Tandem Deep Features for Text-Dependent Speaker Verification

Tianfan Fu, Yanmin Qian, Yuan Liu, Kai Yu

Key Lab. of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering
MOE-Microsoft Key Lab. for Intelligent Computing and Intelligent Systems
Department of Computer Science and Engineering, Shanghai Jiao Tong University
{erduo, yanminqian, liuyuanthelma, kai.yu}@sjtu.edu.cn

Abstract

Although deep learning has been successfully used in acoustic modeling of speech recognition, it has not been thoroughly investigated and widely accepted for speaker verification. This paper describes an investigation of using various types of deep features in a Tandem fashion for text-dependent speaker verification. Three types of networks are used to extract deep features: restricted Boltzmann machine (RBM), phone discriminant and speaker discriminant deep neural network (DNN). Hidden layer outputs from these networks are concatenated with the original acoustic features and used in a GMM-UBM classifier. The systems with Tandem deep feature were evaluated on RSR2015, a short-term text dependent speaker verification task. Experiments showed that the best Tandem deep feature obtained more than 50% relative EER reduction over the traditional feature in a GMM-UBM framework.

Index Terms: Speaker Verification, Tandem Feature, Feature Extractor, Deep Neural Network.

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 have been deployed in commercial applications in recent years. 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 and the knowledge of lexicon is integrated in the modeling, it is easier to recognize than text-independent systems and the accuracy of the recognition result is higher. Considering that the drastic limitation in terms of speech duration is especially demanded in the real scenario, the text-dependent speaker verification is more appropriate to be implemented in the real applications to address the problem of lack of data. However, recent breakthroughs on speaker verification mainly focused on the text-independent technologies which may not benefit text-dependent applications. Therefore, the research on text-dependent speaker verification has attracted much interest from both academia and industry, and this is also the main work of this paper.

The general procedure of speaker verification can be mainly divided into three phases, including front-end feature extraction, modeling and back-end classification. In the first step, mel-frequency cepstral coefficients (MFCCs) or perceptual linear prediction (PLP) coefficients feature are widely used as the front-end cepstral features. Then the various approaches are applied to build models, commonly including 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 GMM-based approaches are proposed, such as GMM-UBM [5], Joint Factor Analysis (JFA) [6], and the state of the art technology-iVector [7]. However, despite the success of the iVector paradigm, its applicability in text-dependent speaker verification remains questionable, and some work shows that conventional approaches, such as GMM-UBM, are superior to the iVector-based ones in this scenario [8][9][10][11].

The history of neural network based Speaker Verification can be traced back to the late of last century: The work in [4] proposed a method named Modified Neural Tree Network (MNTN), which is a hierarchical classifier that combines the property of decision trees and feedforward neural networks . Moreover, Auto-Associative Neural Network (AANN) is widely used in speaker verification, such as [12] and [13]. In recently work in [14], a mixtures of auto-associative neural network for speaker verification is proposed. Besides, the neural network is also applied on the feature level. Such as in [15], a nonlinear discriminant analysis (NLDA) which uses multi-layer perceptron (MLP) was proposed to aim at transforming acoustic feature into low-dimensional speaker-discriminative feature. However, the shallow architecture of these neural networks leads to its limited ability to fit nonlinear features.

In recent years, DNN achieves breakthrough in speech recognition [16], and the field of speaker recognition also begins to pay attentions to DNN. The initial attempts are performed in some work, such as the work in [17] and our previous work in [18], however, the results is not comparable to the state of the art system, so the application of DNN on the speaker verification needs more exploration. In this paper, the DNN-based techniques are investigated here and some feature level based deep modeling approaches are proposed. The scenario focused here is text-dependent which is more realistic in the real applications as described above. And it is noted that the proposed deep models based feature extraction can also be compatible with other modeling approaches including JFA and i-Vector, which can be combined in the future to get additive improvement.

The remainder of this paper is organized as follows. Section II reviews the RBM and DNN briefly. Section III describes our proposed deep feature extractors in detail. The experimental comparisons and results are presented in Section IV and the whole work is summarized in Section V.

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.

