2014 International Joint Conference on Neural Networks (IJCNN)
July 6-11, 2014, Beijing, China

Speaker Verification with Deep Features

Yuan Liu Tianfan Fu Yuchen Fan Yanmin Qian Kai Yu
Institute of Intelligent Human-Machine Interaction
MOE-Microsoft Key Lab. for Intelligent Computing and Intelligent Systems
Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China
Email: {liuyuanthelma, erduo, fyc0624, yanminqian, kai.yu}@sjtu.edu.cn

***Abstract*—Due to great success of deep learning in speech recognition, there has been interest of applying deep learning to speaker verification. Previous investigations usually focus on using deep neural network as new classifiers or to extract speaker dependent features. They are either not compatible with existing speaker verification approaches, or not able to achieve significant performance gain in large scale tasks. Also, all the previous approaches have not addressed the issue of how to make use of extra unsupervised data. This paper proposes a novel feature engineering approach within the deep learning framework for speaker verification. Hidden layer output of deep neural network or deep belief network trained on large amount of speech recognition data are extracted as *deep features*. These features are then used in a Tandem fashion or concatenated with the original acoustic features for GMM-UBM speaker verification. The proposed approach can make use of large amount of existing speech recognition data without speaker labels and is easy to be combined with other mature classification approaches. Experiments on the core condition of NIST 2006 SRE showed that, in a text independent task, the proposed approach can achieve 12.8% relative EER improvement compared to the standard GMM-UBM systems. In addition, text-dependent speaker verification experiments were also performed and yielded similar significant gain.**

I. Introduction

Speaker recognition is a form of biometric personal recognition. Since everybody has his or her unique voice, in speaker recognition discrete feature vectors from people's voices through several steps of signal processing are extracted and are used to recognize speaker ID via subsequent modelings. Usually there are two modes of recognition: verification and identification. Speaker identification aims at identifying who is the speaker while speaker verification focuses on whether the claimed speaker is the true speaker, a yes or no problem. In this paper, only speaker verification is discussed. In order to distinguish whether the speech is truly said by the claimed speaker, the speaker's speech needs to enroll in advance which is called enrollment data. And the speech to be distinguished is called test data. According to whether the text of test speech is the same as that of enrollment speech, two kinds of speaker verification systems are involved: text-dependent and text-independent. Since text-dependent speaker verification systems strictly constrain the speech text of speaker, it is easier to recognize than text-independent systems and the accuracy of the recognition result is higher. In order to strengthen our conclusion, both of text-dependent and text-independent speaker verification experiments will be carried out.

In generally, speaker verification consists of three stages: frond-end feature extraction, modeling, and back-end scoring or classification. In the front-end cepstral feature extraction, mel-frequency cepstral coefficients (MFCCs) are usually used. But in our experiments features represented by perceptual linear prediction (PLP) coefficients were found to obtain better recognition results. Hence, PLPs were adopted as the acoustic features in this paper. Widely studied speaker modeling approaches include Vector Quantization (VQ) model [1], Gaussian Mixture Model (GMM) [2], Support Vector Machine (SVM) [3], Artificial Neural Networks (ANNs) [4] et.al. Over the past decades, GMMs have been the dominant approaches for modeling speakers. A series of creditable methods are GMM-based: Gaussian Mixture Model-Universal Background Model (GMM-UBM) [5], Joint Factor Analysis (JFA) [6] and i-Vector [7]. In the back-end scoring or classification method, likelihood ratios and SVMs are usually used. Cosine Distance classifier is also put forward along with the advent of i-Vector. Although JFA and i-Vector have demonstrated the state-of-the-art performance for text-independent speaker recognition in the NIST speaker recognition evaluations (SREs) [7][8], they are complicated and all based on GMM modelling with conventional acoustic features. Since feature engineering is the focus of this paper, GMM-UBM is adopted as the primary modelling approach here. It is worth noting that the proposed deep features can easily fit into the JFA or iVector framework in the future.

In the early years of speaker recognition research, neural network has been applied to speaker recognition tasks as a classifier or to strengthen other classifiers [9][10]. Similar ideas have been extended in recent years. In [11], one neural network with feature after Z-norm is trained for each speaker for verification; in [12], hierarchical neural network is used to improve performance. These approaches all require some forms of speaker-specific network to be trained and are usually not easy to scale up to tasks with large number of speakers. Another category is to use neural network to help the extraction of i-Vector [13][14][15][16]. These approaches are usually complicated and the gain sometimes is limited [14]. Using neural network to extract Tandem features is an effective approach in speech recognition [17]. Motivated by this, researchers have also tried to use speaker id as the target to train neural network based feature extractor [18]. However, the performance gain is not significant and is hard to reproduce. In this paper, a novel deep feature extraction approach is proposed. To take advantage of large amount of speech recognition data, features

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 No. 201302060012.

