



# Investigation of Bottleneck Features and Multilingual Deep Neural Networks for Speaker Verification

Yao Tian, Meng Cai, Liang He, Jia Liu

National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China

{tianyao11, cai-m10}@mails.tsinghua.edu.cn, {heliang, liuj}@mail.tsinghua.edu.cn

## Abstract

Recently, the integration of deep neural networks (DNNs) with i-vector systems is proved to be effective for speaker verification. This method uses the DNN with senone outputs to produce frame alignments for sufficient statistics extraction. However, two types of data mismatch may degrade the performance of the DNN-based speaker verification systems. First, the DNN requires transcribed training data, while the data sets used for i-vector training and extraction are mostly untranscribed. Second, the language of the training data for DNN is limited by the pronunciation lexicon, making the model unsuitable for multilingual tasks. In this paper, we propose to use bottleneck features and multilingual DNNs to narrow the gap caused by the data mismatch. In our method, a DNN is first trained with senone labels to extract bottleneck features. Then a Gaussian mixture model (GMM) is trained with the bottleneck features to produce frame alignments. Additionally, bottleneck features based on multilingual DNNs are explored for multilingual speaker verification. Experiments on the NIST SRE 2008 female short2-short3 telephone task (multilingual) and the NIST SRE 2010 female core-extended telephone task (English) demonstrate the effectiveness of the proposed method.

**Index Terms:** speaker verification, i-vectors, deep neural networks, bottleneck features

## 1. Introduction

The performance of speaker verification gains significant improvements due to the recently proposed i-vector/Probabilistic Linear Discriminant Analysis (PLDA) framework [1, 2, 3, 4]. In this model, acoustic features (e.g., mel-frequency cepstral coefficients (MFCC), perception linear prediction features (PLP)) are first converted into high-dimensional sufficient statistics (SS) using the occupancy posterior probabilities generated by a Gaussian Mixture Model (GMM) known as Universal Background Model (UBM). Then, these statistics are mapped into a low-dimensional space and each utterance is represented as a fixed-length vector called i-vector in this subspace. Finally, PLDA is used to give verification scores between i-vectors.

In recent years, DNNs have been successfully applied in speech recognition as a replacement of GMM and have brought significant performance improvements [5, 6]. The advantages of DNNs include being able to handle longer segments as inputs and its superior learning ability derived from the multi-linear-layer structure. The speaker verification community has also done plenty of researches on DNNs [7, 8, 9, 10] but most of their work did not perform very well since a direct transition of DNNs to speaker verification is much more challenging where the target speakers are task-varying and each speaker

usually has very little training data. Recently, Lei, et.al [11] and Kenny, et.al [12] proposed a method which incorporates the DNN used in speech recognition with the i-vector model and shows promising results for speaker verification. It replaces the GMM with a DNN to compute frame posterior probabilities with respect to each class during the extraction of SS. While in the case of the GMM, each class corresponds to an individual Gaussian and has no inherent meaning since the model is trained in an unsupervised way, in the case of the DNN, the transcribed data is taken into consideration during the supervised model training and each class corresponds to certain content (tied triphone states). As a result, the DNN aligns speech frames to a sub-phonetic categories and a comparison is able to be made between the same phonetic content.

Generally, speech recognition and speaker recognition are two individual tasks and the data sets of speech recognition (source data) and speaker recognition (target data) might be different in many aspects such as channel, gender, language and *etc.* As a result, the DNN model trained with the source data might not reflect the target phonetic space so well. Finding a way to effectively narrow the gap caused by data mismatch is thus a meaningful issue. In this paper, we investigate using bottleneck (BN) features and multilingual DNNs for speaker verification. In our method, the DNN trained with source data is used to extract BN features. Then a GMM is trained in the traditional unsupervised way with BN features of target data. This GMM is used to calculate frame posterior probabilities instead of using DNN directly while collecting the SS. The BN features derived from the DNN can reflect the property of training labels which correspond to senones (tied triphone states) in our work, thus containing rich phonetic information. Consequently, the GMM trained with BN features of target data can depict the target phonetic space more accurately. In addition, the phonetic information contained in BN features is limited by the pronunciation lexicon, making it unsuitable for multilingual speaker verification tasks. So DNNs trained with multilingual data sets (English and Mandarin) are also explored in this paper in order to get better generalization for multilingual tasks. Experiments on the NIST SRE 2008 female short2-short3 telephone task (multilingual) and the NIST SRE 2010 female core-extended telephone task (English) verify the effectiveness of the proposed method and consistency improvements can be obtained over corresponding DNN based approaches.

