Multilingual Tandem Bottleneck Feature For Language Identification

Wang Geng, Jie Li, Shanshan Zhang, Xinyuan Cai, Bo Xu

Interactive Digital Media Technology Research Center
Institute of Automation Chinese Academy of Sciences, Beijing 100190, China
{wang.geng,jie.li,shanshan.zhang,xinyuan.cai.xubo}@ia.ac.cn

Abstract

The deep bottleneck (BN) feature based ivector solution has been recognized as a popular pipeline for language identification (LID) recently. However, issues such as how to extract more effective BN features and how to fully utilize features extracted from deep neural networks (DNN) are still not well investigated. In this paper, these issues are empirically tackled by means as follows: First, two novel types of deep features, phone-discriminant and triphone-discriminate are extracted. Then, DNNs are trained both separately and jointly on multilingual corpuses to produce different BN features. Finally, tandem fashion on deep BN features is applied to build enhanced deep features. Experiment results show that systems built on top of tandem deep features obtain 19% and 42% relative equal error rate reduction on average on NIST LRE 2007 over the counterpart built on traditional deep BN features and the cepstral feature based LID system, respectively.

Index Terms: Language Identification, Deep Bottleneck Feature, Tandem Feature, Multi-deep Feature, multi-training procedure.

1. Introduction

For language identification task, the useful information in speech is latent and weak, and the performance of LID system is degraded largely due to short segment duration of the speech recordings in addition to the significant noise caused by different speakers, various channels and backgrounds [1]. The key to alleviate the above problems is to investigate effective representation of language information. Generally, shift delta cepstral (SDC) feature is the most popular acoustic representation of the existing LID Systems. It can be considered as an linear extension of the traditional perceptual linear prediction (PLP), mel-frequency cepstral coefficients (MFCCs) [2, 3, 4] which can exploit long-term contextual information among PLPs .On the basis of the above spectral features, a series of creditable gaussian mixture model (GMM) based modeling approaches learned from speaker verification are developed such as GMM suppport vector machine (SVM), joint factor analysis(JFA) and total variability modeling(TVM) [5, 6].

However, the fixed structure and linear operation on the conventional acoustic features may not be adaptive enough to fit the diverse variations of the speech recordings [1]. In this paper, we propose various types of deep bottleneck features based on the i-vector representation to boost the LID system performance. We use the tagged data corresponding to phoneme states to train a special structured DNN which contains a narrow internal layer, the so-called BN-DNN. Since the number of hidden units in the bottleneck layer is normally much smaller than the non-bottleneck layers, BN-DNN training process will force the activation signals in the bottleneck (BN) layer to form a low-dimensional compact representation of the original inputs [7]. The output feature extracted from the BN layer is termed deep bottleneck feature, and is used for total variability modeling (TVM). The BN feature has been found to be a powerful representation of the underlying phonemes or phoneme states [8, 9, 10] which allows the BN feature to be inherently more robust to variations such as different channels, speakers, and background noise that are irrelevant to phonemes or phoneme states.

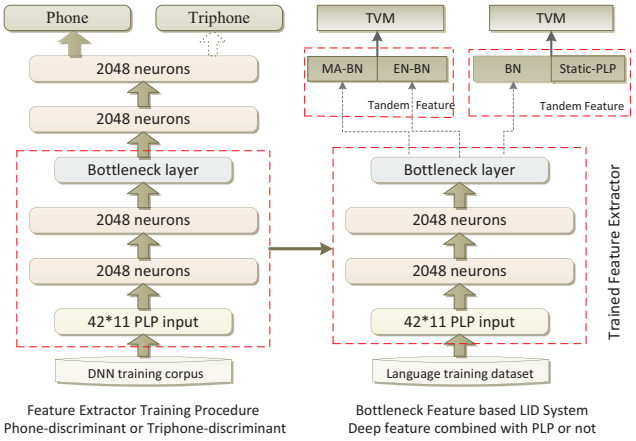

Figure 1: Framework of using deep BN feature for ivector-based LID system. The left is the BN-DNN training procedure, the right is the diagram of the BN feature based LID system

Although much research work has been done on DNN-based language identification [5, 1, 7], there is still good potential for applying deep feature on LID task. As the work in [7] shows, BN-DNN training process is not directly aim to language identification task since the output labels are phonemes or phoneme states rather than the target languages [7]. However, the deep BN feature yields promising results . Inspired by the work in [11], there can be a variety of feature choices depending on the type of neural network supervision labels and, the tandem fashion on features can also obtain additional gains. Multilingual research indicates that the bottleneck features are in possession of the language independent property to some degree [12, 13], so the available transcribed data resource from other languages can be utilized to train multilingual acoustic models such as multilingual neep neural networks [14, 15]. These techniques implement data borrowing on the model level and opens up research on gaining deeper insight in the under-

The work was supported by 973 program in china, grant No. Y3A2011D81.

standing of the influence of all the transcribed data resources for language identification.