978-1-4799-1484-5/14/$31.00 ©2014 IEEE

are extracted from RBM or DNN trained for large vocabulary continuous speech recognition. Output of hidden layers from the neural networks are used as the raw features. These features are processed in a Tandem way as in speech recognition. The extracted Tandem features are then concatenated with the original acoustic features to form a new frame-level features. Once the new features are obtained, they can be used with any classifiers. In this paper, GMM-UBM framework is adopted for the experiments. To our best knowledge, this is the first systematic study of using deep Tandem features for speaker verification.

The structure of this paper is as follows. Section II reviews the basic concepts and procedures of GMM-based speaker verification and DNN and RBM. Section III describes motivations and details of the proposed approaches. Section IV gives detailed description of experiments and section V concludes the whole work.

II. Review of mainstream GMM-based methods in speaker recognition and deep learning

In this section, firstly traditional feature extraction procedures are described. Then popular GMM-based approaches like adapted universal background model (GMM-UBM), joint-factor analysis, i-Vector are briefly introduced. The fundamental principals will be stated and basic formula will be listed. Besides, the basic concepts and theories of DNN and RBM will also be reviewed.

A. Short-term spectral feature extraction

It is important to extract feature vectors from each speech frame which can capture the speaker's specific characteristic. Usually short-term spectral features are extracted in speech and speaker technologies, as is known that speech signal changes continuously. Within a short time about 20-30ms, it is assumed to remain stationary. In speech and speaker fields, speech is cut into short frames usually about 10ms and the feature vector is extracted from each frame.

Generally before signal transforming, the frame is pre-emphasized and multiplied by a smooth window function such as hamming. The window function is needed because of the finite length of the discrete Fourier transform (DFT). DFT decomposes the speech signal into frequency components and usually only the magnitude spectrum is retained. Then band-pass filters is got with energy integration over neighboring frequency bands.

Although the sub-band energy values are a kind of features, usually they are further transformed into a lower dimensionality feature, the so-called mel-frequency cepstral coefficients (MFCCs). MFCCs are widely used in speech and speaker recognitions and were introduced in early 1980s for speech recognition and then adopted in speaker recognition [19]. Through decades MFCCs are all the way the first priority features in speaker technologies. MFCCs can be further transformed into another feature,called perceptual linear prediction (PLP) coefficients. In our previous experiments, it has been found that PLP features yielded slightly better results than MFCC features. Hence, in this paper, PLP is used as the basic acoustic feature.

B. GMM-UBM based speaker verification system

In recent years, GMM-based approaches have received considerable attention in speaker verification. A widely accepted approach is the classical Maximum A Posterior (MAP) adaptation of Universal Background Model (UBM) parameters(GMM-UBM)[5]. The GMM-UBM approach for speaker verification consists of three stages. Different stages require different type/amount of data.

1) **UBM Training:** Each speaker can be modeled as a GMM model using feature vectors extracted from his or her speech. But in general for a single person, there is not that much data enough to train a complete GMM which covers this person's all possible speech data he or she would say. Thus a speaker-independent background GMM model trained with data from large amount of speakers which represents the general speaker independent distribute of speech acoustic features, called UBM, is needed. The UBM parameter is trained with the iterative Expectation-Maximization (EM) algorithm. This stage normally requires large amount of unlabelled data.

2) **Enrollment stage:** 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 as figure 1 shows. This adaptation is similar to EM algorithm and is identical to EM in the first step. But in the second step, unlike the EM algorithm, the new sufficient statistics estimates are combined with sufficient statistics from UBM parameter using a data-dependent mixing coefficient [5]. This adaptation would tune the parameters of GMM mixtures for the data which can be observed in the speaker's enrollment speech and mixtures parameters for those which is not seen in the speaker's enrollment speech are kept unchanged, copied from the UBM. This stage requires as much as possible speaker-specific data. The output of this stage is a number of speaker-dependent models. Since the target number of speakers can be very large, the output model can not be complicated. Since a GMM model is relatively simple, it is possible to scale up to large tasks with GMM-UBM.