The remainder of this paper is organized as follows. Section 2 reviews the DNN based i-vector framework. Section 3 presents the structure of our proposed BN features and multilingual DNNs for speaker verification. Experimental setup and results are given in Section 4. Finally, conclusions are presented in Section 5.

10.21437/Interspeech.2015-300

## 2. The DNN based i-vector framework

### 2.1. i-vector model

The traditional i-vector model is based on GMM-UBM and assumes that most relevant speaker information lives in a low-dimensional space called total variability space. Each utterance can be represented as a fixed-length vector called i-vector in this subspace. Given a speech segment  $i$ , the following SS (Baum-Welch statistics) need to be calculated first for i-vector modeling

$$N_c^{(i)} = \sum_t \gamma_{c,t}^{(i)} \quad (1)$$

$$\mathbf{F}_c^{(i)} = \sum_t \gamma_{c,t}^{(i)} \mathbf{o}_t^{(i)} \quad (2)$$

$$\mathbf{S}_c^{(i)} = \sum_t \gamma_{c,t}^{(i)} \mathbf{o}_t^{(i)} \mathbf{o}_t^{(i)T} \quad (3)$$

Where  $N_c^{(i)}$ ,  $\mathbf{F}_c^{(i)}$  and  $\mathbf{S}_c^{(i)}$  are the zero-order, first-order and second-order statistics of speech segment  $i$  corresponds to the  $c$ -th Gaussian component of UBM.  $\mathbf{o}_t^{(i)}$  is the acoustic feature of segment  $i$  at time  $t$ .  $\gamma_{c,t}^{(i)}$  is the posterior probability of the  $c$ -th Gaussian component given  $\mathbf{o}_t^{(i)}$  and is defined as

$$\gamma_{c,t}^{(k)} = \frac{\varpi_c \mathcal{N}(\mathbf{o}_t^{(k)}; \boldsymbol{\mu}_c, \boldsymbol{\Sigma}_c)}{\sum_{c'=1}^C \varpi_{c'} \mathcal{N}(\mathbf{o}_t^{(k)}; \boldsymbol{\mu}_{c'}, \boldsymbol{\Sigma}_{c'})} \quad (4)$$

Where  $\varpi_c$ ,  $\boldsymbol{\mu}_c$  and  $\boldsymbol{\Sigma}_c$  are the weight, mean vector and covariance of  $c$ -th Gaussian component respectively. Once SS are extracted, they are further whitened using the UBM's means and covariances before model training and i-vector extraction.

### 2.2. Roles of the DNN

In the DNN based i-vector framework, the DNN for speech recognition replaces GMM to provide frame posteriors with respect to each class (eq (4)) for SS extraction. In the traditional i-vector model, each mixture of the GMM represents a class. Since the GMM are trained in an unsupervised fashion, each mixture (class) has no inherent meanings, only representing a certain region of the acoustic space. In speech recognition, however, the DNN effectively leverages transcribed data and is trained aiming to discriminate between senones (classes), thus explicit relations exist between classes and phonetic contents. Consequently, using DNNs to calculate posterior probabilities can be seen as a phonetic-dependent frame alignments and content comparison is able to be made between different speakers afterwards. It is worth noting that Lei, et.al also experimented on using supervised GMM to do frame alignments where the performance are not as good as DNNs because of worse classification ability [11].

Figure 1 presents the processes of UBM based and DNN based sufficient statistics extraction. From the diagrams we can see that the features for frame alignments and statistics computation used to be the same and now are efficiently decoupled with the introduction of DNNs. As a result, we can use different features for the calculation of frame posteriors (Fbank) and statistics (MFCC is used in [11] where we use PLP instead) in order to obtain optimal performance in each case. Once the SS are extracted, an ancillary UBM is needed to pre-whiten the SS [11, 12]. The following model training and i-vectors extraction are the same as the traditional approach.

