INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Improved Gender Independent Speaker Recognition Using Convolutional Neural Network Based Bottleneck Features

Shivesh Ranjan, John H. L. Hansen

Center for Robust Speech Systems (CRSS)
The University of Texas at Dallas, Richardson, TX, USA
{Shivesh.Ranjan, John.Hansen}@utdallas.edu

Abstract

This paper proposes a novel framework to improve performance of gender independent i-Vector PLDA based speaker recognition using convolutional neural network (CNN). Convolutional layers of a CNN offer robustness to variations in input features including those due to gender. A CNN is trained for ASR with a linear bottleneck layer. Bottleneck features extracted using the CNN are then used to train a gender-independent UBM to obtain frame posteriors for training an i-Vector extractor matrix. To preserve speaker specific information, a hybrid approach to training the i-Vector extractor matrix using MFCC features with corresponding frame posteriors derived from bottleneck features is proposed. On the NIST SRE10 C5 condition pooled trials, our approach reduces the EER and minDCF 2010 by +14.62% and +14.42% respectively compared to a standard mfcc based gender-independent speaker recognition system.

Index Terms: gender independent, i-vector, speaker recognition, convolutional neural network.

1. Introduction

State-of-the-art text independent speaker identification (SID) systems employ i-Vectors with Probabilistic Linear Discriminant Analysis (PLDA) back-ends [1, 2, 3, 4, 5]. I-Vectors offer a compact representation of speaker-specific attributes of an utterance [1]. I-Vectors have also found widespread use in language identification [6, 7, 8]. More recently, Deep Neural Network (DNN) based approaches have found usage in SID systems either to compute the posteriors used in i-Vector extractor matrix training/i-Vector extraction, or to generate bottleneck features that can be used in a GMM-UBM based i-Vector PLDA based SID approach [9, 10, 11, 12, 13].

Being an important speaker-specific attribute, gender information is well preserved in the i-Vectors [14]. Most state-of-the-art SID systems use a gender dependent approach where either part, or all of the SID pipeline can be gender dependent to obtain more competitive results. A mixture of gender-dependent PLDA models was proposed for speaker verification in [15]. In [12], gender-dependent approach was used to train the UBM, i-Vector extractor, and PLDA back-end. The approach involved extracting stacked bottleneck features (SBN) from a second DNN trained for ASR with bottleneck features extracted from the first DNN. For generating the monolingual SBN features, the system was trained with about 250 hours of Fisher English Part 1 data. In [13], a gender-independent approach was considered to train the UBM, i-Vector extractor matrix, and both gender-dependent and gender-independent PLDA back-ends. The approach used posteriors obtained from a Time Delay Neural Network (TDNN) trained on 1800 hours of Fisher English. In [11], a unified framework for speaker and language recognition was presented where a DNN with linear bottleneck layer was trained using around 100 hours of Switchboard data. The bottleneck features were then used to train a standard i-Vector PLDA based SID framework using a GMM-UBM.

Recently, Convolutional Neural Networks (CNN) have become state-of-the-art in ASR [16, 17, 18]. The convolutional layers of a CNN can offset minor distortions present in input features, and have been widely used in computer vision [19, 20], and were later adopted for ASR. CNNs have been used for SID in [21], where a CNN trained for ASR was used to extract posteriors for training an i-vector PLDA system on the noisy and severely degraded DARPA RATS data [22]. The frame posteriors obtained using CNN were used with MFCC features to train an i-Vector PLDA based SID system.

We propose to use bottleneck features extracted from a CNN trained for ASR in a hybrid fashion to improve the performance of gender-independent SID systems. The input convolutional layer of the CNN uses 1-D convolution along the frequency axis. Bottleneck features extracted with the CNN are used to train a gender-independent UBM to obtain frame posteriors of the training data. Next, using MFCC features and the frame posteriors obtained using CNN based bottleneck features, an i-Vector extractor is estimated and used subsequently in a gender-independent i-Vector PLDA back-end.

To examine the impact of available training data on the proposed approach, two sets of data for training the SID systems are used: a limited data-set of randomly chosen 16,000 utterances from SRE training data, and a full dataset comprising of SRE , Switchboard and Fisher. CNNs corresponding to limited and full data systems are trained with 110hrs and 300 hours of Switchboard respectively. Results in terms of equal error rate (EER) and minDCF 2010 are reported for the NIST SRE 2010 C5 extended condition. For both limited and full data systems, our approach is shown to significantly outperform an MFCC based gender-independent i-Vector PLDA SID system.