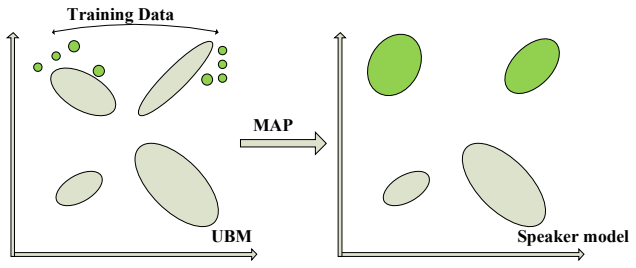

Fig. 1. The MAP process

3) **Test stage:** Likelihood ratio decision method is used in the stage. Given an observation O which represents the feature extracted from a test utterance , and a speaker S, there can be two hypotheses:

$$\begin{aligned} H_0 &: O \text{ is from speaker } S \\ H_1 &: O \text{ is not from speaker } S \end{aligned} \quad (1)$$

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

$$\Lambda = \frac{1}{T} Log \frac{p(O|H_0)}{p(O|H_1)} = \begin{cases} \geq \theta & \text{accept } H_0 \\ < \theta & \text{accept } H_1 \end{cases} \quad (2)$$

Where $P(O|H_i), i = 0, 1$, is the probability of hypothesis H_i being true, which can be represented by computing the probability density function for O given the target speaker GMM model or the impostor GMM model. Usually UBM model acts as impostor model in the test stage. T represents the number of frames of Observation O.

The whole GMM-UBM framework is shown in figure 2.

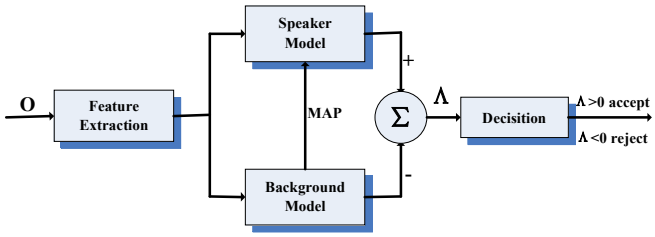

Fig. 2. The GMM-UBM framework

C. Factor analysis based speaker verification

With the speech channels becoming increasingly complicated and audios recorded in various of circumstances, many techniques for channel compensation are applied to speaker recognition. Joint factor analysis is a model of speaker and session variability in GMM's. A GMM is estimated for each target speaker and attempt to remove the session variability hence to compensate for inter-session variability and for channel mismatches between enrollment data and test data [20]. Generally, a speaker utterance is represented by a supervector (M) which derives from the association of mean vectors of all mixture components in the speaker GMM. This speaker-dependent supervector can be decomposed as:

$$M = m + Vy + Ux + Dz \quad (3)$$

Where m is a speaker- and session-independent supervector generated from UBM. Matrix V and D define speaker subspace and U defines a session subspace. The likelihood of a test speech feature vectors can be computed using the channel-compensated model $(M - Ux)$.

In i-Vector approach, a single subspace called total variability is proposed. And new speaker- and channel-dependent GMM supervector is redefined as:

$$M = m + Tw \quad (4)$$

Where T is a low rank variability matrix of speaker and session and the total factors w is called i-Vector. I-Vectors are considered as front-end low dimension features and a fast scoring method of deciding whether two utterances come from the same person is comparing the angle between the two utterances' i-Vectors (Cosine Distance classifier) to a threshold:

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

From the above, either JFA or i-Vector are based on GMM models with standard acoustic features. Hence, any new features which are compatible with GMM models, can also be incorporated into JFA and i-Vector. Since in this paper, the focus is deep feature engineering, JFA and i-Vector will not be further tested.

D. Deep Neural Network and Restricted Boltzmann Machine

In the year 2006, Hinton proposed a deep learning algorithm of neural network, which made it possible to train neural networks of at least 7 hidden layers. And the multi-layer neural network is called Deep neural network. Deep learning exhibits strong representational power to nonliear modelings [21]. Recently, DNN is widely used in automatic speech recognition as well as many other fields in machine learning. Context-Dependent Pre-trained Deep-Neural-Network HMMs, or CD-DNN-HMMs hybrid model [22], achieved a dramatic performance for large vocabulary continuous speech recognition (LVCSR) and becomes the state-of-the-art method and attract prodigious attention.

DNN-HMM is taking the prime place of GMM-HMM in recent acoustic modeling research work. DNN-HMM can automatically model the long-span deep features while GMM-HMM is more stable and robust.To take the convenience of both, several methods which extract features from deep neural network and use the features in GMM-HMM acoustic modeling have be proposed, such as, tandem method [23] uses posterior probability which is the output of the last soft-max layer in DNN, or weighted sum of the output in last hidden layer, bottleneck-feature [24] method forcedly sets a bottleneck in the middle of DNN and uses the weighted sum of output before the bottleneck layer. However, none of this kind of approaches has shown an outperformed result than the best DNN-HMM, until a so-called scalable approach [25] has been proposed, which contains three stages as following, training a DNN feature extractor, deriving features with DNN, modeling DNN-derived features with GMM-HMM model.