Motivated by the above techniques, we propose two approaches to make full use of the available multilingual resources. The first method is to apply multilingual joint training procedure. The other one is to train mono-lingual BN-DNNs separately on each language and concatenate the individual BN features. At the same time, when concatenating the BN feature with conventional acoustic features for GMM-based i-vector system, we obtain additional gains. Under this framework, we first experimentally explore the impacts of different feature types. Then, different approaches of utilizing data from multilingual are evaluated in a more holistic manner. Furthermore, the necessity of delta processing is discussed. At the end, we give a detailed description of the influence of feature combination. It confirms that the proposed deep model features can be compatible with the classical acoustic features in the GMM-based i-vector framework

The remainder of the papers is organized as follows: Section 2 gives a brief description of the structure of BN-DNN, our proposed deep bottleneck features are introduced in Section 3. The experiment comparisons and results are analyzed in Section 4 and our whole work is summarized in Section 5.

2. Framework Description

The proposed deep BN feature based language identification framework is illustrated in Figure 1. It shows that deep bottleneck features are generated from a specially structured BN-DNN. The BN-DNN contains five hidden layers and one of the internal layers possesses a relative smaller amount of hidden units, termed the bottleneck layer and, the bottleneck layer exists in the middle of BN-DNN model structure. This small layer can reduce the redundant information in the original input acoustic features. At the same time, it creates a constriction in the DNN that can make the useful information for classification into a low dimensional and compact representation. Typically, to get better features, BN-DNNs are trained to predict tied states of triphone and monophone states. After pre-trained and fine-tuned on language-irrelevant data, BN-DNNs are used to process language-relevant acoustic data to create deep bottleneck features. The deep BN features are applied on the i-vector based LID system. A 2048 diagonal covariance universal background model (UBM) is trained to compute frame posteriors for sufficient statistics collection. Then the total variability matrix is trained for i-vector extraction [16]. Once the i-vectors were extracted, linear discriminative analysis (LDA) followed by cosine distance backend is applied to perform the final language classification.

3. Deep Bottleneck Feature Extraction for Language Identification

There exists inconsistency between the conventional acoustic features and the language identification task, as these spectral features extracted in a short time can not well represent sound characteristics of a relatively long duration, such as language identification task [7][11]. Hence, it is necessary to construct new features that are more specific for the task of language identification.

To address the above problem, the proposed BN features extracted from the hidden layer of deep neural network are regarded as high level features. It is a nonlinear discriminant transformation of the original acoustic inputs and this makes the BN feature to be inherently more robust to variations in language identification. 11 frames raw spectral features are fed into the BN-DNN to generate deep BN features. This context span improves the long duration problem and may also lead to more task-specific features. Hence, deep BN features are usually regarded as better representation than original acoustic features for specific tasks like language identification.

Although much exploration has been done on applying deep features on language identification, it has not been thoroughly studied for LID task. More importantly, depending on the supervised labels of the neural network and training procedure, there can be a number of feature choices. But unfortunately, there has not been a detailed investigation on further research. In this paper a variety of deep bottleneck features are proposed to improve the performance of language identification systems. This scenario gives a detailed description of different supervision labels of BN-DNN , various training strategies of utilizing the available resources and diverse feature combination approaches.

3.1. Two Types of Deep Bottleneck Features

As indicated in the previous section, we have examined the impact of choices of target labels used for BN-DNN training on system performance. The BN-DNN supervised labels can be triphone and monophone states.

♢ Triphone-discriminant BN-DNN

Triphone states are closely linked to context information and have been widely accepted in automatic speech recognition (ASR) system. Here, we train BN-DNNs to predict tied states of context-dependent triphone. After pre-trained and fine-tuned, triphone-discriminant BN-DNNs are treated as feature extractor to generate deep bottleneck features.

♢ Phone-discriminant BN-DNN

It was hypothesized that using context-independent monophone state labels was sufficient enough for training BN-DNN to extract BN feature [10][17]. It is believed that phone-discriminant BN-DNN can retain enough information about the inputs which is especially useful in language identification task. As presented in the following experiment section, Deep features extracted from phone-discriminant BN-DNN capture more language specific information and actually show promising results.

3.2. BN-DNN Training on Utilizing Multilingual Resources

Two straight-forward means are applied to make use of the available multilingual resources. The first method is to apply multilingual joint training procedure. The second is to separately train mono-lingual BN-DNNs on each kind of training materials and then concatenate the individual BN feature. Multilingual joint training makes the BN-DNN model more comprehensive as the supervised labels become larger and here we get the Multi-BN model (trained using about 750h language irrelevant Mandarin and English hybrid training materials). The second method is based on the idea that deep features extracted from mono-lingual BN-DNNs trained respectively may have complementary information. Here, we obtain the MA-BN model (trained using about 370h language irrelevant Mandarin training data) and the EN-BN model (trained using about 370h language irrelevant English training materials).

