Available online at www.sciencedirect.com

ScienceDirect

Speech Communication 73 (2015) 1–13

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Deep feature for text-dependent speaker verification

Yuan Liu[a], Yanmin Qian[a,*], Nanxin Chen[a], Tianfan Fu[a], Ya Zhang[b], Kai Yu[a]

[a] Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering SpeechLab, Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

[b] Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China

Received 9 April 2015; received in revised form 27 June 2015; accepted 15 July 2015
Available online 21 July 2015

Abstract

Recently deep learning has been successfully used in speech recognition, however it has not been carefully explored and widely accepted for speaker verification. To incorporate deep learning into speaker verification, this paper proposes novel approaches of extracting and using features from deep learning models for text-dependent speaker verification. In contrast to the traditional short-term spectral feature, such as MFCC or PLP, in this paper, outputs from hidden layer of various deep models are employed as *deep features* for text-dependent speaker verification. Fours types of deep models are investigated: deep Restricted Boltzmann Machines, speech-discriminant Deep Neural Network (DNN), speaker-discriminant DNN, and multi-task joint-learned DNN. Once deep features are extracted, they may be used within either the GMM-UBM framework or the identity vector (i-vector) framework. Joint linear discriminant analysis and probabilistic linear discriminant analysis are proposed as effective back-end classifiers for identity vector based deep features. These approaches were evaluated on the RSR2015 data corpus. Experiments showed that deep feature based methods can obtain significant performance improvements compared to the traditional baselines, no matter if they are directly applied in the GMM-UBM system or utilized as identity vectors. The EER of the best system using the proposed identity vector is 0.10%, only one fifteenth of that in the GMM-UBM baseline.

© 2015 Elsevier B.V. All rights reserved.

Keywords: Text-dependent speaker verification; Deep neural networks; Deep features; RSR2015

1. Introduction

Speaker verification is the identification of the person who is speaking by characteristics of their voices (voice biometrics), and the relative technologies have reached maturity and been deployed in commercial applications in recent years. According to whether the text of test speech is the same as the one in the enrollment stage, there are two types of speaker verification tasks: *text-dependent* and *text-independent* speaker-verification. Since text-dependent speaker verification systems strictly constrain the speech phrase of a speaker and the knowledge of the lexicon is integrated in the modeling, the verification result is much more accurate compared to text-independent systems and the application is much safer. Besides, in many real scenarios, the duration of the user speech is usually short and text-independent verification is not robust. Accordingly the text-dependent speaker verification is more appropriate to be implemented in real applications to obtain an accurate verification result. This is also the research focus of this work.

In general, speaker verification system construction consists of three stages: *frond-end feature extraction*, *modeling*, and *back-end scoring or classification*. In the first step,

[*] Corresponding author.
E-mail addresses: liuyuanthelma@sjtu.edu.cn (Y. Liu), yanminqian@sjtu.edu.cn (Y. Qian), bobchennan@gmail.com (N. Chen), erduo@sjtu.edu.cn (T. Fu), ya_zhang@sjtu.edu.cn (Y. Zhang), kai.yu@sjtu.edu.cn (K. Yu).

http://dx.doi.org/10.1016/j.specom.2015.07.003
0167-6393/© 2015 Elsevier B.V. All rights reserved.

spectral based features are extracted, for example mel-frequency cepstral coefficients (MFCCs) or perceptual linear prediction (PLP) coefficients are widely used as the front-end cepstral features. Then various approaches are applied to build models,such as Gaussian Mixture Model (GMM) (Reynolds, 1995), Support Vector Machine (SVM) (Campbell et al., 2006) and so on. In particular, GMM-based methods, such as the classical Gaussian Mixture Model-Universal Background Model (GMM-UBM) (Reynolds et al., 2000) and the state-of-the-art i-vector (Dehak et al., 2011) approach, are most popular for speaker modeling.

Neural network has been applied to speaker recognition for a long time. In early years, it is used as a classifier or to strengthen other classifiers (Farrell et al., 1994; Wouhaybi and Adnan Al-Alaoui, 1999). Similar ideas have been extended in recent years. In (Turajlic and Bozanovic, 2012), one neural network with feature after Z-norm is trained for each speaker for verification. In (Ghosh et al., 2004), hierarchical neural network is used to improve performance. All these approaches require some forms of speaker-specific networks to be trained and are usually not easy to scale up to tasks with large number of speakers. Another category is to use the neural network to assist in the i-vector extraction (Senoussaoui et al., 2012; Burget et al., 2011; Vasilakakis et al., 2013; Thomas et al., 2012).

In recent years, deep learning, especially deep neural network (DNN), became a hot research topic in machine learning and achieved a breakthrough in speech recognition (Hinton et al., 2012), however it has not been widely accepted for speaker verification. Some systems were proposed (Variani et al., 2014), but did not achieve the best single system performance. Considering that the DNNs possess strong capability of nonlinear modeling representation (Le Roux and Bengio, 2008), it is believed that deep neural networks can be a better choice to extract discriminative features. In this paper, the use of deep models, including DNNs and deep RBM, is investigated in detail for text-dependent speaker verification. Deep features extraction using deep structures are proposed to improve the text-dependent speaker verification system.

The remainder of this paper is organized as follows. Section 2 reviews the developments of text-dependent speaker verification and the popular technologies used in this task. Section 3 and 4 describe the proposed deep feature extraction approaches and the back-end classifier construction. The detailed experimental results and comparisons are presented in Section 5 and the whole work is summarized in Section 6.

2. Text-dependent speaker verification

Throughout the history of speaker verification, from nonparametric template matching methods such as Dynamic Time Warping (DTW) (Yu et al., 1995) and Vector Quantization (VQ) (Burton, 1987) to parametric modeling methods such as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM) (Larcher et al., 2012b; Matsui et al., 1996), Artificial Neural Network (ANN), and most recently Deep Neural Network (DNN) (Variani et al., 2014); from the simple and clear data environments to more complicated noisy environments, the technologies of speaker verification show considerable advanced developments. Text-dependent speaker verification constrains the speech phrase in the enrollment stage the same as the phrase in the test stage, and it performs much better than text-independent speaker verification. The constrain of phrases makes the verification more accurate, because the decision can be made by only analyzing how a speaker produces the text-specific sounds rather than requiring to compare among different speakers over large lexical variations.

Text-dependent speaker verification needs to focus on both speaker characteristics and lexical contents. Early works usually utilized DTW, a dynamic programming method, which aligns two sequences of different lengths and performs the temporal template matching (Furui, 1981). But this frame level alignment is computationally expensive. Another commonly used model is the hidden Markov model which typically uses GMMs to generate sequences of acoustic vectors.

From the year 1996 on, the Speaker Recognition Evaluations (SRE) are held by National Institute of Standard and Technology (NIST) every one or two years, which leads to the fast development of speaker verification technologies, especially the text-independent systems. Fortunately most of these techniques proposed for text-independent speaker verification can also be applied to text-dependent ones, such as the classical GMM-UBM (Reynolds et al., 2000), the state of the art system i-vector (Dehak et al., 2011). Although it does not take into account the lexical information, GMM-UBM system still shows a promising performance in text-dependent speaker verification (Fu et al., 2014). As for the i-vector system, the direct application on the text-dependent condition is not as satisfactory as the text-independent ones (Larcher et al., 2012b). Accordingly researchers should try to find more advanced and suitable techniques for the text-dependent speaker verification application. In this paper, the deep features extracted from deep models are applied in the GMM-UBM or identity vector framework to get improved performances of the text-dependent speaker verification. The recently popularly used standard text-dependent-task-oriented RSR2015 database will be chosen to evaluate all the systems.