In DNN-HMM modeling in speech recognition, DNN is trained using cross entropy criterion according to the force alignment of frame level state prediction from the state-of-art ASR system. Hence, DNN is a supervised training approach. In the extracting stage, all the data are fed into the DNN to get the weighted sum of the output before the last hidden layer. After that, principal component analysis are used to project the high dimensional DNN output into a low dimensional space, and then cascaded with original spectral features which is the DNN input. In the application stage, the spectral and DNN-derived tandem features are directly put into the well-tuned conventional GMM-HMM ASR system to get a better performance compared to both DNN-HMM and GMM-HMM system.

Since DNN employs supervised training, the derived features are regarded as *supervised feature* which reflects the property of training labels (target). For speech recognition,

the features are inclined to discriminate between phones or context-dependent phone states. In contrast, deep belief network is an unsupervised approach to train neural networks. It is usually constructed as the initial network for DNN training using the stack of Restricted Boltzmann Machine (RBM).

Restricted Boltzmann Machine derives from Boltzmann Machine, a bidirectionally connected network of stochastic processing units, inheriting its merits on learning important aspects of an unknown probability distribution based on samples from distribution [26].

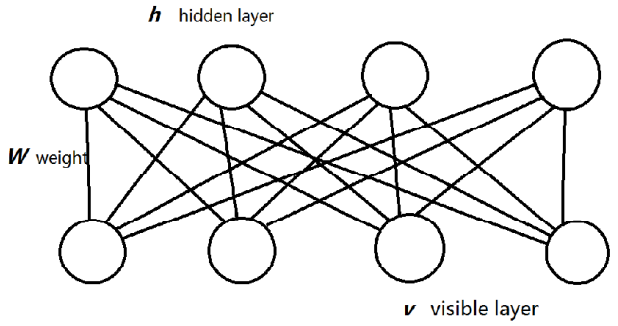

Fig. 3. Restricted Boltzmann machine

RBM is an undirected model, one being visible layer (v) and the other hidden layer (h). In graph theory it can be regarded as a bipartite graph, as the figure 3 shows, each edge in the bipartite graph being attached with a weight, noted as a matrix W.

Suppose RBM system has n vertexes in visible layer and m vertexes in hidden layer. vector h and v stand for the state in hidden layer and visible layer respectively, among which h_i and v_j stand for the state of the $i-th$ vertex in hidden layer and the state of the $j-th$ vertex in visible layer. Then the energy of the RBM is defined as follows:

$$E(\mathbf{v},\mathbf{h}|\theta) = -\sum_{i=1}^{n} a_i v_i - \sum_{j=1}^{m} b_j h_j - \sum_{i=1}^{n}\sum_{j=1}^{m} v_i W_{ij} h_j \quad (6)$$

The joint probabilistic distribution shown in (7) can be infered from the formula above, among which Z is called partition function and only relevant to the parameter which is needed to estimate:

$$p(\mathbf{v},\mathbf{h}|\theta) = \frac{e^{-E(\mathbf{v},\mathbf{h}|\theta)}}{Z(\theta)} \quad (7)$$

$$Z(\theta) = \sum_{\mathbf{v},\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h}|\theta)} \quad (8)$$

To determine this distribution, the partition function is needed to computed, which need exponential time computation, above usual ablility. Thus, these parameter instead of direct computation, including a_i, b_j, W_{ij} can only be estimated. And what is useful to us is only the marginal distribution of v, which can be represented as:

$$p(\mathbf{v}|\theta) = \frac{1}{Z(\theta)} \sum_{\mathbf{h}} e^{-E(\mathbf{v},\mathbf{h}|\theta)} \quad (9)$$

The activation state of the hidden vertex only depends on the vertex in the visible layer,

$$P(h_j = 1|\mathbf{v},\theta) = sigmoid(b_j + \sum_{i=1}^{n} v_i W_{ij}) \quad (10)$$

and vice versa for symmetry.

$$P(v_i = 1|\mathbf{h},\theta) = sigmoid(a_i + \sum_{j=1}^{n} h_j W_{ji}) \quad (11)$$

It can be seen that each layer is related only with the previous layer and the whole procedure can be regarded as layer-wise train. Since the training does not require any label information, the derived features is considered as *unsupervised features*.

III. DNN AND RBM AS FEATURE EXTRACTORS IN GMM-UBM SPEAKER VERIFICATION SYSTEMS

A. Previous research on using neural network for speaker verification

Due to the significant performance of DNN achieved in speech recognition, research of applying DNN to speaker verification has drawn special attention. There are mainly two ways to apply deep learning to speaker verification: model-based or feature-based.