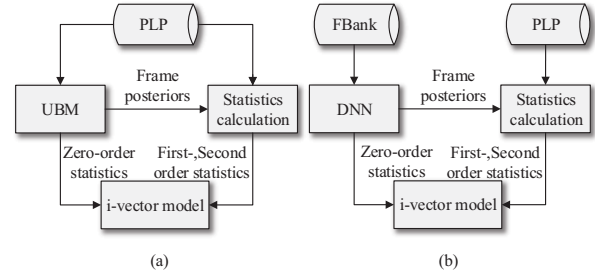


Figure 1: The flow diagrams of (a) UBM based (b) DNN based i-vector framework.

## 3. Bottleneck features and multilingual DNNs for speaker verification

In this section, we present BN features based i-vector framework. Besides, a multilingual DNN structure is proposed in order to make the feature more suitable for multilingual tasks.

### 3.1. Bottleneck features based i-vector framework

The DNN based BN features are increasingly being used to improve the performance of speech related applications [13, 14, 15]. BN here means a hidden layer placed in the middle of a DNN which has a relative small number of hidden units compared to the size of the other layers. The linear outputs of the neurons from this layer is referred to the BN feature and can be seen as a compact low-dimensional representation of the inputs which contains information pertinent to classification. BN features are most often used in autoencoders where the network is trained to reconstruct the input features [16]. While in speech recognition, the network with BN layer is trained to discriminate between senones (tied triphone states) and thus the BN feature contains rich phonetic information.

In our work, we propose to use a GMM trained with BN features to provide frame alignments in order to make the alignment model more relevant and reliable for the speaker verification data sets. Let's denote the training data sets for speech recognition and speaker verification as source data and target data respectively. In our method, we first train a DNN using source data as a feature extractor. Then a GMM is trained in the traditional unsupervised way using BN features of target data and is used to calculate frame posterior probabilities while collecting SS. Each mixture of the GMM is more connected to phonetic content due to the discriminative information contained in BN features. In addition, compared with the DNN based method, this model can depict the target phonetic space more accurately to some extent by utilizing the target data. An ancillary UBM is also needed here as in the DNN based system to pre-whiten the SS since the features for alignment and statistics computation are different.

### 3.2. Multilingual DNNs

In speech recognition, multilingual model is an efficient technique for addressing resource constraints such as data, time and processing power where data from multiple languages are put together to train a single system [17, 18]. DNNs are naturally for multilingual training by the characteristic of parameters sharing. In this paper, we utilize a DNN trained with both English and Mandarin data to extract BN features in order to improve the feature's flexibility on multilingual speaker verifi-

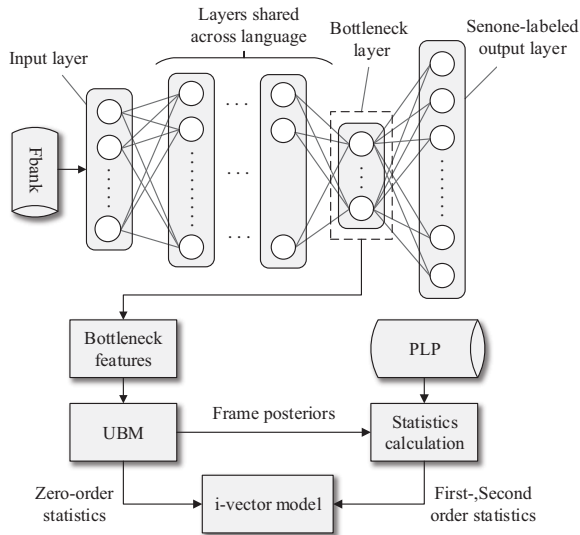


Figure 2: The structure of bottleneck feature based i-vector framework.

cation tasks. In our model, the two languages' phone sets and question sets of decision trees are put together. A GMM-HMM system is first trained on both English and Mandarin data to generate the transcription for senones. Then a DNN with bottleneck layer is trained with these senone-labeled data where all data is presented to the model at the same time and joint optimization take place on the feature extraction and the classifiers during model training. Since the hidden layers are shared across languages, the multilingual DNN provides a more sophisticated and robust basis for BN features which contain phonetic information on both English and Mandarin and would benefit multilingual speaker verification tasks. Figure 2 presents the structure of our proposed BN features based i-vector framework.