2.1. RSR2015 database

RSR2015 data corpus, released by the Human Language Technology (HLT) department at Institute for Infocomm Research (I2R) in Singapore, is designed for text-dependent speaker recognition with scenario based on fixed pass-phrases (Larcher et al., 2012b). It consists of three parts, each dedicated to a specific task involving different lexical and duration constraints (Larcher et al., 2014). In this paper Part I is used, which contains about 72 h of audio. It contains audio recording from 300 people, which include 143 female and 157 male speakers that are between 17 to 42 years old, and the whole set is divided into background (bkg), development (dev) and evaluation (eval) subsets. Among the 300 people, 50 male and 47 female speakers are in the background set, 50/47 in the development set and 57/49 in the evaluation set.

All audios are recorded using three portable devices, divided into nine sessions. Each session contains thirty short phrases. The average duration of these audios is 3.2 s. During testing, a speaker is enrolled with 3 utterances of the same phrase. The corresponding test utterances are also of the same phrase, however all utterances in a trial come from different sessions and are taken from the eval set.

The RSR2015 data corpus is well designed and has become a standard database for text-dependent speaker verification research, such as (Miguel et al., 2014; Scheffer and Lei, 2014; Kenny et al., 2014; Fu et al., 2014). The data configuration in this work is the same as in others (Miguel et al., 2014; Scheffer and Lei, 2014; Kenny et al., 2014; Fu et al., 2014).

2.2. GMM-UBM approach

In the category of the classical GMM-based speaker verification technologies, GMM-UBM plays an important role. The whole GMM-UBM framework can be shown in Fig. 1. It consists of three stages.

- UBM training:
  A speaker-independent background GMM model, is trained with data from large amounts of non-target speakers. It can represent the general speaker-independent distribution of speech acoustic features, and it is called Universal Background Model (UBM). The UBM parameters are trained with the iterative Expectation–Maximization (EM) algorithm and require unlabelled data which covers different people, different lexical contents and different channels.
- Enrollment stage: MAP training
  In this stage, the target speaker model is derived by adapting the parameters of UBM using the target speaker's enrollment speech and a form of Bayesian adaptation which is known as Maximum a Posteriori (MAP) adaptation. This adaptation would tune the parameters of GMM mixtures for the data which can be observed in the speaker's enrollment speech, and the parameters for those which is not seen in the speaker's enrollment speech are kept unchanged as the UBM. The outputs of this stage are a number of speaker-dependent models.
- Verification stageLikelihood ratio decision method is used in the verification stage. Given an observation sequence $\mathbf{O}$ which represents the feature extracted from a test utterance of a speaker s, there can be two hypotheses:

$$\begin{aligned} &H_0 : \mathbf{O} \text{ is from the target speaker } s \\ &H_1 : \mathbf{O} \text{ is not from the target speaker } s \end{aligned} \tag{1}$$

  Then the decision is made according to the likelihood ratio as below:

$$\Lambda = \frac{1}{T}\log\frac{p(\mathbf{O}|H_0)}{p(\mathbf{O}|H_1)} = \begin{cases} \geqslant \theta & \text{accept } H_0 \\ < \theta & \text{accept } H_1 \end{cases} \tag{2}$$

  where $P(\mathbf{O}|H_i), i = 0, 1$, is the probability of hypothesis H_i, which can be computed using the probability density function for $\mathbf{O}$ given the target speaker GMM model or the impostor GMM model. Usually UBM model acts as an impostor model in the test stage. T represents the number of frames in observation $\mathbf{O}$.

The whole procedure of GMM-UBM method is easy to implement, and it can usually obtain satisfactory performance in both text-dependent (Sturim et al., 2002) and text-independent speaker verification system (Liu et al., 2014).

2.3. GMM based i-vector approach

The identity vector (*i-vector*) is developed from the joint factor analysis (JFA) (Kenny et al., 2007a,b), which is a model representing speaker and session variability in GMM's. In JFA, a GMM is estimated for each target speaker, and the session variability is removed which helps to compensate for the inter-session variability and the channel mismatches between enrollment data and test data (Kenny et al., 2007a). In general, a speaker utterance is represented by a supervector (M) which derives from the

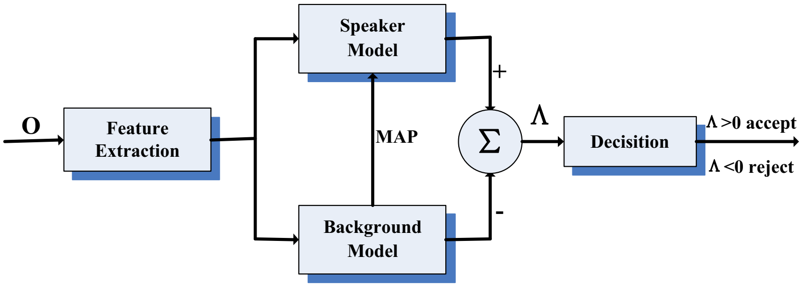

Fig. 1. The GMM-UBM speaker verification framework.

cascading of mean vectors of all mixture components in the speaker GMM. This speaker-dependent supervector can be decomposed as:

$$M = m + Vy + Ux + Dz \tag{3}$$

where m is a speaker and session-independent supervector, generated from UBM. Both matrix V and D define speaker subspace, and U defines a session subspace. y and x represent speaker factors and channel factors respectively. Dz serves as a residual to compensate for the speaker information that may not be caught by Vy.

The work in (Dehak, 2009) found that channel factors also contain speaker information. Thus a single subspace called total variability is proposed, which is known as the i-vector approach (Dehak et al., 2011). The new speaker and session-dependent GMM supervector is redefined as:

$$M = m + Tw \tag{4}$$

where T is a low rank matrix of speaker and session variability, and the total factor w is called identity vector, named i-vector. I-vectors are considered as front-end low dimension features, and normally the cosine similarity classifier is used to do fast scoring and decision:

$$\mathrm{Score}(w1, w2) = \frac{< w1, w2 >}{||w1||\ ||w2||} \tag{5}$$

where $< w1, w2 >$ is the inner product of two i-vectors and $||w_1||$ or $||w_2||$ the length of the respective i-vector.

Besides a more efficient back-end scoring method called PLDA (Probabilistic Linear Discriminant Analysis Model) is proposed (Jiang et al., 2012). It is similar to JFA but implemented in the i-vector space rather than supervector space. For a speaker whose i-vectors of all his speeches are defined as $D_1, D_2, \ldots D_N$, the i-vector D_n $(n = 1, \ldots, N)$ can be represented as

$$D_n = \mu + U_1x_1 + U_2x_{2n} + \epsilon_n \tag{6}$$

Corresponding with JFA, μ is the mean of the speaker's i-vector distribution and U_1 and U_2 define the speaker and session subspace individually.