2. i-Vector PLDA based Speaker Identification

2.1. i-Vector Extraction

An i-Vector is a fixed low dimensional representation of a speech utterance that preserves the speaker-specific information. In the i-Vector paradigm, a speaker-specific GMM mean

This project was funded by AFRL under contract FA8750-15-1-0205 and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J.H.L. Hansen.

http://dx.doi.org/10.21437/Interspeech.2017-1182

supervector M can be represented in terms of speaker and channel independent supervector m, a low rank *total variability matrix* T, and a vector w as

$$M = m + Tw. \quad (1)$$

In (1), w is a random vector with a standard normal distribution $N(0, I)$. The T matrix is learned using large amounts of (domain matched, when available) training data. The i-Vector of an utterance are its coordinates in the *total variability space* (i.e. space spanned by the columns of T), extracted as the maximum a posteriori (MAP) point estimates of w given the utterance [1, 23].

2.2. PLDA back-end for Speaker Identification

Using i-Vectors extracted from a large labeled-training set, a PLDA model is used to learn the within-class and across-class variabilities using an Expectation Maximization (EM) algorithm [2]. Specifically, in this study, a Gaussian PLDA (G-PLDA) of the form described in [4] is used. Assuming R training utterances of a speaker, the entire collection of i-Vectors may be expressed as $\{\eta_r : r = 1, 2, ..., R\}$. In the G-PLDA formulation , an i-Vector of this collection can be expressed as,

$$\eta_r = m + \Phi\beta + \epsilon_r. \quad (2)$$

In (2), m is a global offset, columns of Φ constitute a basis for speaker subspace , β corresponds to the coordinates in the speaker subspace, and ϵ_r is a Gaussian with zero mean and covariance Σ. The G-PLDA model parameters $\{m, \Phi, \Sigma\}$ are estimated using an EM algorithm on a large collection of speaker-labeled i-Vectors. Given a test utterance, the verification score can be computed using a closed form solution with the G-PLDA model as presented in [4]. A single speaker-identification trial thus requires access to the mean of the speaker's enrollment i-Vectors, the test i-Vector, and the G-PLDA model parameters $\{m, \Phi, \Sigma\}$.

3. CNN based Bottleneck Features for SID

3.1. Bottleneck Feature Generation

CNNs have found widespread use in ASR [16, 17, 18, 24]. In essence, the strength of CNN comes from *convolutional* and *max-pooling* layers which allow it to offer invariance to minor shifts along the frequency axis [24]. Additionally, CNNs may use *weight-sharing* at some level which keeps the number of trainable parameters tractable [24].

For ASR applications, a few convolutional layers followed by fully connected layers are used, and the output is a softmax layer corresponding to the target senones. The network can be trained using cross-entropy objective function with stochastic gradient descent (SGD). Fig. 1 shows the architecture of the network used for generating CNN based bottleneck features. A CNN is trained for ASR using the alignments obtained using an HMM-GMM based ASR system. The first hidden layer is a convolutional layer followed by max pooling, followed by another convolutional layer with no max pooling. This follows a standard CNN based ASR recipe in Kaldi (based on Karel's Nnet1) using 1-D convolution along the frequency axis [25].

The bottleneck layer is introduced as a linear layer at the second last hidden layer. Linear bottleneck layer has been reported to be suitable for several applications such as: ASR, SID, language identification [26, 11], and is used in the present work. After the network is trained for ASR, all layers following the bottleneck layer are removed from the network for feature extraction.

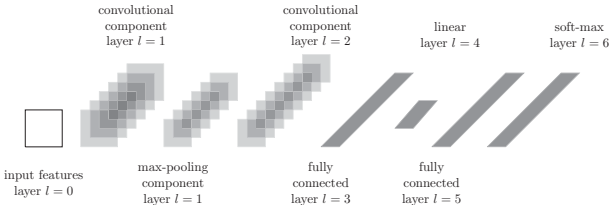

Figure 1: *Architecture of the proposed CNN for ASR trained with a linear bottleneck layer. Max-pooling is used only at the first convolutional layer. Only 2 fully connected layers are shown here.*

3.2. CNN based Bottleneck Features for SID

Here, we propose to use CNN based bottleneck features (CNN BNF) for training a gender-independent UBM. CNN BNF features have been previously used for ASR [26], and LID [27]. Using a CNN BNF features trained UBM m_{BNF} provides better phonetic alignments that are more invariant to speaker gender. However, we believe the CNN can also make the bottleneck features invariant to speaker identity, hence, these features have less speaker-specific information compared to standard features for SID such as MFCC. This is motivated by earlier observations regarding loss of speaker-specific information in DNN based bottleneck features for SID [12].