2. Supervised and Unsupervised Deep Learning

There are two types of deep learning with different optimization methods and usage properties: supervised and unsupervised deep learning. In this section, typical approaches of the two types, namely deep neural network and restricted Boltzmann machine, are re-visited briefly.

Deep Neural Network (DNN) is a feed-forward artificial neural network with more than one *hidden* layers. It is widely used for classification and its output layer usually consists of the classes which it can discriminate. Hence, it is a *supervised* deep learning approach. For example, in state-of-the-art speech recognition, DNN is used to classify cross-word triphone states within a hybrid DNN-HMM [19] framework. Depending on tasks, other class labels, such as speakers or phones from different languages can also be used as the outputs of DNN. Cross entropy is a widely used criterion to optimize DNN parameters:

$$\mathcal{L}(\theta) = -\sum_{s} d_s \log P(s|\mathbf{o}, \theta) \tag{1}$$

where θ denotes the set of DNN parameters, and $P(s|\mathbf{o})$ is the posterior probability of the output class s, d_s is 1 for the target class and 0 for the non-target classes. Since DNN is a supervised learning approach, back-propagation (BP) algorithm [20] is usually used to optimize equation (1). One big problem of BP is that it can easily get trapped in a poor local optima. Hence, the initialization of DNN is particularly important. Unsupervised deep learning is usually used for this occasion.

A well-known approach of *unsupervised* deep learning is *restricted Boltzmann machine* (RBM). It derives from Boltzmann Machine, a bidirectionally connected network of stochastic processing units. RBM contains symmetric hidden and visible layers. It can be regarded as a bipartite graph, where the stochastic units in visible layer only relies on the stochastic units in the hidden layer. The distributions of the visible and hidden layers can be expressed as:

$$p(\mathbf{v}, \mathbf{h}|\theta) = \frac{e^{-E(\mathbf{v},\mathbf{h}|\theta)}}{Z(\theta)} \tag{2}$$

where $E(\mathbf{v}, \mathbf{h}; \theta)$ is an energy function and $Z = \sum_v \sum_h e^{-E(\mathbf{v},\mathbf{h};\theta)}$ is the normalization term. By choosing a specific energy function, the conditional probability of h and v can be efficiently calculated. Contrastive divergence algorithm can be used to optimize parameters by maximising the log likelihood $\log p(\mathbf{v}|\theta)$. Since there is no class labels involved in RBM training, it is an unsupervised approach and widely used for initializing a deep neural network.

Both DNN and RBM are multi-layer neural networks which transforms input features using complicated non-linear transformations. Although they can be regarded as *model* for classification or clustering, a widely accepted opinion is to view them as *feature extractors*. It is worth noting that, from the feature extraction viewpoint, RBM and DNN have different properties. Due to the unsupervised learning nature, RBM extracts global high-level features without any bias. In contrast, DNN extracts classification task-specific features due to the use of class labels.

3. Deep Feature Extraction for Text-Dependent Speaker Verification

Short-term spectral features extracted from raw waveform signals, such as MFCCs and PLPs, are most widely used in the speech and speaker recognition. However, there are two disadvantages of using these spectral features for speaker verification: 1) The features extracted in a short time can not well represent sound characteristics of a relatively long duration, such as speaker identity; 2) The spectral features are originally designed for speech recognition, not speaker verification. Although they can be used for both speech and speaker recognition, there is still inconsistency between the traditional spectral features and the task of speaker verification. Hence, it is important to construct features which are more specific for the speaker verification task.

To address the above problems, outputs from hidden or output layers of neural network are usually regarded as high-level features and used in speaker verification. These features are not extracted using signal processing algorithms. Instead, raw features are spanned in a context window (e.g. 11 frames, 5 frames on each side) and fed into a neural network to generate these features. The context span effectively improves the duration problem and neural network supervision may also lead to more task-specific features. Hence, neural network features are usually regarded as better representation than raw features for specific tasks. For example, in speaker recognition, *bottleneck features* have been used to build a GMM-UBM system [15]. In this approach, neural network with a bottleneck layer in the middle is trained, and speaker labels are used as the outputs. Once the network is trained, the outputs from the bottleneck layer are utilized as features to build GMM-UBM speaker verification systems. Although individual bottleneck-feature-based systems are slightly worse than the spectral-feature-based systems, when performing score combination, the system combining the bottleneck and the raw-features yields improved performance. Similar idea was enhanced in [21] and showed slight gains.