Now most of the speaker verification systems utilize the i-vector approach, and this method has become the state of the art technology in this application (Matejka et al., 2011; Jiang et al., 2012). However, some work shows that this traditional i-vector framework may not work well in some scenarios, especially the text-dependent speaker verification applications (Larcher et al., 2012a). So more efforts need to be done in investigating this approach to improve the whole framework.

3. Deep features extraction using deep models

Both in speech recognition and in speaker recognition, feature extraction is very important for the system construction. Usually both of them utilize the short-time spectral features despite the completely different task objects (discriminating phones or discriminating speakers). Some disadvantages are obvious: (1) The features extracted in a short time cannot represent sound characteristics of a relatively long duration well, such as speaker identity; (2) The spectral features are originally designed for speech recognition, not speaker verification. Although these features could be used for speaker recognition, they are not optimized for the speaker discrimination, especially for the text-dependent speaker verification. Accordingly, it is important and meaningful to explore new features which are more discriminative and effective for the text-dependent speaker verification.

Neural networks especially deep neural networks have powerful nonlinear modeling abilities. In the early years of speaker recognition research, neural network has tried to be applied on speaker verification tasks. There are mainly two ways to utilize the neural network: model-based or feature-based. Most model-based approaches employ neural network as a classifier or to strengthen other classifiers (Farrell et al., 1994; Wouhaybi and Adnan Al-Alaoui, 1999). Model based approaches normally require speaker-specific network to be trained, which means for each test speaker, there will be a distinct neural network. Feature-based approaches employ neural network to extract compact and representative features for speaker verification. When using supervised nonlinear features, there is an issue of what labels to use as the target for training neural network. Early in the year 1998, Konig (Konig et al., 1998) tried to use bottleneck features to build GMM-UBM system. A neural network was trained with a bottleneck layer in the middle hidden layer. The input is expanded context-frame feature vectors and the output label is speaker id. Experiment results showed that the final combined system consisting of the spectral feature and the bottleneck feature systems outperformed the single feature system, however the individual bottleneck-feature-based system performed still worse than the baseline spectral feature based systems. Moreover the bottleneck approach was enhanced in (Yaman et al., 2012) and showed a slight gain. In (Chen and Salman, 2011), the unsupervised learning model autoencoder, which tries to make the output value equal to the input value and can learn quite good features from unlabelled data in the reconstruction process, could extract the discriminative features from the central hidden layer. However, all of the previous work obtained limited performance improvements or a little worse than baseline system.

Due to the recently large performance improvement in speech recognition by using the deep neural networks, applying deep neural networks to speaker verification has drawn special attentions (Fu et al., 2014). Traditional spectral features such as MFCCs or PLPs pass through deep models and arrive at a specific hidden layer and then projected new features are obtained. These new features, whether their dimensions are then reduced or kept, are called deep features. Focusing on deep feature extraction, this paper will implement several kinds of deep neural networks and RBM to extract effective features comprehensively, and then these novel features are applied in the GMM-UBM framework or directly act as an identity vector similar as the traditional i-vector framework.

3.1. Deep restricted Boltzmann machines

The training of the generative model RBM is an unsupervised process with the approximate contrastive divergence algorithm. No label information of data is needed, so very large amounts of training data can be utilized. Deep RBM can be adopted as a reasonable feature extractor since it models input features to more regular and generative features, shown as the Fig. 2. Considering that no label information is used in RBM, all speech characteristics may be represented in the RBM, including phone-level, speaker-level and channel-level characteristics. Once the RBM is trained, the original spectral features are fed through the neural network and the outputs of a particular hidden layer are extracted. The context information could also be encoded due to the extended inputs with the left and right *n* frames. These unsupervised trained deep features can be obtained from the different hidden layers of RBM, which hope to be beneficial for discriminating speakers.

3.2. Speech-discriminant deep neural network

Considering that the task is text-dependent, the text information should be useful in the modeling phase. Accordingly deep neural network which is trained for speech discrimination can be used as another feature extractor shown as the Fig. 3. Usually there are text labels (phone/state labels) for the text-dependent training data, e.g. borrowing from the speech recognition task corpus, so training a speech-discriminant DNN is feasible. In this speech-discriminant DNN, the layer close to the output layer is much more phrase-discriminative and less speaker-dependent, hence we need to make some trade off on the hidden layer selection considering both the phrase and speaker knowledge. To derive this feature extractor, a DNN is trained in supervised mode using the triphone states labels as targets. The RBM pretraining (Hinton et al., 2006) is used to initialize the neural network. Then deep features are extracted in a similar way as in the RBM extractor. Triphone states are closely related to text information and have been widely utilized for speech recognition. Considering the task is text-dependent speaker verification, we believe that using speech-discriminant DNN should be particularly useful.

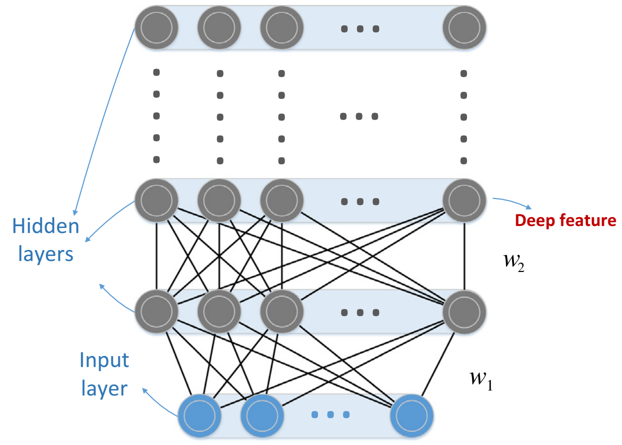

Fig. 2. Deep restricted Boltzmann machines.

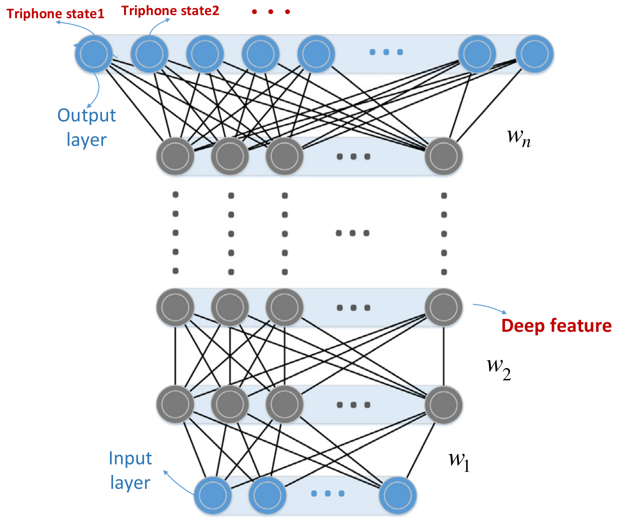

Fig. 3. Speech-discriminant deep neural network.