Most model-based approaches employ neural network as a classifier or to strengthen other classifiers [9][10]. Model based approaches normally require speaker-specific network to be trained, which means for each test speaker, there will be a distinct neural network or similar parameters. Due to the sparse enrolment data, the estimation of the speaker-specific neural network or parameters is not robust. Several approaches are applied to address this issue, for example, in [12], hierarchical neural network is used. Or in i-Vector based approaches, the number of the estimated parameters (i-Vector) is kept small [14][15]. However, the performance gain from model-based approach has not been shown to be significant unless very complicated models are used [14][15].

Feature-based approaches employ deep learning to extract compact and representative features for speaker verification. When using supervised deep feature, there is an issue of what labels to choose as the target for training DNN. Early in the year 1998, Konig [27] tried to use bottleneck features to build GMM-UBM system. Neural network was trained with a bottleneck layer as the middle hidden layer. The input is expanded context-frame feature vectors and the output label is speaker id. The extracted bottleneck features are used to build GMM-UBM speaker verification systems. Experiment results showed that although separate bottleneck-feature-based systems performed worse than the original-feature-based systems, the performance exceeded them after back-end scoring

linear combination of the bottleneck-feature-based and the original-feature-based systems. Similar idea was enhanced in [18] and showed slight gains. However, the results can not be reproduced by us. To our knowledge, there has not been works on using phone labels to train the neural network features for speaker verification, although they are widely used in speech recognition. As for unsupervised neural network, there has not been reports on using them for feature extraction in speaker verification.

B. Deep Tandem Features for Speaker Verification

As indicated before, deep learning has achieved significant gains within the DNN-HMM framework in various speech recognition tasks [28]. Motivated by the Tandem feature processing in speech recognition, in this paper, Tandem deep features are proposed to be used for speaker verification. Two types of deep features are investigated as below:

1) **Supervised deep feature: DNN Tandem feature:** In this paper, deep neural network, which is widely used for speech recognition, is used to extract tandem features to apply in GMM-UBM speaker verification systems. Figure 4 displays the structure of the our work. We do so because of following reasons. As far as we know, much audio data with text labels can be obtained for speech recognition while a great deal of data is lack of information about speakers. Further more, the input nodes of DNN are expanded multi-frame features, which can carry more reliable information and more distinguishable. The phone-dependent network also carries speaker information and in that way is speaker-dependent. Of course, the layer nearer to the output layer is less speaker-dependent because the speaker information is omitted to be phone-dependent. Thus, the performance of features from the last hidden layer is worse than that of features from the middle hidden layer. Experiments will be set up to confirm this.

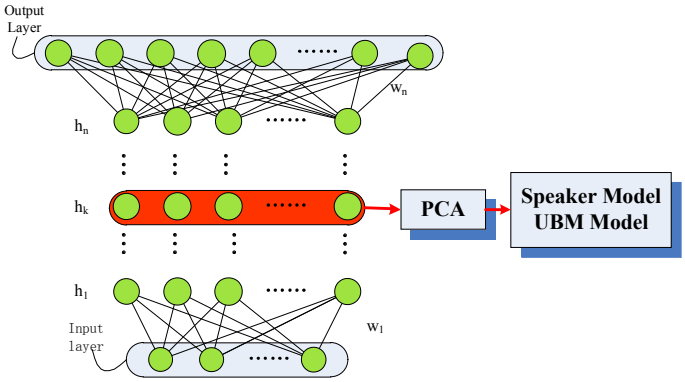

Fig. 4. The framework of speaker verification with deep features

2) **Unsupervised deep feature: RBM Tandem feature:** As unsupervised RBM models the input features to more regular and discriminative features, and unlike speech DNN tuned by back propagation, it is not more phone-oriented,and also its inputs are multi-frame features, we believe that RBM feature extractor is more applicable. The Tandem processing of the RBM features is similar to that of DNN.

Once the Tandem deep features are extracted, they can be combined with the original acoustic features at either front-end feature level or back-end score level. At feature level, Tandem features are simply concatenated to the original acoustic features. At score-level, two systems with Tandem and original acoustic features are built separately and the scores from the two systems are weighted averaged to form the final score for classification. Experiments of both combination approaches will be investigated.

IV. Experiments

To evaluate the performance of the proposed approach, experiments of both text-independent and text-dependent speaker verification were performed. First, text-independent experiments on NIST Speaker Recognition Evaluation (SRE) data were carried out.