## 4. Experiments

### 4.1. Experimental setup

Experiments are carried out on the NIST SRE2010 female core-extended telephone-telephone task (Condition 5, English) and NIST SRE2008 female short2-short3 telephone-telephone task (Condition 6, Multilingual). Equal error rate (EER) and minimum decision cost function (minDCF) are selected as metrics for evaluation [19, 20].

The training data for English DNNs are 300 hours English telephone speeches from Switchboard data sets. The training data for Mandarin DNNs are 300 hours Mandarin telephone speeches from self-collected data sets. These data sets are all used when training multilingual DNNs. A GMM-HMM is first trained to generate transcriptions for senones. The GMM-HMM system uses 13-dimensional PLP features with speaker-based mean-covariance normalization. The basic features are then concatenated with their first-, second- and third-order derivatives and further reduced to 39 dimensions by HLDA. The DNN used to provide the posterior probability has five-hidden layers and is trained with cross-entropy criterion using the transcriptions from the GMM-HMM. The input layer of the DNN has 1320 nodes composed of 11 frames (5 frames on each side of the frame) where each frame consists of 120 log Mel-filterbank coefficients (40 basic + first order + second order). Each hidden

layer has 1200 nodes. The configurations of the DNN with BN layer are the same except that the number of the fifth hidden layer is changed to 39. Different number of output senones are evaluated and the details are in the results section.

The training data for UBM based systems includes data from Switchboard, NIST SRE 2004, 2005, 2006 where the Switchboard data is all English and NIST data is mostly English but contains four other languages including Arabic, Mandarin, Russian and Spanish. When training UBM, only NIST data sets are used. When training i-vector model and PLDA model, all the training data sets are used. 39-dimensional (13 basic + first order + second order) PLP is extracted as the raw acoustic feature. Then a gender-dependent diagonal covariance UBM with 2048 mixtures is trained. The dimensionality of i-vectors is set to 400. Simplified Gaussian PLDA is used to given verification scores [4] and the dimensionality of speaker subspace in PLDA model is 200.

In the DNN based systems, DNNs with senone outputs are used to provide frame alignments during SS extraction. Additionally, these frame posteriors are combined with PLP features to estimate a diagonal ancillary UBM in one pass in order to pre-whiten SS. The number of mixtures is determined by the number of senones. Other model configurations including the usage of training data and the parameter settings are the same with UBM based systems.

In the BN features based systems, DNNs with senone outputs are used to extract BN features. Then a UBM is trained using these BN features and used to provide frame posteriors during sufficient statistics extraction. A diagonal ancillary UBM is also estimated to pre-whiten the statistics. The number of mixtures is set to 2048 for all conditions. Other model configurations are the same with UBM based systems.

### 4.2. Experimental results

Table 1: The performance of UBM based, DNN based and BN based systems on the NIST SRE2010 coreext tel-tel condition.

system	EER(%)	minDCF08	minDCF10
UBM-2048	2.91	0.134	0.487
DNN-Eng-2227	2.58	0.123	0.407
DNN-Man-2178	3.24	0.149	0.491
DNN-Multi-2232	2.88	0.129	0.435
BN-Eng-2227	2.28	0.110	0.355
BN-Man-2178	3.03	0.142	0.426
BN-Multi-2232	2.61	0.121	0.412

Table 1 presents the results of UBM based (UBM-), DNN based (DNN-) and BN features based (BN-) systems on the NIST SRE 2010 female core-extended tel-tel condition. The senone numbers are 2227, 2178 and 2232 for English (Eng), Mandarin (Man) and English+Mandarin (Multi) related DNNs. From the results we can see that both DNN-Eng and DNN-Multi systems outperform the UBM approach while the DNN-Man system is slightly worse, since the first two systems are more relevant to the task and can provide more accurate alignments. In addition, the BN features based systems can provide further improvements over DNN based systems on all language conditions. The Switchboard and NIST data sets are generally considered to be consistent, but the BN-Eng system still shows improvements over DNN-Eng system due to more data relevant factors such as gender-dependent, similar recording envi-

ronment and *etc.* Compared to DNN-Eng based method, the relative improvements of BN-Eng are 11.6% in EER, 12.8% in minDCF10. Even though the BN-Multi system is inferior to BN-Eng system, it is competitive to DNN-Eng system.

