is indeed more powerful than the simple and shallow discriminative function such as the cosine kernel and the LDA. When compare with the PLDA, we see that in condition 1 to 3, the DNN-based scoring obtain better performance, however in condition 4 to 7, the PLDA-based approach is clearly superior. This discrepancy on different conditions may be attributed to the training process of the DNN model. Specifically, since the discriminative feature selection (see the previous section) is based on the CV set which is sampled from all the conditions, the DNN model tends to optimize on the entire evaluation set instead of a particular condition.

Additionally, we notice that condition 1 to 3 are all microphone conditions and the contents are interview. This is highly different from the condition under which

the DNN is trained, where the training data are telephone speech, and the content is conversation. The superiority of the DNN-based approach on these three conditions seems to indicate that the DNN model is more powerful to learn discriminative patterns in complex conditions such as with acoustic mismatch. For matched conditions (between model training and evaluation), the PLDA model can deal with the discrimination really well so the DNN model can not beat. This is consistent to our argument that the DNN can learn complex decision boundaries in heterogeneous speaker space. Nevertheless, the relatively low performance of the DNN model in some conditions needs more investigation.

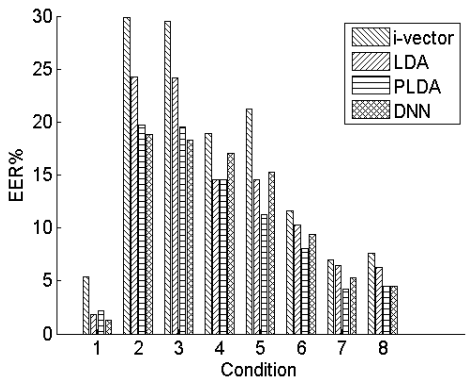

Figure 7 The results in EER with the DNN-based scoring on different conditions of the evaluation set.

4.5 Combination of DNN and PLDA

Considering the relative advantages of the DNN-based and PLDA-based scoring approaches, one can combine the two approaches to get better performance. Two combination methods are studied in this section, one is in the feature-domain and the other is in the score-domain.

4.5.1 Feature-domain combination

The first combination method is a feature-domain approach which augments the PLDA score to the input vector of the DNN model. Following the same training process, we obtain the performance on the evaluation set with different number of hidden layers L, as shown in Fig 8. For comparison, the results without the PLDA score are also shown. We see that more hidden layers offer better performance, given that L is not over large. And it is clear that involving the PLDA score consistently reduces the EER.

The fifth row in Table 3 (denoted by 'DNN(L=2):feature') presents the numerical result with L=2. The results on different conditions of the evaluation are shown in Figure 9. It can be observed that augmenting the PLDA score improves the DNN-based scoring in almost all the conditions. With the PLDA score involved, the DNN-based scoring equals to or outperforms the PLDA-based scoring in most of the conditions, except condition 4, 5 and 7. In the conditions that involve the most trials, i.e., condition 3 and 6, the DNN-based scoring outperforms the PLDA-based scoring.

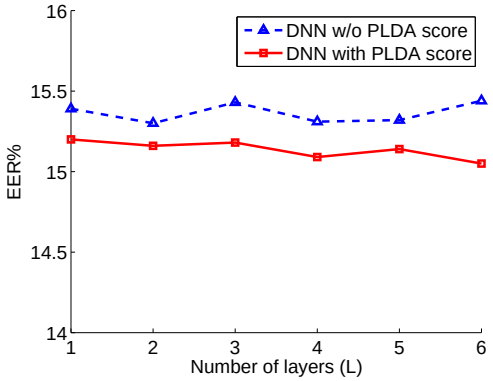

Figure 8 The results in EER with the DNN-based scoring on the evaluation set with/without the PLDA score.

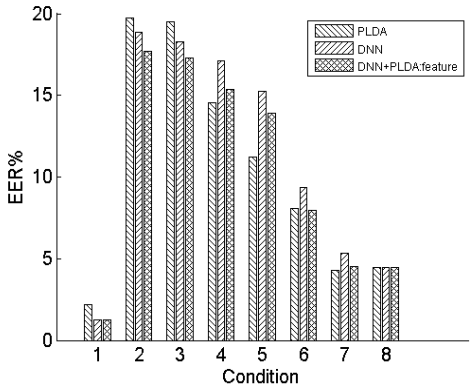

Figure 9 The results in EER with the DNN-based scoring on different conditions of the evaluation set, with the PLDA score augmented to the DNN input.