A. Experimental Setup and Baseline

1-conversation data from NIST 2005 SRE was chosen for the UBM training data and only the male data was used. Each audio is a five-minute telephone conversation between two speakers, containing roughly two minutes of speech after removing unvoiced segments by voice activity detection (VAD) for each speaker. In total, there are about 9 hours of data consisting of 274 speakers. Evaluation was carried out on the core condition of NIST 2006 SRE. The characteristics of this data is similar to that of NIST 2005 SRE and also only the male data was chosen. There are about 352 target speakers with two minutes of speech for each person and 645 test audios. The total number of tests is 10037.

13-dimension PLP coefficients were extracted at every 10ms for the baseline system. Then, the first and second derivative were calculated to form a 39-dimensional feature. The PLP segments belonging to the same person were concatenated together to an integrated feature file after VAD. Mean and variance normalization were applied to each file.

First, a universal background model (UBM) was trained with NIST 2005 SRE containing 512 Gaussian mixtures. Then, 352 target speaker models were trained using standard MAP adaptation on top of the UBM with the enrollment data from NIST 2006 SRE (enrollment data are different from test data although from the same speaker). In the test stage, each test file will get two sets of average log likelihood scores per frame using the target model and the UBM model respectively. Then the final score is calculated as the difference of the two scores. With different threshold of the final score, it is possible to obtain a number of different false acceptance or rejection errors. False rejection rate represents the proportion of true speakers being incorrectly rejected and false acceptation rate indicates that of impostor speakers being incorrectly accepted. The evaluation metric being used was *Equal Error Rate* (EER) which is the error rate where false rejection rate equals false acceptance rate. The baseline system, i.e. GMM-UBM with PLP features, yielded an EER of 11.18%.

B. Text-Independent Speaker Verification with Deep Features

As indicated before, deep features from existing speech recognition system can be borrowed for speaker verification. The data used for training the DNN and RBM was a 309-hour switchboard English data set [29]. It consists of 4869 speakers which are different from the NIST SRE data. Since

the DNN and RBM were trained on large amount of data, it is possible to get better deep feature representation using these deep networks.

Both DNN and RBM have 7 hidden layers with 2048 nodes per layer. The input layer has 429 nodes with 11 PLP feature frames (Each frame was expanded to 11 frames with the left and right 5 frames). The output layer has 9296 nodes corresponding to 9296 tri-phone states. The DNN was trained on top of the RBM using back-propagate algorithm with cross-entropy objective function, along with a L2-norm weight-decay term of coefficient 10^{-6}. Hence, DNN generated phone related supervised deep features while RBM generated unsupervised deep features.

1) Projection Dimension of Deep Features: The expanded 39-dimensional PLP features were passed through the network (RBM or DNN) and generate 2048-dimensional new features at each layer of the network. After applying PCA projection to the 2048-dimensional feature, a new low-dimension tandem feature was obtained and after mean and variance normalization the final low-dimension tandem feature was obtained. As the dimension reduction process is a common practice in speech recognition, it is useful to first investigate the effect of dimensionality. The performance of deep feature extracted at layer 7 with different dimensions are shown in table I:

TABLE I. EER (%) OF DEEP FEATURES OF DIFFERENT DIMENSIONS AFTER PCA

Dimension	RBM	DNN
20	11.17	15.54
39	10.33	15.69
78	12.19	14.82

It can be observed that there is no consistent performance change trend on dimensions. With appropriate dimension (20 or 39), unsupervised deep feature (RBM feature) already outperformed the PLP baseline system (11.18%). The supervised deep features (DNN feature) showed worse performance and larger dimension did not help much. To achieve the best performance, 39 dimension was used in later experiments.

2) Combination of PLP and Deep Features: Although deep features can yield performance gain as shown in the previous section, the gain is small. It has been observed that the errors of the PLP baseline and the errors of the systems with deep features have different patterns. This implies that combination of the two systems may be helpful. There can be two ways of system combination. The first is feature-level combination, i.e. concatenating the original PLP feature and the deep tandem features to form an augmented feature. The second is score-level combination. Two scores of a test speech from the deep-feature GMM-UBM and PLP-feature GMM-UBM were linearly added together with a weight. The weight of the better score was set 0.7 and that of the worse score 0.3. The two combination approaches were investigated with the same setup as the previous section. The results are shown in the below table.

From table II, all system combination approaches obtained significant performance improvement and outperformed the baseline PLP system. Score-level combination is more effective for the DNN feature, while feature-level combination is more effective for the RBM feature. The best performance is still from systems with unsupervised features. Feature-level combination will be used in later experiments.

TABLE II. EER (%) OF SYSTEM COMBINATION

System	Single System	Combined System	
		Feature	Score
PLP	11.18	—	—
+ DNN	15.69	11.03	10.74
+ RBM	10.33	10.02	10.31

3) Deepness of Features: There has been an assumption that "*deep*" is an important factor to achieve good performance. It is then interesting to investigate the performance of features from different layers of DNN or RBM. The same experimental set up as the previous section was used. Performance of the systems with feature-level combination is shown in table III.