Although there has been some works on using neural network features for speaker verification, the reported improvements are still slight and most previous works are based on shallow networks. More importantly, depending on the type of neural network supervision labels, there can be a variety of feature choices. There has not been a detailed investigation on this. In this paper, three types of deep features are proposed to improve the performance of text-dependent speaker verification systems. A detailed description of various neural network choices and feature combinations are given in this section.

3.1. Three Types of Deep Features

As indicated in the previous section, deep neural network can be either supervised and unsupervised. In supervised neural networks, the supervision label can also vary, depending on the classification tasks. In the context of speaker verification, there are three main types:

- **RBM - Unsupervised Feature**

  Here, the neural network parameters are optimized in an unsupervised fashion with the approximate contrastive divergence algorithm. Once the RBM is trained, the original spectral feature is fed into it and the outputs of a particular hidden layer are extracted. The principal component analysis (PCA) is then applied to orthogonalize the outputs and only the most important components, which account for the 95% of the total variance, are retained. Considering that no label information is used in RBM, all speech characteristics may be represented in the RBM feature, including phone-level, speaker-level and channel-level characteristics.

- **Phone-discriminant DNN - Supervised Feature**

  Here, a DNN is trained in supervised mode using the triphone states labels as targets. The RBM pretraining [22] is used to initialize the neural network. Then deep features are extracted in a similar way as for RBM. Triphone states are closely linked to text information and have been widely accepted for speech recognition. Considering that the task in this paper is *text-dependent* speaker verification, using phone-discriminant DNN is believed to be useful.

- **Speaker-discriminant DNN - Supervised Feature**

  This is a natural choice for *speaker verification* as 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. Although it has been used in some previous work [15], it has not shown effectiveness before. As shown later in the experiment section, in this paper, speaker-discriminant DNN can actually yield significant improvement.

3.2. Tandem Deep Feature and Feature Combination

Deep features described above are pure neural network features. As shown in speech recognition, combining neural network features with the original spectral features in a Tandem fashion can yield additional gains. Hence, in this work, Tandem deep features are constructed as shown in Figure 1.

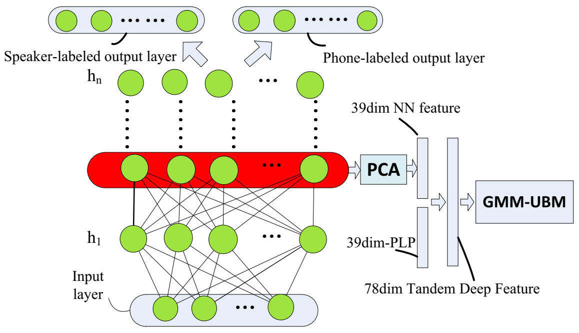
Speaker-labeled output layer
Phone-labeled output layer
h_n
39dim NN feature
PCA
GMM-UBM
h_1
39dim-PLP
Input layer
78dim Tandem Deep Feature

Figure 1: The framework of speaker verification with deep features

In addition to the Tandem combination shown in Figure 1, different types of deep features can also be combined to form new features. This is done by simply concatenating neural network features from different types of networks to obtain multi-deep features combined version. As shown in the below experiments, the combined deep features can indeed obtain additive gains.

4. Experiments

4.1. Experimental Setup and Baseline Sysmtem

To evaluate and demonstrate the effectiveness of the proposed feature extractor approaches in text-dependent speaker verification, the RSR2015 data set [8], which is released by I2R, is used. The corpus consists of 300 speakers, and is divided into background (bkg), development (dev) and evaluation (eval) subsets. We used the part I of the dataset, where each of the speakers is speaking 30 different phrases, into 9 different sessions recorded with three portable devices. When 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 taken from the eval set.