To overcome the loss of speaker-specific information in CNN BNF features for SID, we propose a hybrid approach by using them together with MFCC features. In this, the UBM trained with bottleneck features is used only to obtain the alignments of the training data to be used in training the i-Vector extractor. Once the posteriors are obtained for the training data, we use these posteriors with MFCC features to train the i-Vector extractor. Similar hybrid approaches have been used in DNN based SID techniques, where the DNN trained using ASR features is used to obtain posteriors together with MFCC features to train the i-Vector extractor matrix [9, 10].

Frame posteriors obtained using a CNN BNF feature UBM m_{BNF} are also used to initialize an MFCC UBM m_{mfcc}. Estimation of the MFCC UBM is similar to obtaining an ancillary UBM as done in [10, 13, 28]. The MFCC UBM m_{mfcc} is then used to initialize the i-Vector extractor matrix ie_{init}, and also during i-Vector extraction for computing frame posteriors. Thus, the role of CNN BNF features in our hybrid approach is limited to obtaining frame posteriors used to train the i-Vector extractor matrix, and to obtain an MFCC UBM. This is a deviation from prior studies using DNN based bottleneck features for SID, since we do not use the CNN BNF UBM to compute the posteriors for i-Vector extraction.

We hypothesize that adopting the hybrid approach of using bottleneck features only to obtain frame posteriors of the training data results in improved i-Vector extractor training, with better phonetic alignments that are less influenced by the gender of speakers. Since the MFCC UBM m_{mfcc} is constructed with the same CNN BNF posteriors, it is less impacted by the gender of the speakers as well. Nonetheless, using MFCC features for m_{mfcc} allows us to avoid pitfalls of losing speaker-specific information that can occur due to CNN BNF features. The following Alg. 1 highlights the steps of our proposed hybrid approach to gender-independent SID.

Algorithm 1: Hybrid approach to gender-independent SID using CNN based bottleneck features and MFCC features.

1. Train a UBM m_{BNF} using CNN-BNF features for the training data.
2. Obtain frame posteriors for the training data using the CNN-BNF UBM m_{BNF}.
3. Construct an ancillary UBM m_{mfcc} using MFCC features and posteriors obtained from step 2.
4. Initialize the i-Vector extractor ie_{init} matrix using the ancillary UBM m_{mfcc} from step 3.
5. Train i-Vector extractor matrix using posteriors from step 2, the i-Vector extractor ie_{init} initialized in step 4, and MFCC features.
6. For extracting i-Vectors, use ancillary UBM m_{mfcc} from step 3 to compute the posteriors.

4. Experiments, Results and Discussions

4.1. Datasets and SID Systems

For training the SID systems reported in this study, we use a standard Kaldi based SID recipe as outlined in [13]. To observe the effect of data for the proposed approach, we use two sets of data for the experiments reported in this study: a full data set, and a limited data set. Log-mel filterbank features of dimension 40 with delta,delta-delta, and a context of 5 frames each on either side were used as input to the CNN for both datasets. For both CNNs, we used the default parameters for the convolutional components from Kaldi's 1-D CNN set-up (Karel's Nnet1 recipe) without any layer-wise pre-training.

4.1.1. Limited Data SID Systems

For the limited data systems, a subset of 16,000 utterances was randomly chosen from the SRE training data, and used to train the UBM, i-Vector extractor matrix, and PLDA back-end using Kaldi toolkit. Both MFCC based baseline system, and CNN BNF based SID system were trained using the limited dataset in a gender-independent fashion. To train the baseline MFCC system, 20-dim features with delta, and delta-delta were used. For training the full covariance gender-independent UBM with 2048 components, 4 iterations of EM algorithm was used. The i-Vector extractor matrix was trained using 5 iterations of EM algorithm, and 600 dimensional i-Vectors were used.

To train the CNN for ASR used to extract bottleneck features, approximately 110 hours of Switchboard data was used. The bottleneck layer was a linear layer of dimension 60 at the 2nd last hidden layer.The CNN had 2 convolutional layers (as mentioned in Sec 3.1), 4 fully connected layers with 1024 units per layer, and the softmax layer had 2884 output nodes corresponding to the senones. After extracting CNN BNF features, the SID system was trained using the hybrid approach outlined in Sec 3. Rest of training set up was same as used for the baseline system.