3.3. Speaker-discriminant deep neural network

This is a natural choice for speaker verification, as shown in the Fig. 4. The speaker discriminative ability will be enhanced in this type of DNN and other information such as phone variability and channel variability are constrained at a relative lower level. As mentioned above, works in (Konig et al., 1998; Chen and Salman, 2011) have

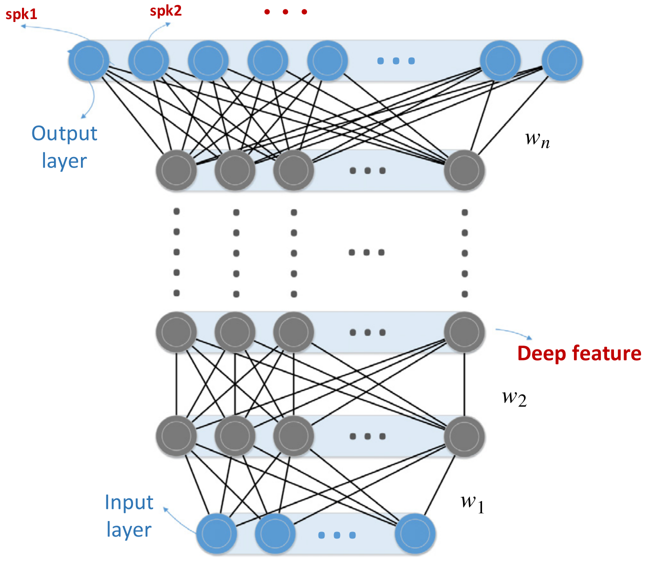

Fig. 4. Speaker-discriminant deep neural network.

used this speaker discriminative neural network feature extractor, however these features are both from the shallow NN model and they finally performed a little worse than the traditional spectral features. Speaker-discriminant DNN, which is also initialized by RBM, is more powerful in information reconstruction and must be more reliable in speaker information extraction. This DNN structure is similar as the speech-discriminant DNN except that the output classes represent individual speakers.

3.4. Multi-task joint-learned deep neural network

In the scenario of text-dependent speaker verification, both the discriminative knowledge from the speakers and the texts are very important. Thus, the combination of speaker-discriminant DNN and speech-discriminant DNN is the straightforward thinking. To learn the useful knowledge from different levels simultaneously, a multi-task joint-learned training process is applied. In other words, only one network is trained but the target is optimization on several levels at the same time, shown as the Fig. 5.

The output nodes consist of both speakers and texts. Here we consider two types of multi-task joint training: speaker + phrase, speaker + phone. In part I of RSR2015 database, there are 30 distinct phrases spoken by all the speakers. Hence in the first type of multi-task joint training, the number of text nodes can be 30. For training data, each speaker has 270 speeches and each text is spoken by all the speakers and through all channels. For simplicity we use the sum of the two original loss function $C_1(y_1, y'_1), C_2(y_2, y'_2)$ as the total loss function:

$$C([y_1, y_2], [y'_1, y'_2]) = C_1(y_1, y'_1) + C_2(y_2, y'_2) \qquad (7)$$

where C_1, C_2 are the two cross-entropy criteria for speakers and phrases. y_1, y_2 indicate the true labels for speakers and phrases individually, while y'_1, y'_2 are the outputs of the two targets respectively. According to the linearity of the gradient, the gradient of each parameter can be calculated individually, and the new parameters on common layers can be updated by the gradient for the sum of two loss functions. The learning rate is reduced when the classifying accuracy of the two tasks is not improving any more. Joint learning avoids over-fitting for DNN training, and also enhances the functionality of the DNN.

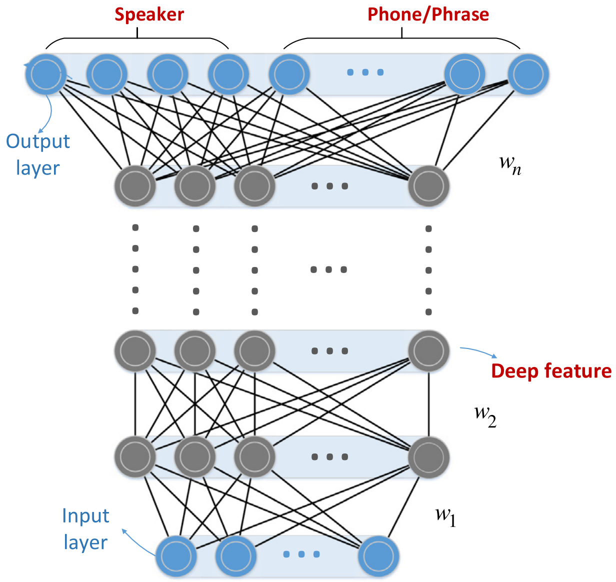

Fig. 5. Multi-task joint-learned deep neural network.

For the second type of multi-task joint training, named speaker + phone training, the text nodes represent the clustered context-dependent triphones states which are the basic modeling units in speech recognition (Lee, 1990). The training process is similar to the first type but notice that the number of clustered triphones states (usually over two thousand) are much more than the number of phrases.

Once the neural network training process is finished, the output layers of the two multi-task joint-learned DNNs can be removed, and the rest of each of the neural networks (common hidden layers) is used to extract the speaker-text joint representative features.

4. Back-end classifiers with deep features

These deep features derived from the above described deep models can then be used to replace the traditional spectral features. In this paper, they are firstly applied in GMM-UBM framework, and then these deep features can also serve as the identity vectors, similar as the i-vector framework, to get more sophisticated systems.

4.1. Deep features used in GMM-UBM framework

The deep features can be used directly in the GMM-UBM framework. Also they can be combined with the spectral features to form the tandem features in GMM-UBM.

4.1.1. Single deep features

The single deep features can be directly used in GMM-UBM system. After obtaining the high-dimension features output from the hidden layer in different deep models, the principal component analysis (PCA) is applied to orthogonalize the high-dimension features and only the most important components, which account for over 95% of the total variance, are retained. Usually the deep features are kept in the same dimension as the original spectral features. Also mean and variance normalization are applied to these new features. Finally, these deep features can be used to replace the original spectral features to train the normal GMM-UBM model.

4.1.2. Tandem deep features

Deep features described above are pure neural network based features. Demonstrated by many cases in speech recognition, combining neural network features with the

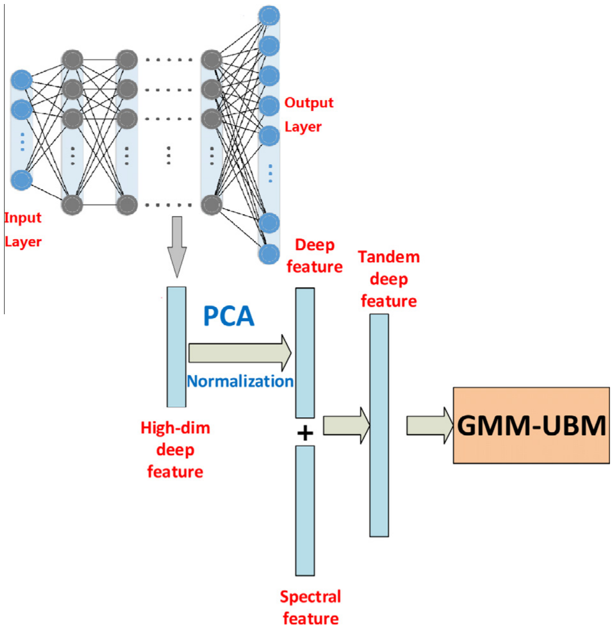

Fig. 6. Tandem deep features used in GMM-UBM.

original spectral features in a tandem fashion as the Fig. 6 shows can yield additional gains. The same idea is applied in GMM-UBM speaker verification. Described in Fig. 6, once the neural network is trained, raw spectral features are spanned in a context window (e.g. 11 frames, 5 frames on each side) and fed into a refined deep neural network to generate new features. The outputs from the specific hidden layers are utilized as the high-level features, optionally combined with the original spectral features, to build GMM-UBM speaker verification systems. The outputs from the hidden layer of the network always have high dimension due to the shape of neural network (indicated as the high-dim deep feature in the figure), so dimension is reduced to the same size as the normal spectral features using PCA algorithm. Then mean and variance normalization are used to normalize both spectral features and new deep features.