The baseline system is constructed using the gender-independent GMM-UBM approach. A 39-dimensional PLP features with short-term mean and variance normalization is used as the spectral features in our experiments. 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 both bkg and dev data. In test data set there are 19052 tests for true speaker and 1548956 tests for imposture, and no score normalization has been applied. The Equal Error Rate (EER) of the baseline system is shown in Table 1 and it is comparable to other researchers' work such as [23]. We can see that the EER is relatively very low compared to the text-independent tasks, which is relative hard to improve.

4.2. Evaluation of proposed Tandem Deep Features

To evaluate the proposed three tandem deep feature extractors, different neural networks are trained firstly, including RBM, phone-discriminant and speaker-discriminant DNN. All the deep models have 7 hidden layers with 1024 nodes per layer, and a context window of 11 frames 39-dim PLP was used as the NN inputs. The bkg and dev data are used in the NN training. The state alignment for the phone DNN training was generated using a GMM model with 3001 tied-triphone-states, which is built on a 50-hour SWB English task (ref to our previous work in [24] for more description about this system), and 194 classes (194 speakers in bkg and dev set) are used in the speaker DNN training. The contrastive divergence algorithm is used in the RBM training and SGD based back-propagation is applied to training the DNN. The learning rate annealing and early stopping strategies as in [19] was used in the BP process and the DNN was finetuned with cross-entropy objective function, along with a L2-norm weight-dacay term of coefficient 10^{-6}.

When finishing the model training, the deep models are utilized to transform the original spectral features into new tandem deep features. For each speech frame, the principal component analysis (PCA) is applied on the outputs of hidden layers and reduce the dimension to 39 dims as the original PLP feature. This can just be treated as the deep feature for following modeling, or be connected with the original PLP to form the new concatenated Tandem deep feature. The GMM-UBM framework is implemented in all the experiments.

4.2.1. Evaluation of Individual Deep Features

In this section, the proposed three deep features are firstly investigated individually. For detailed investigation, the comparison of deep feature extraction based on different hidden layers of the RBM or the DNN is performed. Besides the experiments those both using the single deep neural network features and the concatenated tandem deep features are implemented. The system performance are shown in Table 1.

From Table 1, it is observed that most of the neural network based deep features get better performance than the PLP-based baseline. Regarding the RBM-based deep features, the feature from the middle layer obtains the best EER, and the relative lower layer (the 2nd layer here) achieves the best position when using the phone or speaker based DNNs. Although the single deep feature can also obtain some obvious improvement, the concatenated mode with the PLP get a large EER re-

Table 1: Performance EER (%) of Individual Deep Features

Layer Index	RBM		Phone-DNN		Speaker-DNN	
	Deep Feature	+ PLP	Deep Feature	+ PLP	Deep Feature	+ PLP
Baseline	**1.50**					
2nd-layer	1.25	0.99	**1.45**	**0.89**	**1.08**	**0.80**
4th-layer	**1.23**	**0.94**	1.86	1.04	1.48	0.97
7th-layer	1.46	1.06	1.94	1.15	2.15	0.96

duction in all three types of tandem deep features. Moreover the supervised DNNs are superior than the unsupervised RBM method, and it confirms the previous speculation that the phone-discriminant DNN retains much more information about the text which is especially useful in this text-dependent task, and more speaker-dependent knowledge can be enhanced in the speaker-discriminant DNN, which makes the features more speaker discriminatively. All of the individual best systems in each tandem deep feature get more than 30% relative EER reduction.

4.2.2. Evaluation of Different Deep Features Combination

The next step is to combine the different deep features to get more improved performance. Based on the results shown in Table 1, we selected the relative best system for individual deep feature respectively, including the 4th-layer RBM deep feature and both the 2nd-layer phone or speaker discriminant DNN deep feature. The different deep features tried to be concatenated to form new tandem deep feature (PLP is always connected). After this, the GMM-UBM modeling is performed as before. These systems not only comprises complementarity of different unit-based DNNs, phone vs. speaker, but also combines different criteria training strategies, including the unsupervised strategy and supervised strategy.