4.1.2. Full Data SID Systems

For the full data baseline MFCC SID system, a full covariance gender-independent UBM with 2048 components was trained with 4 iterations of EM algorithm using SRE, Switchboard-I and Fisher with approximately 57,000 utterances. The i-Vector extractor matrix was trained using the same data. Gender independent PLDA back-end was trained with around 36,000 utterances of in-domain labeled SRE data. A corresponding system using CNN BNFs was trained using the full dataset using approach outlined in Sec 3. To train the CNN for ASR used in extracting bottleneck features, around 300 hrs of data from Switchboard was used. The CNN trained for ASR had a linear bottleneck of dimension 60 at the 2nd last hidden layer. The architecture of the full-data CNN was same as the limited data CNN, except at the output layer which now had 8907 nodes corresponding to the senones.

Table 1: *EER (%), and min DCF10 of the MFCC i-Vector PLDA based SID system vs. our proposed CNN BNF based SID approach on the NIST SRE 2010 C5 condition using limited dataset.*

Trials	EER (%)		min DCF10	
	MFCC	CNN BNF	MFCC	CNN BNF
gender-ind female	3.51	3.02	0.060	0.050
gender-ind male	3.17	2.74	0.053	0.047
pooled	3.47	2.90	0.058	0.049

4.2. Results

Table 1 shows the results for the limited data MFCC based i-Vector PLDA SID system compared against the proposed CNN BNF based SID approach on the NIST SRE 2010 C5 condition trials. As can be seen from the results, our proposed gender-independent CNN BNF based SID approach outperforms MFCC based i-Vector PLDA SID system across female, male and pooled trials in terms of EER and minDCF10 for the limited data systems. Table 2 shows the results for the full data MFCC based i-Vector PLDA SID system compared against the proposed CNN BNF based SID approach on the SRE 2010 C5 condition trials. We can observe that for the full data systems as well, our proposed CNN BNF based SID approach outperforms MFCC based SID system in terms of EER and minDCF10 for the female, male and pooled trials.

Figure 2 and Fig. 3 show the relative reduction in EER and minDCF10 respectively, obtained by our CNN-BNF based approach compared to MFCC based SID systems for both limited and full data. For the limited data systems, we observe a relative reduction in EER of +16.42% for the pooled trials for SRE 2010 C5. Similarly, the proposed approach is highly effective in reducing minDCF10 with a relative reduction of +15.83% for the pooled trials. For the full data systems, our proposed CNN BNF based approach reduces the EER by +14.62(%) , and, the minDCF10 by +14.42% relative to the baseline MFCC based SID system on the pooled trials of NIST SRE 2010 C5.

4.3. Discussion

From the results for both limited and full data systems presented, we observe that the proposed CNN BNF based SID approach consistently outperforms MFCC based i-Vector PLDA

Table 2: *EER (%), and min DCF10 of the MFCC i-Vector PLDA based SID system vs. our proposed CNN BNF based SID approach on the NIST SRE 2010 C5 using full dataset.*

Trials	EER (%)		min DCF10	
	MFCC	CNN BNF	MFCC	CNN BNF
gender-ind female	2.43	2.34	0.050	0.040
gender-ind male	2.42	2.04	0.041	0.039
pooled	2.53	2.16	0.047	0.040

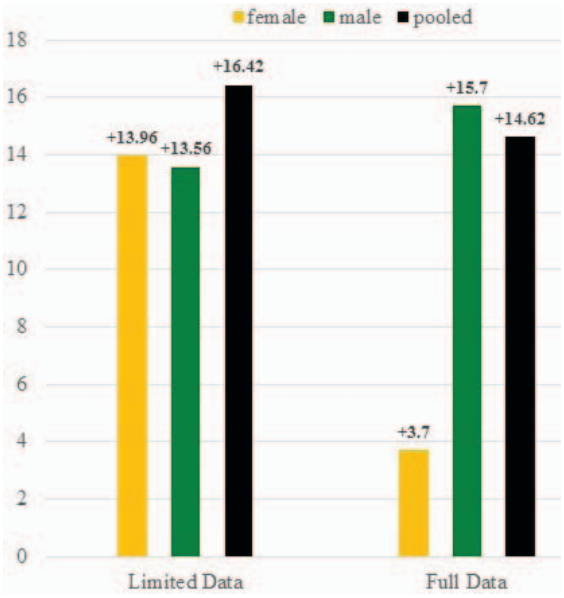

Figure 2: *Rel. reduction in EER (%) of our proposed CNN BNF based SID approach against corresponding MFCC baseline SID system on female, male, and pooled trials of the NIST SRE 2010 C5 using limited and full dataset.*