4.5.2 Score-domain combination

The second combination method uses linear interpolation to combine the DNN-based and PLDA-based approaches in the score-domain, according to (6). In order to choose the best interpolation factor α, we examine the EERs on the CV set by varying α from 0.0 to 1.0. The results show that 0.99 is a good trade-off. We choose this value to interpolate the DNN-based score and the PLDA score and test the performance on the evaluation set. For easy comparison, the DNN with 2 hidden layers is used in the experiment. The EER result on the entire evaluation set is presented in Table 3, at the row denoted by 'DNN(L=2)+LDA: score'. It can be seen that the number is a bit higher than the result obtained by the feature-domain combination, but it is lower than those obtained with the PLDA and the baseline DNN system.

The EER results on various conditions are shown in Figure 10. Interestingly, we observe that the score-domain combination generally outperforms the feature-domain combination, although its result on the entire evaluation set is slightly worse. With this combination, the DNN-based scoring outperforms the PLDA baseline in most conditions, except in condition 5 and 7 which composes only a small propor-

tion of the test trials. The inferior of the score-domain combination on the entire evaluation set and its superior on the multiple conditions again indicate that the DNN-based scoring over-fits to the 'entire-set optimization', less considering individual conditions.

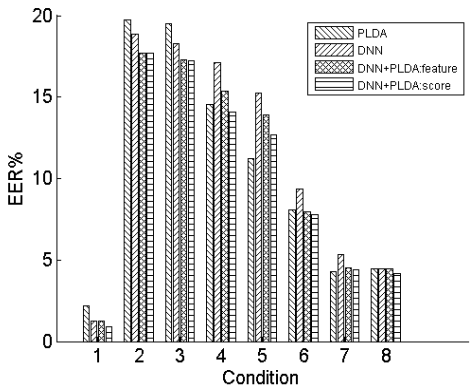

Figure 10 The results in EER on different conditions of the evaluation set, with the PLDA score interpolated with the DNN score.

4.6 Comparison between DNN and SVM

The last experiment compares the DNN, a deep model, and the SVM, a representative shallow model. As discussed in Section 3, the deep model such as DNN can learn flexible hierarchical features, while the shallow model such as the SVM relies on a kernel function that is predefined. It is then expected that the DNN can learn complex patterns in the heterogeneous speaker space.

We experimented the SVM model with three kernels: the linear kernel, the polynomial kernel, and the radio basis function (RBF) kernel. The feature is the same as the one used in the feature-domain DNN+PLDA combination (ref. (eq:feat-ext)). The EER results on the entire evaluation set are presented in Table 3. It can be seen that the SVM-based scoring also works and obtains better performance than the i-vector (using the cosine distance) and the LDA baseline. The RBF kernel is the most powerful and it actually outperforms the PLDA baseline. However, the SVM-based scoring can not beat the DNN-based scoring, even the baseline DNN without combination with the PLDA.

The EER results on different conditions of the evaluation set are presented in Figure 11. It can be observed that with different kernels, the SVM-based scoring behaves quite different. The most powerful RBF kernel, although obtains the best performance on the entire evaluation set, performs rather bad in some conditions. The polynomial kernel shares the same characteristics though it looks more stable than the RBF kernel. The linear kernel, although the most simple, is the most stable. With this kernel, the SVM-based scoring obtains better performance than the DNN-based scoring in conditions where the DNN-based scoring does not perform well, e.g., in condition 4 and condition 5. But in most conditions, the SVM-based scoring can not beat the DNN-based scoring, and it is generally worse than the PLDA baseline.

The above results demonstrate that the SVM is not so powerful to learn the complicated decision boundary of i-vector pairs, and the DNN is more suitable in such tasks. As in the DNN experiments, we again find a clear discrepancy between the result on the entire dataset and the results in individual conditions: it seems that a powerful model tends to gain a good performance on the entire evaluation data, but may not work in some conditions. This double confirms our conjecture that the criterion that we used in feature selection and model training, i.e., to optimize the global performance, may be biased. A condition-dependent discriminative training would be helpful if we want to optimize for a particular condition.