Although spectral features are not well discriminative, they may contain some useful information that has been omitted from deep features. In that way the tandem features are not only more discriminative but also comprehensive.

In addition, different types of deep features can also be combined to form new tandem deep features. This is done by simply concatenating different neural network features to obtain multi-deep features combined system.

4.2. Deep features used in identity vector framework

Different from the above described method which uses a deep model to extract deep features applied in GMM-UBM framework, here we regard the deep features extracted from the neural networks directly as the speaker identity representations, which is similar as the i-vector idea. In Google's recent work, a speaker classified DNN is trained to map frame-level features in a given context to the corresponding speaker identity target. During enrollment, the speaker identity vector is computed as the average of outputs derived from the last DNN hidden layer, which is defined as a deep vector or "d-vector" (Variani et al., 2014). In the evaluation phase, decisions are made according to the distance between the target d-vector and the test d-vector, which is similar as in the i-vector speaker verification systems.

Inspired by this, all the types of proposed deep features described in the Section 3 can serve to form the identity vectors. Considering that the deep features are frame-level features, to obtain the identity vector of a speaker (speaker here represents joint class – speaker and phrase because of text-dependent condition), all of these deep features belonging to the same class are averaged to form the target vector. In this work, besides extracting the identity vectors from the last hidden layer as the previous d-vector work (Variani et al., 2014), we also explored the identity vectors extraction from different hidden layers to investigate better performance. For the convenient representation, the identity vectors extracted using the above four (actually five) deep models (deep restricted boltzmann machines, speech-discriminant deep neural network, speaker-discriminant deep neural network, multi-task joint learned (speaker + phrase) deep neural network, multi-task joint learned (speaker + phone) deep neural network) are named r-vector, p-vector, d-vector (the same as the name in Google's recently work (Variani et al., 2014)), j-vector (spkr-phr-vector, spkr-pho-vector) respectively.

These identity vectors can be used in several different back-end classifications, such as cosine similarity, linear discriminant analysis (LDA) (McLaren and Van Leeuwen, 2012, 2011), and probabilistic linear discriminant analysis (PLDA) (Matejka et al., 2011; Kenny et al., 2013), all of which are usually used in the classic i-vector framework. Previous work on d-vector (Variani et al., 2014) only used cosine similarity for verification, however we investigate other classifiers in detail on these different types of NN based identity vectors.

4.2.1. Joint linear discriminant analysis

Linear discriminant analysis (LDA) provides good generalization capability even with limited number of training samples. The motivation for using this model is that LDA attempts to define new special axes that minimize the intra-class variance caused by channel effects, and to maximize the variance between classes. Due to these reasons it was used on many tasks related to speaker verification and speaker identification (Matejka et al., 2011; Jin and Waibel, 2000). It assumes that each class density can be modelled as a multivariate gaussian:

$$\mathcal{N}(x|\mu_k,\Sigma_k) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{\frac{1}{2}}}\exp^{-\frac{1}{2}(x-\mu_k)\Sigma_k^{-1}(x-\mu_k)} \quad (8)$$

where Σ_k and μ_k is the covariance and mean for class k. LDA(Scholkopft and Mullert, 1999) model assumes that $\Sigma_k = \Sigma, \forall k$. And the posterior probability is given as:

$$P(Class = k|X = x) = \frac{\mathcal{N}(x|\mu_k, \Sigma_k)\pi_k}{\sum_{l=1}^{K}\mathcal{N}(x|\mu_l, \Sigma_l)\pi_l} \tag{9}$$

where K denotes the total number of classes and π_k is the prior of class k (uniform distribution is used as prior in this work).

It is obvious to define the class in other types of identity vectors except the multi-task joint learned vector. It is noted that we need to define the LDA class as the joint class considering both speaker and phrase information in j-vector, similar to the joint learned process in deep neural networks.

It is worthy to note that the posterior calculation limits the usage of LDA algorithm. It assumed that the test segments are given by one of the enroll speaker, which condition is defined as closed-set evaluation.

4.2.2. Probabilistic linear discriminant analysis

LDA uses gaussian mixture model, which can be regarded as a latent variable model, where the observed node x represents the example and the latent variable μ_k is the center of a mixture component representing class k. The class-conditional distribution is $P(x|y) = \mathcal{N}(x|\mu_k, \phi)$ where ϕ is shared by all classes. PLDA (Matejka et al., 2011; Kenny et al., 2013) is proposed to make the latent variable prior continuous. Particularly, to enable efficient inference and closed-form training, a Gaussian prior is imposed: $P(\mu_k) = \mathcal{N}(\mu_k|m, \phi_b)$.

One advantage of PLDA is that it is not constraint to the closed-set in testing, so it also can deal with those "unseen speakers", who are not in the enrolled speakers.[1]

5. Experiments and results

To fully explore the effectiveness of the proposed deep features for the text-dependent speaker verification, experiments and comparisons about these four types of NN based features are designed, and the evaluations are implemented in both the GMM-UBM framework and identity vector framework.

5.1. Experimental setup and baseline systems

In all the experiments of this paper, the bkg and dev data of RSR2015 part I are merged as new bkg data (194 speakers, 100 male/94 female). In the test data set there are 19,052 tests for true speaker and 1,548,956 tests for imposture.

Table 1
Performance EER (%) of basline systems.

System	EER(%)
GMM-UBM	1.50
i-vector	5.02

Two normal systems are constructed as the baseline systems in this work: one is the spectral feature based GMM-UBM system and the other is the traditional i-vector system. 39-dimensional PLP features with mean and variance normalization are used as the spectral features in the baselines. An energy-based Voice Activity Detection (VAD) is utilized to detect the speech segments, and a gender-independent UBM of 1024 components is trained using the new bkg data for the GMM-UBM baseline. In the traditional classic i-vector system construction, parts of NIST SRE 2005 and NIST SRE 2008 data are used as the development data to train the T matrix and cosine similarity is directly used after LDA in back-end processing. The Equal Error Rate (EER) of the baseline systems are illustrated in Table 1. We can see that the EER is relatively low compared to the usual text-independent tasks, and it is relatively hard to improve on this good point baseline. The traditional i-vector system is not as good as the GMM-UBM system in this text-dependent scenario, which is consistent to the conclusions in others' work (Larcher et al., 2014).

5.2. Neural network training configuration

To evaluate the proposed four types of deep feature extractors, different neural networks are trained firstly, including Deep RBM, speech-discriminant DNN, speaker-discriminant DNN and multi-task joint-learned DNN. All the deep models have 7 hidden layers with 1024 nodes per layer, and a context window of 11 frames 39-dim PLP is concatenated to be used as the NN input. The new bkg data is used in the NN training.