The results of deep feature combination are shown in Table 2. It can be observed that compared to the individual deep feature system in Table 1, the multi-deep feature combination obtain additive improvement. The best tandem deep feature obtain approaching another 10% relative EER reduction when compared to the best individual system. The DET curves in Figure 2 shows a performance comparison of all the proposed methods investigated in this paper. Compared to the traditional system, the proposed deep features show substantial improvements. Moreover the gains from the individual strategy are additive, and the best multi-deep features combination approach achieves the best overall performance, and the improvement is even more than 50% relative EER reduction over the traditional PLP feature in the GMM-UBM framework. To our best knowledge, this result is so far the best reported performance on the same RSR2015 corpus under the GMM-UBM framework [23].

Table 2: Performance of Different Deep Features Combination

PLP	RBM	Phone-DNN	Speaker-DNN	EER (%)
√	—	—	—	1.50
√	√	√	—	0.80
	√	—	√	**0.73**
	—	√	√	0.82
	√	√	√	0.74

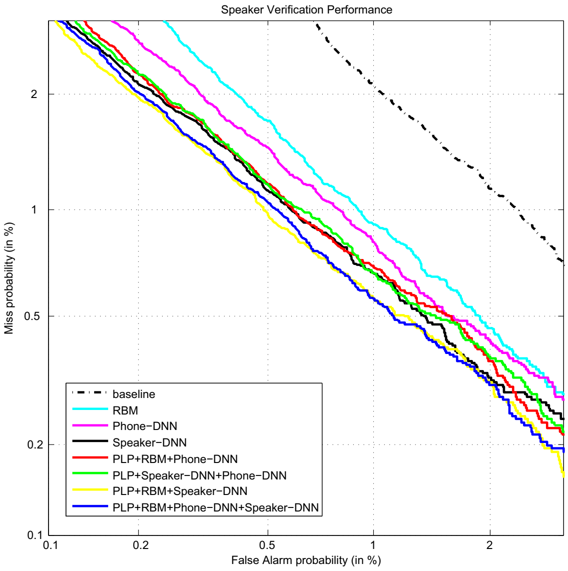

Figure 2: The DET comparison of different deep features

5. Conclusions and Future Work

This paper presents the detailed researches on using various types of deep features for text-dependent speaker verification. Three types of deep feature extractors are proposed, including restricted Boltzmann machine (RBM), phone discriminant and speaker discriminant deep neural network (DNN). The outputs of the hidden layers are extracted and concatenated with the original acoustic features to be used in a GMM-UBM framework, and large improvements are obtained in all Tandem deep feature systems. Particularly the phone discriminant and speaker discriminant DNN, which enhanced the useful text-dependent and speaker-dependent knowledge, are more effective for the text-dependent speaker verification applications. Considering the potential complementarity among the different deep features, the multi-deep features combination is implemented to get more refined feature representation. Experiments show that the final best combined feature achieved more than 50% relative EER reduction over the traditional PLP system. To our best knowledge, this result is so far the best reported performance on the same RSR2015 corpus under the GMM-UBM framework.

In the future we hope to combine these feature-level ideas with other model-level approaches, such as JFA [6] or iVector [7], to get a more improved system.

6. References

[1] Ahsanu Kabir and Sheikh Mohammad Masudul Ahsan, "Vector Quantization in Text Dependent Automatic Speaker Recognition using Mel-frequency Cepstrum Coefficient," in *International Conference on circuits, systems, electronics, control signal processing*, 2007.

[2] Douglas A. Reynolds, "Speaker Identification and Verification using Gaussian Mixture Speaker Models," *Speech communication*, vol. 17, no. 1, pp. 91–108, 1995.

[3] William M. Campbell, Douglas E. Sturim, and Douglas A. Reynolds, "Support Vector Machines using GMM Supervectors for Speaker Verification," *Signal Processing Letters, IEEE*, vol. 13, no. 5, pp. 308–311, 2006.

[4] Kevin R. Farrell, Richard J. Mammone, and Khaled T. Assaleh, "Speaker Recognition using neural networks and conventional classifiers," *Speech and Audio Processing, IEEE Transactions on*, vol. 2, no. 1, pp. 194–205, 1994.