Table 2: *The performance of BN based systems with different number of senones on the NIST SRE2010 female corext tel-tel condition.*

system	EER(%)	minDCF08	minDCF10
UBM-2048	2.91	0.134	0.487
BN-Eng-2227	2.28	0.110	0.355
BN-Eng-4126	2.37	0.116	0.360
BN-Eng-8223	2.42	0.118	0.363
BN-Multi-2232	2.61	0.121	0.412
BN-Multi-4155	2.62	0.122	0.412
BN-Multi-8269	2.68	0.124	0.415

An attractive benefit of BN features based systems is that the DNN is used as a feature extractor rather than a classifier which makes it feasible to utilize DNNs with larger number of senone outputs. The results of DNNs with different number of senone outputs on the NIST SRE 2010 female core-extended tel-tel condition are presented in Table 2. The senone numbers are 4126 and 8223 for English DNNs, 4155 and 8269 for English+Mandarin DNNs. Our original thoughts was that DNNs with larger senones provide finer granularity, making the BN features more flexible for the target phonetic space modeling and resulting in better performance. However, the results suggest that the performance is getting worse by enlarging the senone numbers. We thought the relative poor classification accuracy of DNNs with larger number of senones might be the reason for performance degradation where the frame classification accuracies of DNNs on the development sets are 67.2%, 63.8% and 60.3% with respect to BN-Eng-2227, BN-Eng-4126 and BN-Eng-8223, 64.8%, 61.2% and 57.1% with respect to BN-Multi-2232, BN-Multi-4155, BN-Multi-8269. Since the mixture of UBMs in BN features based systems is fixed to 2048, the BN features extracted from DNNs with around 2000 senones might be more accurate for the phonetic space.

Table 3: *The performance of UBM based, DNN based, BN based systems on the NIST SRE2008 short2-short3 tel-tel condition.*

system	EER(%)	minDCF08	minDCF10
UBM-2048	5.55	0.290	0.927
DNN-Eng-2227	7.26	0.408	0.962
DNN-Man-2178	7.54	0.426	0.974
DNN-Multi-2232	7.15	0.398	0.933
BN-Eng-2227	5.86	0.336	0.955
BN-Man-2178	6.07	0.356	0.976
BN-Multi-2232	5.58	0.325	0.951

Table 3 presents the results of different systems on the NIST SRE 2008 female tel-tel condition. From the results we can see that the performance of DNN based systems degrades sharply compared to the UBM based systems for all language conditions. The performance of the DNN-Multi system is slightly better than the DNN-Eng and DNN-Man systems due to the multilingual characteristics, but the performance is still very poor because of the data mismatch. However, results show that BN

features based systems perform significantly better than DNN based systems. The BN-Multi system performs best among the three BN features based systems which confirms the superiority of multilingual DNNs over monolingual DNNs on multilingual task. Compared to DNN-Eng based method, the relative improvements of BN-Eng are 19.3% in EER, 17.6% in minDCF08, the relative improvements of BN-Multi are 23.1% in EER, 20.3% in minDCF08. However, the performance of the BN-Multi system is still inferior to UBM based system especially on the minDCF08. Actually, the training data contains five languages and the number is even larger on the test data. BN features trained with merely two languages can hardly capture the pronunciation patterns of so many languages, while PLP features without phoneme-specific tuning might be more neutral for different languages thus leading to a more blurred clustering which might somehow benefit multilingual tasks.

Table 4: *The performance of BN based systems with different number of senones on the NIST SRE2008 short2-short3 tel-tel condition.*

system	EER(%)	minDCF08	minDCF10
UBM-2048	5.55	0.290	0.927
BN-Eng-2227	5.86	0.336	0.955
BN-Eng-4126	6.10	0.348	0.965
BN-Eng-8223	6.27	0.360	0.957
BN-Multi-2232	5.58	0.325	0.951
BN-Multi-4155	5.76	0.334	0.952
BN-Multi-8269	5.98	0.341	0.958

The results of BN based systems with different number of senones are presented in Table 4. The same phenomenon is observed as in the NIST SRE2010 task that the performance becomes worse by enlarging the senone sizes.