The state alignment for the speech-DNN training is performed using a GMM-HMM model with 3001 tied-triphone-states, which is built on a 50-h SWB English task (refer to our previous work in (Fu et al., 2014). Totally 194 classes (194 speakers in the new bkg set) are used in the speaker DNN training. In the first type of multi-task joint learned neural network, speaker + phrase, 224 classes (194 speakers and 30 phrases) are used in the joint training. In the second type, there are 3195 classes (194 speakers and 3001 triphone states).

The contrastive divergence algorithm is used in the Deep RBM training, and SGD based back-propagation is applied to train the other DNNs. The learning rate annealing and early stopping strategies as in (Dahl et al., 2012) are used in the BP process and the DNNs are fine-tuned with cross-entropy objective function, along with an L2-norm weight-decay term of coefficient 10^{-6}.

[1] In this paper, all the experiments are the closed-set evaluation, and the testing utterances are all from one of the enrolled speakers, i.e. non unseen speakers exist in the experiments.

5.3. Evaluation of the deep features in GMM-UBM framework

When finishing the model training, the deep models are utilized to transform the original spectral features into the new deep features. For each speech frame, the principal component analysis (PCA) is applied on the outputs of hidden layers and reduces the dimension to 39 as the original PLP features. After mean and variance normalization, these new deep features can be used for the following modeling, or be connected with the original PLP to form the new concatenated tandem deep features. The later GMM-UBM construction is built as usual.

5.3.1. Evaluation of individual deep features

The proposed four types of deep features are firstly investigated individually. For detailed comparison, the experiments of deep features extracted from different hidden layers of the Deep RBM or DNNs are performed. Besides the experiments using the single deep features and the concatenated tandem deep features are also implemented. A system performances are shown in Table 2.

From Table 2, it is observed that most of the neural network based deep features get much better performance than the PLP baseline. Regarding the RBM-based deep features, the feature from the middle layer obtains the best EER, and the relatively lower layer (the 2nd layer here) achieves the best position when using the speech, speaker or multi-task based DNNs. Although the single deep features already can obtain the obvious improvement, the concatenated features with PLP get a much larger EER reduction in all types of neural networks. Moreover the supervised DNNs could use more information for model training, and they are all superior to the unsupervised RBM in tandem deep feature based GMM-UBM. The speech-discriminant DNN retains much more information about the text which is especially useful in this text-dependent task, while more speaker-dependent knowledge can be enhanced in the speaker-discriminant DNN, which makes the features more speaker discriminable. Considering the multi-task joint-learned DNNs, no matter speaker + phrase DNN or speaker + phone DNN, they are the best choices in deriving deep features. Moreover there are relatively big differences among the performance of distinct layers in speech/speaker discriminative DNNs, however the differences are very small when in the multi-task joint-learned ones, which illustrates the robustness and effectiveness of the multi-task joint-learned approach again.

In the RSR2015 evaluation, the speech texts are relatively fixed and there are only small number of phrases (30 distinct phrases). Accordingly speaker + phrase DNN is a good choice to be used as a deep feature extractor. In other situations where the texts are more flexible, speaker + phone DNN can be more suitable. Besides the unsupervised RBM method is a good alternative to make use of large quantity of unlabelled data. To summarize, neural network based deep features show much better performance in text-dependent speaker verification.

5.3.2. Evaluation of different deep features combination

Different types of deep features can also be combined to form new tandem discriminative deep features. In this section, we select the relative best system for individual deep feature respectively, such as the 4th-layer RBM deep feature, the 2nd-layer speaker discriminant DNN deep feature, and the 2nd-layer speaker + phrase multi-task learned DNN. The different deep features are tried to be concatenated to form new tandem deep features (PLP is always connected). After this, the GMM-UBM modeling is performed as before. The results of different deep feature combinations are shown in Table 3. These systems not only comprise complementarity of different target based DNNs, e.g. phone v.s. speaker, but also combine different criteria training strategies, including the unsupervised strategy and supervised strategy.

It can be observed that compared to the individual deep feature systems in Table 2, the multi-deep feature combinations obtain additive improvements.The best tandem deep feature approach obtains another 10% relative EER reduction when compared to the best individual system. Compared to the baseline, the multi-deep features show obvious advantages in discrimination. The DET curves in Fig. 7 show a performance comparison of some of the proposed features in GMM-UBM framework.

5.4. Evaluation of the deep features in identity vector framework

As described in Section 4.2, the individual identity vectors are firstly extracted by the average of the last hidden layer's outputs in each type of deep models, as named

Table 2
Performance EER (%) of individual discriminative deep features. The bold fonts denote the best performance in individual deep feature type.

Layer index	RBM		Speech-DNN		Speaker-DNN		Speaker + phrase DNN		Speaker + phone DNN	
	Deep fea	+ PLP	Deep fea	+ PLP	Deep fea	+ PLP	Deep fea	+ PLP	Deep fea	+ PLP
Base GMM-UBM	**1.50**									
2nd-layer	1.25	0.99	**1.45**	**0.89**	**1.08**	**0.80**	**1.06**	**0.80**	1.06	**0.85**
4th-layer	**1.23**	**0.94**	1.86	1.04	1.48	0.97	1.07	0.84	**0.95**	0.93
7th-layer	1.46	1.06	1.94	1.15	2.15	0.96	1.13	0.92	1.19	0.93

Table 3
Performance of different deep features combination.

PLP	RBM	Speech-DNN	Speaker-DNN	Speaker + phrase DNN	Speaker + phone DNN	EER (%)
√	–	–	–	–	–	1.50
√	√	√	–	–	–	0.80
	√	–	√	–	–	0.73
	–	√	√	–	–	0.82
	√	√	√	–	–	0.74
	√	–	–	√	–	0.74
	√	–	–	–	√	0.80
	–	–	–	√	√	0.76
	√	–	–	√	√	0.73

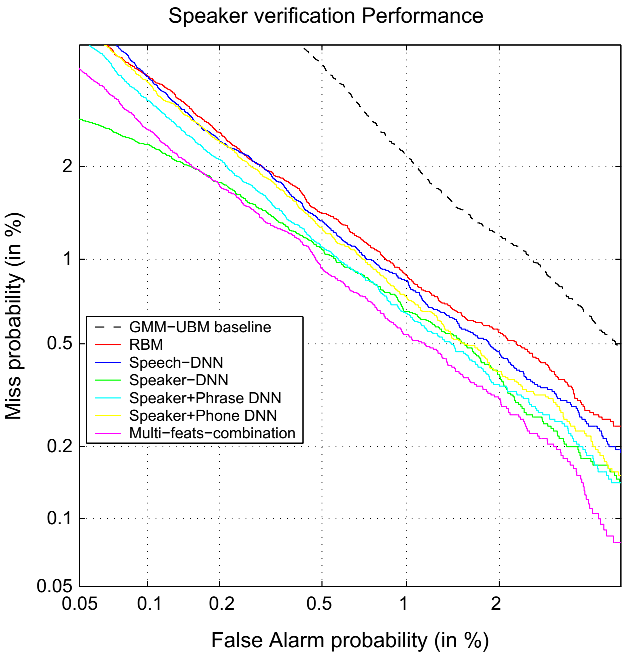

Fig. 7. The DET comparison of different deep features in GMM-UBM framework.

r-vector, p-vector, d-vector, j-vector (spkr-phr-vector and spkr-pho-vector) respectively.