TABLE III. EER (%) OF SYSTEMS WITH FEATURES FROM DIFFERENT LAYERS

# Layer	RBM	DNN
1	10.16	10.31
2	10.02	10.32
4	**9.75**	10.45
7	10.02	11.03

From table III, all combined systems obtained significantly better performance than the PLP baseline. However, for supervised feature, deep features are not necessarily better than shallow features. The deeper the network is, the worse the performance is. This might be because that DNN was tuned to phone state posteriors while the aim of speaker verification is to discriminate between speakers. In contrast, deeper feature obtained better performance for the unsupervised RBM features. It is interesting to note that the best performance was actually achieved in the middle layer. This implies that deep is not always useful. Nevertheless, table III has shown the power of incorporating deep features into speaker verification. Compared to the PLP baseline, the best deep feature system obtained about 12.8% relative performance improvement, which is significant.

C. Text-dependent Speaker Verification with Deep Features

To strengthen the reliability of our judgments, experiments of text-dependent speaker verification systems were also performed on a Chinese task. About 12.5 hours mobile audio data were chosen as the training data to build a 512-mixture GMM-UBM system. 51 target speakers were chosen to be enrolled. In the enrollment stage, there were only 3 utterances of the same text to be recorded for each speaker, each about just 3 seconds. In the test stage, each speaker had 7 test utterances, which consist of 1 true utterance and 6 impostors' utterances. Among the 6 impostors' utterances, 2 utterances were said by the true speaker but the texts were different from this speaker's enrollment utterances, 2 utterances were said by impostors but the texts were the same as the enrollment utterances, and 2 utterances were from impostors with texts different from enrollment data. There were 7140 test files in total.

In this experiment, the best setup from the pervious section was used for deep feature extraction. From table IV, similar

TABLE IV. EER (%) OF TEXT-DEPENDENT SPEAKER VERIFICATIONS SYSTEMS WITH DEEP FEATURES

System	EER
PLP	0.98
+ RBM	**0.88**
+ DNN	0.98

trends can be observed that supervised deep features performed worse than unsupervised features. Even with combination, PLP+DNN system still did not outperform the baseline. This might be because the equal error rate of the baseline is already very low. However, in contrast, combined systems with RBM features can still significantly outperform the PLP baseline. It is worth noting that, with a good PLP baseline, performance improvement is hard. This again shows that unsupervised deep features can yield better and complementary feature representation for speaker verification.

V. CONCLUSIONS

This paper proposes a novel approach to extract deep features for speaker verification. Well trained DNN and RBM in speech recognition are employed as the feature extractor, which shows the potential of using extra large amount of data. Tandem features are then extracted and combined with original acoustic features. The proposed approach is compatible with most existing speaker verification algorithms and is easy to apply. Experiments showed that, within a GMM-UBM framework, GMM-UBM with unsupervised deep features achieved signifiant performance improvement compared to the standard GMM-UBM approach in both text dependent and text independent tasks. The gain reported is a lot more significant than previous neural network based speaker verification. In the future, we will investigate how to apply deep features to more complex speaker verification framework such as i-Vector.

REFERENCES

[1] Ahsanu Kabir,Sheikh Mohammad Masudul Ahsan, "Vector Quantization in Text Dependent Automatic Speaker Recognition using Mel-Frequency Cepstrum Coefficient", *6th WSEAS International Conference on circuits, systems, electronics, control signal processing*, Cairo, Egypt. 2007: 352-355.

[2] Douglas A.Reynolds, "Speaker identification and verification using Gaussian mixture speaker models", *Speech communication* 17.1 (1995): 91-108.

[3] William M.Campbell, Douglas E.Sturim, Douglas A.Reynolds, "Support vector machines using GMM supervectors for speaker verification", *Signal Processing Letters*, IEEE 13.5 (2006): 308-311.

[4] Kevin R.Farrell, Richard J.Mammone, Khaled T.Assaleh, "Speaker recognition using neural networks and conventional classifiers" *Speech and Audio Processing*, IEEE Transactions on 2.1 (1994): 194-205.

[5] Douglas A.Reynolds, Thomas F.Quatieri, Robert B.Dunn, "Speaker verification using adapted Gaussian mixture models", *Digital signal processing*, 10.1 (2000): 19-41.

[6] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, Pierre Dumouchel, "Speaker and session variability in GMM-based speaker verification", *Audio, Speech, and Language Processing*, IEEE Transactions on 15.4 (2007): 1448-1460.