All these results clearly illustrate the effectiveness of the proposed BN features and multilingual DNNs based models. Compared to the DNN based approach, the BN features based model provides a more flexible and effective pattern to narrow the gap caused by data mismatch.

## 5. Conclusions

In this paper we investigate using BN features and multilingual DNNs for speaker verification. This approach uses the DNN for senone classification to extract BN features rather than to provide frame alignments directly. Then an GMM is trained using BN features to provide frame alignments in order to make the alignment model more relevant and reliable for the speaker verification data sets. Experiments on the NIST SRE 2008 female short2-short3 telephone task (multilingual) and the NIST SRE 2010 female core-extended telephone task (English) demonstrate the effectiveness of this method.

In the future, we will continue investigating the use of BN features for speaker verification, such as using different combinations of languages and parameter settings for DNNs training. Besides, the recently proposed network structures such as max-out networks [21, 22] will be explored for speaker verification.

## 6. Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant No. 61273268, No. 61370034 and No. 61403224.

## 7. References

- [1] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on 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," in *Computer Vision, 2007. IC-CV 2007. IEEE 11th International Conference on*. IEEE, 2007, pp. 1–8.
- [3] P. Kenny, "Bayesian speaker verification with heavy-tailed priors." in *Odyssey*, 2010, p. 14.
- [4] D. Garcia-Romero and C. Y. Espy-Wilson, "Analysis of i-vector length normalization in speaker recognition systems." in *Interspeech*, 2011, pp. 249–252.
- [5] G. E. Dahl, D. Yu, L. Deng, and A. 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.
- [6] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *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.
- [7] S. Yaman, J. Pelecanos, and R. Sarikaya, "Bottleneck features for speaker recognition," in *Odyssey 2012-The Speaker and Language Recognition Workshop*, 2012.
- [8] T. Yamada, L. Wang, and A. Kai, "Improvement of distant-talking speaker identification using bottleneck features of DNN." in *Interspeech*, 2013, pp. 3661–3664.
- [9] O. Ghahabi and J. Hernando, "Deep belief networks for i-vector based speaker recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 1700–1704.
- [10] E. Variansi, X. Lei, E. McDermott, I. L. Moreno, and J. Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 4052–4056.
- [11] Y. Lei, N. Scheffer, L. Ferrer, and M. McLaren, "A novel scheme for speaker recognition using a phonetically-aware deep neural network," *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014.
- [12] P. Kenny, T. Stafylakis, P. Ouellet, V. Gupta, and J. Alam, "Deep neural networks for extracting Baum-Welch statistics for speaker recognition," in *Odyssey 2014*, 2014, pp. 293–298.
- [13] D. Yu and M. L. Seltzer, "Improved bottleneck features using pre-trained deep neural networks." in *Interspeech*, 2011, p. 240.
- [14] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.-R. Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters*, vol. 49, no. 24, pp. 1569–1570, 2013.
- [15] J. Gehring, Y. Miao, F. Metze, and A. Waibel, "Extracting deep bottleneck features using stacked auto-encoders," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 3377–3381.
- [16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [17] H. Bourlard, J. Dines, M. Magimai-Doss, P. N. Garner, D. Imseng, P. Motlicek, H. Liang, L. Saheer, and F. Valente, "Current trends in multilingual speech processing," *Sadhana*, vol. 36, no. 5, pp. 885–915, 2011.
- [18] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7319–7323.
- [19] "The NIST year 2008 speaker recognition evaluation plan," <http://www.itl.nist.gov/iad/mig//tests/sre/2008/>, 2008.
- [20] "The NIST year 2010 speaker recognition evaluation plan," <http://www.nist.gov/speech/tests/spk/2010/index.html>, 2010.
- [21] I. J. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," *arXiv preprint arXiv:1302.4389*, 2013.
- [22] M. Cai, Y. Shi, and J. Liu, "Deep maxout neural networks for speech recognition," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 291–296.