The LDA model is then trained using these identity vectors respectively. The class defined in the LDA method is the joint label of speaker and phrase. For each test audio we extract identity vector using the same steps and then we use the decision function from LDA algorithm to distinguish among different models. Similarly the PLDA based approach is also applied, here we set the within class covariance smoothing parameter[2] to be 0.1 and then estimate the PLDA model with 25 iterations. Besides, the two methods can also be combined, which means that firstly PLDA is applied and then LDA is used to make the final decision. For easy comparison, the cosine similarity based decision function is also implemented, which is used in Google's d-vector work (Variani et al., 2014). The performance comparison is summarized in Table 4.

From Table 4, it shows that the classifier is important for these NN-based identity vectors. As for the proposed LDA, PLDA or the LDA + PLDA approach, there is a very large performance decline in all the networks when compared to the cosine similarity based decision. Different from the traditional i-vector method (Dehak et al., 2011), the simple cosine similarity based classifier is not appropriate for these neural network based identity vectors. The LDA method gets the best position in almost every types of NN-based identity vector. It may be benefit from the strict closed-set condition in this evaluation. Among these identity vectors, d-vector derived from speaker-discriminant neural network is relatively the worst choice, and the multi-task joint-learned neural networks are much better and more robust than the others.

Similar as the experiments on the deep features in GMM-UBM framework in Section 5.3, the experiments on NN-based identity vector from different hidden layers are also investigated on the multi-task joint-learned DNNs, which is better than the others according to Table 4. The identity vectors extracted from different layers using LDA and PLDA are shown in Table 5. It shows that there is another significant improvement when moving the layer from close to the output, to close to the input, and this is also consistent with the conclusion in the previous GMM-UBM framework. The best systems using the joint-learned identity vectors from the second hidden layer achieve 0.1% EER on both two type j-vectors.

5.5. The final system comparison for the proposed methods

The proposed novel approaches are summarized and illustrated in this section, including the best deep feature based GMM-UBM system and the deep feature based identity vector system. Fig. 8 shows a performance comparison of the proposed systems. Compared to the traditional baselines, the novel deep feature based systems show substantial improvement, and they are both using discriminative deep features. Particularly the EER of the system using j-vector is reduced 15 times when compared to the GMM-UBM baseline (i-vector baseline is much worse).

[2] In order to get a good estimate of the within-class covariance, the production of this parameter and the between-class covariance is adding to the within-class covariance.

Table 4
Performance EER (%) of different NN based identity vector systems. The bold fonts denote the best performance in individual identity vector type.

DNN	Classifier	EER	minDCF
r-vector	Cosine sim.	17.61	0.7817
	LDA	**0.33**	**0.0151**
	PLDA	1.06	0.0476
	PLDA + LDA	0.22	0.0105
p-vector	Cosine sim.	4.67	0.2172
	LDA	**0.27**	**0.0131**
	PLDA + LDA	0.30	0.0155
d-vector	Cosine sim.	7.55	0.3473
	LDA	**1.12**	**0.0459**
	PLDA	2.01	0.0984
	PLDA + LDA	1.15	0.0448
spkr-phr-vector	Cosine sim.	4.41	0.2163
	LDA	**0.15**	**0.0080**
	PLDA	1.25	0.0491
	PLDA + LDA	0.20	0.0117
spkr-pho-vector	Cosine sim.	4.41	0.2163
	LDA	**0.19**	**0.0081**
	PLDA	1.18	0.0450
	PLDA + LDA	0.22	0.0105

Table 5
Performance EER (%) of NN-based identity vectors from different hidden layers. The bold fonts denote the best performance.

Layer index	spkr-phr-vector		spkr-pho-vector	
	LDA	PLDA	LDA	PLDA
2nd-layer	**0.10**	0.90	**0.10**	0.90
4th-layer	0.11	0.94	0.13	1.04
7th-layer	0.15	1.25	0.19	1.18

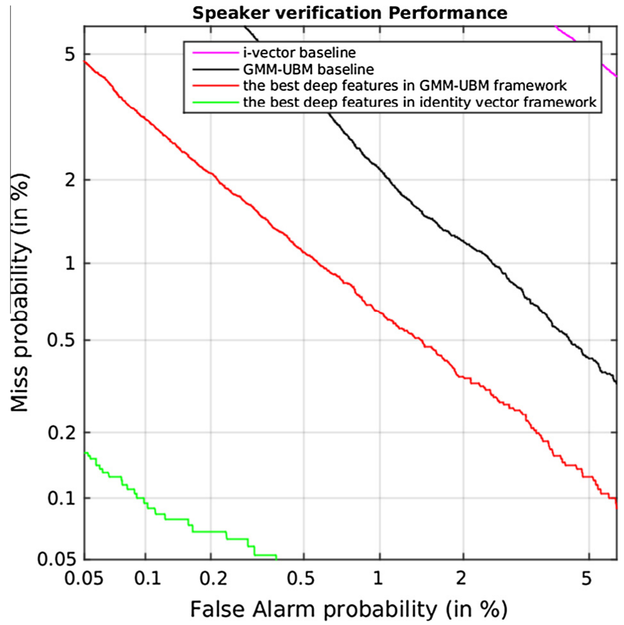

Fig. 8. The DET comparison of the best results of the proposed methods and the baseline results.

6. Conclusion and future work

This paper presents the detailed work on using various types of deep features for text-dependent speaker verification. Four types of deep feature engineering are proposed, including deep restricted boltzmann machines, speech-discriminant deep neural networks, speaker-discriminant deep neural networks, and multi-task joint-learned deep neural networks. These deep features are all implemented in both the GMM-UBM and identity vector framework, and they are much more effective and robust than the traditional methods in both frameworks. Among the proposed four types of deep features, the features from the multi-task joint-learned DNN outperform the others and show superior advantages. The best system using j-vector with LDA classifier achieves the EER 0.1%, which clearly outperforms the baseline.

Based on these experimental results, we could see that using deep models is very promising in speaker verification. In the future, other ways to extract deep features or other deep structures will be developed.

Acknowledgments

This work was supported by the Program for Professor of Special Appointment (Eastern Scholar) at Shanghai Institutions of Higher Learning, the China NSFC Project No. 61222208, and JiangSu NSF Project No. 201302060012.

References

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

Burton, D., 1987. Text-dependent speaker verification using vector quantization source coding. IEEE Trans. Acoust. Speech Signal Process. 35 (2), 133–143.

Campbell, W.M., Sturim, D.E., Reynolds, D.A., 2006. Support vector machines using gmm supervectors for speaker verification. IEEE Signal Process. Lett. 13 (5), 308–311.

Chen, K., Salman, A., 2011. Learning speaker-specific characteristics with a deep neural architecture. IEEE Trans. Neural Netw. 22 (11), 1744–1756.

Dahl, G.E., Yu, D., Deng, L., Acero, A., 2012. Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. IEEE Trans. Audio Speech Language Process. 20 (1), 30–42.