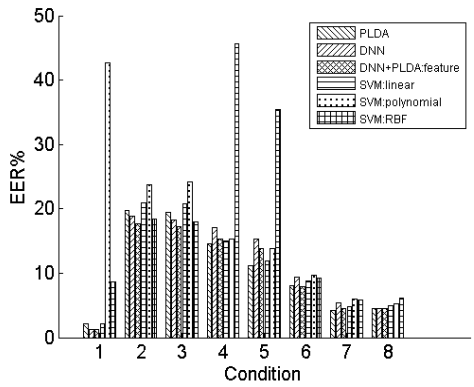

Figure 11 The results in EER on different conditions of the evaluation set, with the SVM-based scoring.

5 Conclusions

This paper present a DNN-based discriminative scoring approach to speaker recognition based on i-vector. We argue that by relaxing the Gaussian assumption of general models such as PLDA and optimizing the model with respect to the decision task directly, the DNN-based approach may achieve better performance than the PLDA in situations where the i-vector space is complicated. Furthermore, the DNN and PLDA approaches are complementary and so can be combined to acheive further gains. The experiments conducted on the SRE08 core test demonstrated that the DNN-based scoring outperforms the PLDA-based scoring on the entire evaluation data, and it also outperforms a popular shallow model, the SVM. The results on individual conditions are not very consistent and the DNN-based approach exhibits obvious advantage in conditions with acoustic mismatch between model training and evaluation. The combination of the DNN-based and the PLDA-based approaches, particularly the combination in the score-domain, has resulted in further performance improvement on the entire evaluation set and most of the individual conditions.

There is much work left. First of all, we found significant discrepancy with the DNN-based scoring with the entire evaluation and individual conditions, and discrepancy among different conditions. This discrepancy, as we argued, is probably attributed to the global optimization criterion we used in feature selection and model training. Nevertheless, this needs to be verified by testing condition-dependent DNN

models, for which the data sparsity is a problem right now. Second, the discriminative features used in this work are still rather simple. It is interesting to know better features that can offer more discriminative information, e.g., the covariance matrices accompanied to i-vectors. Third, better approaches to combining the PLDA or other generative models deserve careful study.

6 Acknowledgements

This work was supported by National Basic Research Program (973 Program) of China under Grand No. 2013CB329302 and the National Science Foundation of China (NSFC) under the project No. 61371136 and No. 61271389.

Author details
[1]Center for Speech and Language Technologies, Tsinghua University, ROOM 1-303, Information Sci & Tech Building, Tsinghua University, Beijing, 100084 China. [2]Center for Speech and Language Technologies, Tsinghua University, ROOM 4-416, Information Sci & Tech Building, Tsinghua University, 100084 Beijing, China.