[7] Najim Dehak, Patrick J.kenny, Reda Dehak, Pierre Dumouchel, Pierre Ouellet, "Front-end factor analysis for speaker verification" *Audio, Speech, and Language Processing*, IEEE Transactions on 19.4 (2011): 788-798.

[8] Martin, Alvin F., and Craig S. Greenberg. "The NIST 2010 speaker recognition evaluation." *INTERSPEECH*. 2010.

[9] Kevin R.Farell, Richard J.Mammone, Khaled T.Assaleh, "speaker recognition using neural network and conventional classifier",*Speech and Audio Processing, IEEE Transactions on*,1994: 194-205.

[10] Wouhaybi, Rita H, Adnan Al-Alaoui, "Comparison of neural networks for speaker recognition", *Electronics, Circuits and Systems*, 1999: 125-128.

[11] E.Turajlic, O.Bozanovic, "neural network based speaker verification for security system", *Telecommunications Forum*,(TELFOR), 2012 20th.IEEE, 2012: 740-743.

[12] Joydeep Ghosh , Brian J.Love , Jennifer Vining , Xuening Sun, "Automatic Speaker Recognition Using Neural Networks[J]." 2004.

[13] Mohammed Senoussaoui, Najim Dehak, Patrick Kenny, Reda Dehak, Pierre Dumouchel, "First attempt of Boltzmann Machines for Speaker Verification.", Odyssey 2012-The Speaker and Language Recognition Workshop. 2012.

[14] Lukas Burget, Oldrich Plchot, Sandro Cumani, Ondrej Glembek, Pavel Matejka, Niko Brummer, "Discriminatively trained probabilistic linear discriminant analysis for speaker verification." *Acoustics, Speech and Signal Processing (ICASSP)*, 2011 IEEE International Conference on. IEEE, 2011: 4832-4835.

[15] Vasileios Vasilakakis, Sandro Cumani, Pietro Laface, "Speaker recognition by means of Deep Belief Networks.",2013.

[16] Samuel Thomas, Sri Harish Mallidi, Sriam Ganapathy, Hynek Hermansky, "adaptation transform of auto-associative neural networks as feature for speaker verification", *Odyssey 2012-The Speaker and Language Recognition Workshop*. 2012.

[17] Marc Ferras, Herv Bourlard, "MLP-based factor analysis for tandem speech recognition[C]",*Acoustics, Speech and Signal Processing (ICASSP)*, 2013 IEEE International Conference on. IEEE, 2013: 6719-6723.

[18] S. Yaman, J. Pelecanos, and R. Sarikaya. "Bottleneck Features for Speaker Recognition." *Odyssey*.2012

[19] Tomi Kinnunen, Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors", *Speech communication*, 52.1 (2010): 12-40.

[20] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, Pierre Dumouchel, "Joint factor analysis versus eigenchannels in speaker recognition", *Audio, Speech, and Language Processing*,IEEE Transactions on 15.4 (2007): 1435-1447.

[21] Nicolas Le Roux, Yoshua Bengio, "Representational power of restricted Boltzmann machines and deep belief networks", *Neural Computation*,2006,20(6): 1631-1649.

[22] George E.Dahl, Dong Yu, Li Deng, Alex Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition", *Audio, Speech, and Language Processing*, IEEE Transactions on 20.1 (2012): 30-42.

[23] Hermansky, Hynek, Daniel PW Ellis, and Sangita Sharma. "Tandem connectionist feature extraction for conventional HMM systems." Acoustics, Speech, and Signal Processing, 2000. ICASSP'00. Proceedings. 2000 IEEE International Conference on. Vol. 3. IEEE, 2000.

[24] Grezl, Frantisek, et al. "Probabilistic and bottle-neck features for LVCSR of meetings." Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on. Vol. 4. IEEE, 2007.

[25] Yan, Z., Qiang Huo, Jian Xu, "A scalable approach to using DNN-derived features in GMM-HMM based acoustic modeling for LVCSR.", *INTERSPEECH*, 2013.

[26] Asja Fischer, Christian Igel, "An introduction to restricted Boltzmann machines.",L.Alvarez et al. (Eds): CIARP 2012, LNCS 7441,pp. 2012:14-36.

[27] Yochai Konig, Larry Heck, Mitch Weintraub, Kemal Sonmez, "Nonlinear discriminant feature extraction for robust text-independent speaker recognition." *Proc. RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications*. 1998: 72-75.

[28] Frank Seide, Gang Li, Dong Yu, "Conversational Speech Transcription Using Context-Dependent Deep Neural Networks.", *INTERSPEECH*, 2011: 437-440.

[29] John J Godfrey, Edward Holliman, "Switchboard-1 release 2", *Linguistic Data Consortium*, Philadelphia,1997.