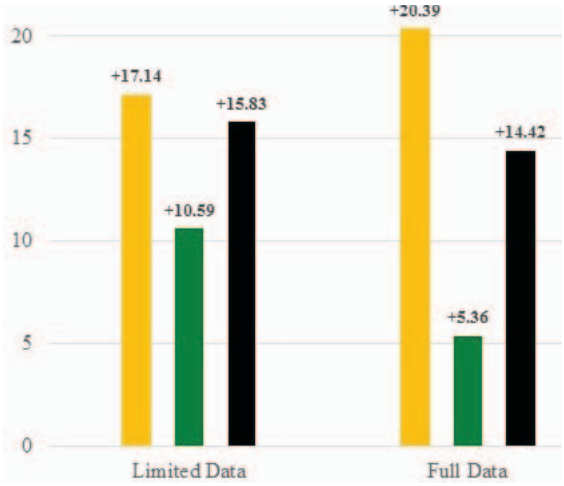

Figure 3: *Rel. reduction in minDCF10 (%) of our proposed CNN BNF based SID approach compared against corresponding MFCC baseline SID system on female, male, and pooled trials of the NIST SRE 2010 C5 using limited and full dataset.*

SID system. The improvements in SID performance can be attributed to better estimation of the i-Vector extractor matrix using frame posteriors obtained using CNN BNF features.

Comparing the results of limited and full data systems, it can also be noted that the relative improvements (measured by corresponding reduction in EER and minDCF10) of gender-specific trials can vary due to any variation in data-set size when using the proposed CNN BNF based SID appraoch. However, our approach obtains almost similar and significant improvements in terms of EER and minDCF10 on the pooled trials for both limited and full data systems. We found a degradation in SID performance by switching to frame posteriors obtained with the CNN BNF UBM m_{BNF} during i-Vector extraction, confirming our original hypothesis that the CNN BNF features attenuate speaker-specific information. Based on these results, the proposed CNN BNF based approach gives consistent, and significant improvements in both EER and minDCF10 for pooled trials across both limited and full datasets. Fig. 4 shows the Detection Error Tradeoff (DET) plots for our proposed CNN BNF based SID approach compared against baseline MFCC based SID systems. Clearly, for both limited and full data systems, our proposed CNN BNF based gender-independent SID approach outperforms corresponding MFCC based baseline SID systems.

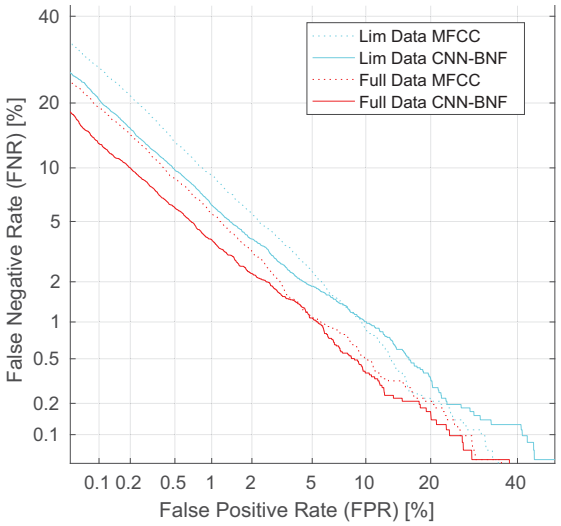

Figure 4: *DET plots for full and limited data baseline and corresponding CNN BNF SID systems for the pooled trials of the NIST SRE 2010 C5.*

5. Conclusions

In this study, we presented a novel approach to improving gender independent speaker recognition by employing bottleneck features extracted using a CNN trained for ASR. The key idea proposed here is to train the i-Vector extractor matrix using MFCC features with corresponding frame posteriors using CNN based bottleneck features. We hypothesized that posteriors obtained using CNN based bottleneck features are more invariant to speaker-gender, and using them to train the i-Vector extractor matrix improves performance of gender-independent i-Vector PLDA based SID systems.