References

1. Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 4, pp. 1435–1447, 2007.
2. Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.
3. Sergey Ioffe, "Probabilistic linear discriminant analysis," in *ECCV 2006*, 2006, pp. 531–542.
4. Andrew O Hatch, Sachin S Kajarekar, and Andreas Stolcke, "Within-class covariance normalization for SVM-based speaker recognition.," in *INTERSPEECH'06*, 2006.
5. A. Solomonoff, C. Quillen, and W. M. Campbell, "Channel compensation for SVM speaker recognition," in *Proc Odyssey, Speaker Language Recognition Workshop 2004*, 2004, pp. 57–62.
6. Craig S Greenberg, Vincent M Stanford, Alvin F Martin, Meghana Yadagiri, George R Doddington, John J Godfrey, and Jaime Hernandez-Cordero, "The 2012 NIST speaker recognition evaluation," in *INTERSPEECH'13*, 2013.
7. W. M. Campbell, D. E. Sturim, D. A. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," in *International Conference on Acoustics, Speech, and Signal Processing*, 2006, vol. 1.
8. Alex Solomonoff, W. M. Campbell, and Ian Boardman, "Advances In Channel Compensation For SVM Speaker Recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2005, vol. 1, pp. 629–632.
9. Andreas Stolcke, Sachin S. Kajarekar, Luciana Ferrer, and Elizabeth Shrinberg, "Speaker Recognition With Session Variability Normalization Based on MLLR Adaptation Transforms," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, pp. 1987–1998, 2007.
10. Sachin S. Kajarekarl and A. Stolche, "NAP and WCCN: Comparison of Approaches using MLLR-SVM Speaker Verification System," in *International Conference on Acoustics, Speech, and Signal Processing*, 2007, vol. 4.
11. Andrew O. Hatch and Andreas Stolcke, "Generalized Linear Kernels for One-Versus-All Classification: Application to Speaker Recognition," in *International Conference on Acoustics, Speech, and Signal Processing*, 2006, vol. 5.
12. Najim Dehak, Patrick Kenny, Réda Dehak, Ondrej Glembek, Pierre Dumouchel, Lukas Burget, Valiantsina Hubeika, and Fabio Castaldo, "Support vector machines and Joint Factor Analysis for speaker verification," in *International Conference on Acoustics, Speech, and Signal Processing*, 2009, pp. 4237–4240.
13. Xiangtao Meng, Chao Liu, Zhiyong Zhang, and Dong Wang, "Noisy training for deep neural networks," in *IEEE ChinaSIP*, 2014.
14. Jun Wang, Dong Wang, Ziwei Zhu, Thomas Fang Zheng, and Frank Soong, "Discriminative scoring for speaker recognition based on i-vectors," in *APASIPA ASC*, 2014.
15. Yun-Fan Chang Hsin-Min Wang Shyh-Kang Jeng Hung-Shin Lee, Yu Tso, "Speaker verification using kernel-based binary classifiers with binary operation derived features," in *International Conference on Acoustics, Speech, and Signal Processing*, 2014.
16. Patrick Kenny, Gilles Boulianne, and Pierre Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.
17. Mitchell McLaren and David Van Leeuwen, "Source-normalised-and-weighted LDA for robust speaker recognition using i-vectors," *IEEE Transactions on Audio, Speech, and Language Processing*, pp. 5456–5459, 2011.
18. Li Deng and Dong Yu, *DEEP LEARNING: Methods and Applications*, NOW Publishers, January 2014.
19. Hervé Bourlard and Nelson Morgan, "Hybrid HMM/ANN systems for speech recognition: Overview and new research directions," in *Adaptive Processing of Sequences and Data Structures*, pp. 389–417. Springer, 1998.
20. Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in *Proc. of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2000, pp. 1635–1638.
21. George E Dahl, Dong Yu, Li Deng, and Alex Acero, "Large vocabulary continuous speech recognition with context-dependent DBN-HMMs," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011, pp. 4688–4691.
22. Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.
23. Z. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines for statistical parametric speech synthesis," in *Proc. of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7825–7829.
24. Z. Ling, L. Deng, and D. Yu, "Modeling spectral envelopes using restricted boltzmann machines and deep belief networks for statistical parametric speech synthesis," *IEEE Transactions on Audio Speech Language Processing*, vol. 21, no. 10, pp. 2129–2139, 2013.
25. P. Hamel and D. Eck, "Learning features from music audio with deep belief networks," in *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*, 2010.
26. E. Batternberg and D. Wessel, "Analyzing drum patterns using conditional deep belief networks," in *Proceedings of International Symposium on Music Information Retrieval (ISMIR)*, 2012.

27. Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proc. of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.
28. Andrew L Maas, Quoc V Le, Tyler M O'Neil, Oriol Vinyals, Patrick Nguyen, and Andrew Y Ng, "Recurrent neural networks for noise reduction in robust ASR," in *Proc. of Interspeech*, 2012, pp. 22–25.
29. Xiao-Lei Zhang and Ji Wu, "Deep belief networks based voice activity detection," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 4, pp. 697–710, April 2013.
30. A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proceedings of Neural Information Processing Systems (NIPS)*, 2013.
31. P Kenny, V Gupta, T Stafylakis, P Ouellet, and J Alam, "Deep neural networks for extracting baum-welch statistics for speaker recognition," in *Speaker Odyssey'14*, 2014.
32. Yun Lei, Nicolas Scheffer, Luciana Ferrer, and Mitchell McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," in *ICASSP-2014*, 2014.
33. Simon JD Prince and James H Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *ICCV'07*. IEEE, 2007, pp. 1–8.
34. Najim Dehak, Reda Dehak, Patrick Kenny, Pierre Ouellet, and Pierre Dumouchel, "Support vector machines versus fast scoring in the low-dimensional total variability space for speaker verification," in *International Conference on Spoken Language Processing - ICSLP*. IEEE, 2009, pp. 1559–1562.
35. Tony Jebara, *Machine Learning: Discriminative and Generative*, Kluwer, 2004.
36. J. Lasserre, C. M. Bishop, and T. Minka, "Principled hybrids of generative and discriminative models," in *CVPR'06*, 2006.
37. NIST, "The NIST Year 2008 Speaker Recognition Evaluation Plan," 2008, http://www.itl.nist.gov/iad/mig/tests/sre/2008/sre08_evalplan_release4.pdf.