Dehak, N., 2009. Discriminative and generative approaches for long- and short-term speaker characteristics modeling: application to speaker verification. Ecole de Technologie Superieure (Canada).

Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., Ouellet, P., 2011. Front-end factor analysis for speaker verification. IEEE Trans. Audio Speech Language Process. 19 (4), 788–798.

Turajlic, E., Bozanovic, O., 2012. Neural network based speaker verification for security system. 20th Telecommunications Forum (TELFOR), 2012. IEEE, pp. 740–743.

Farrell, K.R., Mammone, R.J., Assaleh, K.T., 1994. Speaker recognition using neural networks and conventional classifiers. IEEE Trans. Speech Audio Process. 2 (1), 194–205.

Fu, T., Qian, Y., Liu, Y., Yu, K., 2014. Tandem deep features for text-dependent speaker verification. In: Fifteenth Annual Conference of the International Speech Communication Association.

Furui, S., 1981. Cepstral analysis technique for automatic speaker verification. IEEE Trans. Acoust. Speech Signal Process. 29 (2), 254–272.

Ghosh, J., Love, B.J., Vining, J., Sun, X., 2004. Automatic speaker recognition using neural networks.

Hinton, G., Deng, L., Yu, D., Dahl, G.E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T.N., et al., 2012. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Process. Mag. 29 (6), 82–97.

Hinton, G., Osindero, S., Teh, Y.-W., 2006. A fast learning algorithm for deep belief nets. Neural Comput. 18 (7), 1527–1554.

Jiang, Y., Lee, K.-A., Tang, Z., Ma, B., Larcher, A., Li, H., 2012. Plda modeling in i-vector and supervector space for speaker verification. In: INTERSPEECH.

Jin, Q., Waibel, A., 2000. Application of lda to speaker recognition. In: INTERSPEECH, pp. 250–253.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007a. Joint factor analysis versus eigenchannels in speaker recognition. IEEE Trans. Audio Speech Language Process. 15 (4), 1435–1447.

Kenny, P., Boulianne, G., Ouellet, P., Dumouchel, P., 2007b. Speaker and session variability in gmm-based speaker verification. IEEE Trans. Audio Speech Language Process. 15 (4), 1448–1460.

Kenny, P., Stafylakis, T., Alam, J., Ouellet, P., Kockmann, M., 2014. In-domain versus out-of-domain training for text-dependent jfa. In: Fifteenth Annual Conference of the International Speech Communication Association.

Kenny, P., Stafylakis, T., Ouellet, P., Alam, M., Dumouchel, P., et al., 2013. Plda for speaker verification with utterances of arbitrary duration. 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 7649–7653.

Konig, Y., Heck, L., Weintraub, M., Sonmez, K., et al., 1998. Nonlinear discriminant feature extraction for robust textindependent speaker recognition. In: Proc. RLA2C, ESCA Workshop on Speaker Recognition and its Commercial and Forensic Applications, pp. 72–75.

Larcher, A., Bousquet, P., Lee, K.A., Matrouf, D., Li, H., Bonastre, J.-F., 2012a. I-vectors in the context of phonetically-constrained short utterances for speaker verification. 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4773–4776.

Larcher, A., Lee, K.-A., Ma, B., Li, H., 2012b. Rsr2015: database for text-dependent speaker verification using multiple pass-phrases. In: INTERSPEECH.

Larcher, A., Lee, K.A., Ma, B., Li, H., 2014. Text-dependent speaker verification: classifiers, databases and rsr2015. Speech Commun. 60, 56–77.

Le Roux, N., Bengio, Y., 2008. Representational power of restricted boltzmann machines and deep belief networks. Neural Comput. 20 (6), 1631–1649.

Lee, K.-F., 1990. Context-dependent phonetic hidden markov models for speaker-independent continuous speech recognition. IEEE Trans. Acoust. Speech Signal Process. 38 (4), 599–609.

Liu, Y., Fu, T., Fan, Y., Qian, Y., Yu, K., 2014. Speaker verification with deep features. 2014 International Joint Conference on Neural Networks (IJCNN). IEEE, pp. 747–753.

Matejka, P., Glembek, O., Castaldo, F., Alam, M.J., Plchot, O., Kenny, P., Burget, L., Cernocky, J., 2011. Full-covariance ubm and heavy-tailed plda in i-vector speaker verification. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4828–4831.

Matsui, T., Kanno, T., Furui, S., 1996. Speaker recognition using hmm composition in noisy environments. Comput. Speech Language 10 (2), 107–116.

McLaren, M., Van Leeuwen, D., 2011. Source-normalised and weighted lda for robust speaker recognition using i-vectors. 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5456–5459.

McLaren, M., Van Leeuwen, D., 2012. Source-normalized lda for robust speaker recognition using i-vectors from multiple speech sources. IEEE Trans. Audio Speech Language Process. 20 (3), 755–766.

Miguel, A., Villalba, J., Ortega, A., Lleida, E., Vaquero, C., Agnitio, S., 2014. Factor analysis with sampling methods for text dependent speaker recognition. In: Fifteenth Annual Conference of the International Speech Communication Association.

Reynolds, D.A., 1995. Speaker identification and verification using gaussian mixture speaker models. Speech Commun. 17 (1), 91–108.

Reynolds, D.A., Quatieri, T.F., Dunn, R.B., 2000. Speaker verification using adapted gaussian mixture models. Digital Signal Process. 10 (1), 19–41.

Scheffer, N., Lei, Y., 2014. Content matching for short duration speaker recognition.

Scholkopft, B., Mullert, K.-R., 1999. Fisher discriminant analysis with kernels. Neural networks for signal processing IX.

Senoussaoui, M., Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., 2012. First attempt of boltzmann machines for speaker verification. In: Odyssey 2012-The Speaker and Language Recognition Workshop.

Sturim, D.E., Reynolds, D.A., Dunn, R.B., Quatieri, T.F., 2002. Speaker verification using text-constrained gaussian mixture models. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1. IEEE, pp. I–677.

Thomas, S., Mallidi, S.H., Ganapathy, S., Hermansky, H., 2012. Adaptation transforms of auto-associative neural networks as features for speaker verification. In: Proceedings of Odyssey, pp. 98–104.

Variani, E., Lei, X., McDermott, E., Moreno, I.L., Gonzalez-Dominguez, J., 2014. Deep neural networks for small footprint text-dependent speaker verification. 2014 IEEE International Conference on

Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4052–4056.

Vasilakakis, V., Cumani, S., Laface, P., 2013. Speaker recognition by means of deep belief networks. In: Proc. Biometric Technologies in Forensic Science.

Wouhaybi, R.H., Adnan Al-Alaoui, M., 1999. Comparison of neural networks for speaker recognition. In: Proceedings of ICECS'99. The 6th IEEE International Conference on Electronics, Circuits and Systems, 1999, vol. 1. IEEE, pp. 125–128.

Yaman, S., Pelecanos, J., Sarikaya, R., 2012. Bottleneck features for speaker recognition. In: Odyssey, vol. 12, pp. 105–108.

Yu, K., Mason, J., Oglesby, J., 1995. Speaker recognition using hidden markov models, dynamic time warping and vector quantisation. IEE Proc. – Vis. Image Signal Process. 142 (5), 313–318.