[5] Douglas A.Reynolds, Thomas F.Quatieri, and Robert B.Dunn, "Speaker Verification using adapted Gaussian mixture models," *Digital signal processing*, vol. 10, no. 1, pp. 19–41, 2000.

[6] Patrick Kenny, Gilles Boulianne, Pierre Ouellet, and Pierre Dumouchel, "Joint Factor Analysis versus eigenchannels in Speaker Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 15, no. 4, pp. 1435–1447, 2007.

[7] Najim Dehak, Patrick Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end Factor Analysis for Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, pp. 788–798, 2011.

[8] Authony Larcher, Kong Aik Lee, Bin Ma, and Haizhou Li, "The RSR2015: Database for Text-dependent Speaker Verification using Multiple Pass-Phrase," in *Proc. Interspeech*, 2012.

[9] Hagai Aronowitz, "Text dependent speaker verification using a small development set," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.

[10] Anthony Larcher, P Bousquet, Kong Aik Lee, Driss Matrouf, Haizhou Li, and J-F Bonastre, "I-vectors in the context of phonetically-constrained short utterances for speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*. IEEE, 2012, pp. 4773–4776.

[11] Larcher Anthony, Lee Kong Aik, Ma Bin, and Li Haizhou, "Phonetically-constrained plda modeling for text-dependent speaker verification with multiple short utterances," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7673–7677.

[12] Kishore S.P, Yegnanarayana B, and Gangashetty S.V, "Online Text-Independent Speaker Verification System using Auto-Associative Neural Network Models," in *Proc. IJCNN*, 2001.

[13] Gupta C.S, Prasanna S, and Yegnanarayana B, "Auto-Associative Neural Network Models for Online Speaker Verification using Source Features from vowels," in *Proc. IJCNN*, 2002.

[14] G.S.V.S. Sivaram, Samuel Thomas, and Hynek Hermansky, "Mixture of Auto-Associative Neural Network for Speaker Verification," in *Proc. Interspeech*, 2011.

[15] Yochai Konig, Larry Heck, Mitch Weintraub, and Kemal Sonmez, "Nonlinear Discriminant Feature Extraction for robust Text-independent Speaker Recognition," in *Proc. RLA2C, ESCA workshop on Speaker Recognition and its Commercial and Forensic Applications*, 1998.

[16] Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N. Sainath, and et al, "Deep Neural Networks for acoustic modeling in Speech Recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, vol. 29, no. 6, pp. 82–97, 2012.

[17] Vasilakakis Vasileios, Cumani Sandro, and Laface Pietro, "Speaker Recognition by means of Deep Belief Networks," in *BBfor2*, 2013.

[18] Yuan Liu, Tianfan Fu, Yuchen Fan, Yanmin Qian, and Kai Yu, "Speaker Verification with Deep Features," in *Proc. IJCNN*, 2014.

[19] George E. Dahl, Dong Yu, Li Deng, and Alex Acero, "Context-dependent pre-trained Deep Neural Networks for Large-vocabulary Speech Recognition," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 20, no. 1, pp. 30–42, 2012.

[20] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, *Learning Representations by Back-propagating Errors*, MIT Press, Cambridge, MA, USA, 1988.

[21] Larry P. Heck, Yochai Konig, M Kemal Sönmez, and Mitch Weintraub, "Robustness to telephone handset distortion in speaker recognition by discriminative feature design," *Speech Communication*, vol. 31, no. 2, pp. 181–192, 2000.

[22] Geoffrey E. Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for Deep Belief Nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[23] T. Stafylakis, P. Kenny, P. Ouellet, J. Perez, M. Kockmann, and P. Dumouchel, "Text-dependent Speaker Recognition using PLDA with uncertainty Propagation," *matrix*, vol. 500, pp. 1, 2013.

[24] Tianxing He, Yuchen Fan, Yanmin Qian, Tian Tan, and Kai Yu, "Reshaping Deep Neural Network for fast decoding by Node-pruning," in *Proc. ICASSP*, 2014.