The merits of our proposed approach was demonstrated on trials of the NIST SRE10 C5 condition, where it consistently outperformed an MFCC based i-Vector PLDA SID system. For limited data systems, our approach reduced the EER and minDCF10 by +16.42% and +15.83% respectively. We obtained a reduction of +14.62% in EER, and +14.42% in minDCF10 for full data systems. Thus, the proposed technique was found to give similar gains for pooled trials when using either limited or full data for training the speaker recognition systems. Future work will explore using CNN bottleneck features for speaker recognition in noisy conditions and speaker-independent speech recognition.

6. Acknowledgments

The authors would like to thank David Snyder and Daniel Povey from Johns Hopkins University for helpful feedback on the implementation aspects of this project through Kaldi help group.

7. References

[1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2011.

[2] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," *IEEE ICCV-2007*, pp. 1–8, 2007.

[3] P. Kenny, "Bayesian speaker verification with Heavy-Tailed priors." *Odyssey*, 2010.

[4] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-Vector length normalization in speaker recognition systems." *ISCA Interspeech*, pp. 249–252, 2011.

[5] J. H. L. Hansen and T. Hasan, "Speaker recognition by machines and humans: A tutorial review," *IEEE Signal processing magazine*, vol. 32, no. 6, pp. 74–99, 2015.

[6] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language recognition via i-Vectors and dimensionality reduction." *ISCA INTERSPEECH*, pp. 857–860, 2011.

[7] D. Martınez, O. Plchot, L. Burget, O. Glembek, and P. Matejka, "Language recognition in ivectors space," *Proceedings of Interspeech, Firenze, Italy*, pp. 861–864, 2011.

[8] S. Ranjan, C. Yu, C. Zhang, F. Kelly, and J. H. L. Hansen, "Language recognition using deep neural networks with very limited training data," *IEEE ICASSP 2016*, 2016.

[9] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," *IEEE ICASSP 2014*, pp. 1695–1699, 2014.

[10] P. Kenny, V. Gupta, T. Stafylakis, P. Ouellet, and J. Alam, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," *Proc. Odyssey*, pp. 293–298, 2014.

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

[12] P. Matějka, O. Glembek, O. Novotnỳ, O. Plchot, F. Grézl, L. Burget, and J. H. Cernockỳ, "Analysis of DNN approaches to speaker identification," *IEEE ICASSP 2016*, pp. 5100–5104, 2016.

[13] D. Snyder, D. Garcia-Romero, and D. Povey, "Time delay deep neural network-based universal background models for speaker recognition," *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pp. 92–97, 2015.

[14] S. Ranjan, G. Liu, and J. H. L. Hansen, "An i-vector PLDA based gender identification approach for severely distorted and multilingual DARPA RATS data," *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, pp. 331–337, 2015.

[15] M. Senoussaoui, P. Kenny, N. Brümmer, E. De Villiers, and P. Dumouchel, "Mixture of PLDA models in i-vector space for gender-independent speaker recognition." *ISCA Interspeech*, 2011.

[16] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, and G. Penn, "Applying convolutional neural networks concepts to hybrid NN-HMM model for speech recognition," *IEEE ICASSP 2012*, pp. 4277–4280, 2012.

[17] T. N. Sainath, A.-R. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for LVCSR," *IEEE ICASSP 2013*, pp. 8614–8618, 2013.

[18] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[19] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[21] M. McLaren, Y. Lei, N. Scheffer, and L. Ferrer, "Application of convolutional neural networks to speaker recognition in noisy conditions." *ISCA INTERSPEECH*, pp. 686–690, 2014.

[22] K. Walker and S. Strassel, "The RATS radio traffic collection system," *Odyssey*, 2012.

[23] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE Trans. Speech and Audio Processing*, vol. 13, no. 3, pp. 345–354, 2005.

[24] O. Abdel-Hamid, A.-r. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on audio, speech, and language processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[25] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The kaldi speech recognition toolkit," *IEEE 2011 workshop on automatic speech recognition and understanding*, 2011.

[26] K. Veselỳ, M. Karafiát, and F. Grézl, "Convolutive bottleneck network features for LVCSR," *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*, pp. 42–47, 2011.

[27] S. Ganapathy, K. J. Han, S. Thomas, M. K. Omar, M. Van Segbroeck, and S. S. Narayanan, "Robust language identification using convolutional neural network features." *ISCA INTERSPEECH*, 2014.

[28] D. Garcia-Romero, X. Zhang, A. McCree, and D. Povey, "Improving speaker recognition performance in the domain adaptation challenge using deep neural networks," *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pp. 378–383, 2014.