3.3. Feature Combination

Table 1: Performance EER (%) of deep BN feature extracted from different supervision label type in BN-DNN; comparison between phone label and triphone label

EER(%)	MA-BN_static			MA-BN_delta		
	3s	10s	30s	3s	10s	30s
Phone	16.6	7.6	2.7	17.0	7.1	2.4
Triphone	19.5	10.5	5.1	21.1	10.2	4.4

Motivated by speech recognition, we combine deep neural network features with conventional cepstral features in a tandem fashion. Being regarded as hign-level feature, BN feature can provide substantial complimentary information to the raw acoustic features, and in turn, it can also be augmented with standard cepstral feature for i-vector modeling. Moreover, different types of neural network features can also be concatenated. This is done by simply concatenate deep features from different types of neural networks to obtain combined multi-deep features and the combined multi-deep features can indeed obtain additive gains.

4. Experiments

4.1. Baseline System

The baseline system is a standard GMM-based i-vector framework established by following the classical recipe describe in [18]. The raw acoustic features are 14-dimensional vectors that composed of 13-dimensional PLP and pitch coefficient, and then SDC feature is obtained as a linear extension of the PLP [3, 5]. A diagonal covariance UBM with 2048 components is trained for providing frame alignments and collecting Baum-Welch sufficient statistics. Then the total variability matrix T is trained for i-vectors (600D representative vectors) extraction. Once finishing extracting the i-vectors, we further reduce the dimensionality of i-vectors to 13 by linear discriminative analysis and apply cosine distance to perform final classification.

4.2. Experiment Setup

In this scenario, there are 5 hidden layers in the trained BN-DNNs with 40 units in the bottleneck layer and 2048 units in each non-bottleneck layer. The input to the BN-DNN is the concatenation of 11 frames of 42-dimensional features (13-dimensional PLP and pitch appended with the first and second order derivatives).

A stack of restricted bolzmann machines (RBM) are trained to initialize BN-DNN weight matrices, after pre-training, the BN-DNN is fine-tuned in a supervised manner with the standard error back-propagation to classify the phone states. When finished training, BN-DNNs are treated as feature extractors to extract BN features for GMM-based i-vector system. The experiment test corpus is NIST LRE 2007 dataset including 3s, 10s and 30s conditions. All experiments are carried out with the open source toolkit KALDI.

Table 2: Effects of delta processing; comparison between static BN feature and delta processed BN feature

EER(%)	MA-BN-DNN			EN-BN-DNN		
	3s	10s	30s	3s	10s	30s
BN-static	16.6	7.6	2.7	18.2	7.9	3.5
BN-delta	15.9	7.1	2.4	17.3	7.5	3.0

4.3. Evaluation and Results

4.3.1. Optimizing BN Features

To get optimal performance of BN feature based LID systems, different configurations of BN features are experimentally studied. Here we focus on three aspects, different supervised labels of BN-DNNs, the necessity of delta processing and the training procedure on utilizing multilingual resources.

♢ Type Selection of Deep Feature Extractor

BN-DNNs are trained to predict tied states of context-dependent triphone and context-independent monophone states. As presented in Table 1, phone-discriminant BN-DNN shows absolutely better performance in both static and delta processed BN feature extracted from MA-BN-DNNs while the triphone-discriminant BN-DNN performs not as good as the baseline system. This highlights the importance of supervised labels for BN-DNN training. It is believed that phone-discriminant BN-DNN may capture enough language-specific information about the raw inputs and that we hypothesize that monophone states are sufficient enough for training BN-DNNs. According to the above result, phone-discriminant BN-DNNs are chosen for the following experiments as done in [1][7].

♢ Necessity of Delta Processing

Delta processing is important for deep bottleneck features [19]. As shown in Table 2, BN features extracted from both the MA-BN and EN-BN obtain additional EER reduction on longer duration testset after delta processing. Moreover, delta processing plays a more important role when the test utterance gets longer and more than 10% relative EER reduction over the BN-static feature is obtained on the 30s test condition. Observed from the result, we can deduce that feature with delta processing contains more long-term contextual information which helps to make the LID system more stable. Similar idea was enhanced in [19] and showed slight improvements. Learning from the above result, it is necessary to apply delta processing on the features and we keep this operation for the remainder of the experiments.

♢ Feature Extraction on Utilizing Multilingual Resources

As it shows in Table 3, the multi-BN feature performs a little worse than original acoustic feature, thus it seems that the multilingual joint training procedure is not an optimal method to make full use of the available multilingual resources. This indicates that sufficient language-irrelevant training materials are helpful for DNN training, while hybrid training materials seem to not contribute to improve performance. Since the extraction of the multi-BN feature is disconnected from the manner in which features are utilized during language modeling, the connections among the phone states of different languages may bring some confusing information to supervised labels.

Both the MA-BN and EN-BN outperform the multi-BN

Table 3: Results EER (%) of mono-lingual and multilingual BN feature

EER(%)	MA-BN	EN-BN	Multi-BN	MA-BN+EN-BN	Baseline
3s	17.0	19.1	24.9	14.6	20.39
10s	7.1	7.5	14.2	5.4	9.77
30s	2.34	3.0	6.8	1.9	4.44

feature and the baseline acoustic feature and more than 40% relative reduction in EER are observed on the 3s test condition. We can deduce that the common language specific information exists in speech data across different languages, and each BN-DNNs has similar ability to capture the commonalities and this point ensures the generalization capability of BN feature. Specially, the significant performance of MA-BN+EN-BN (the concatenation of MA-BN feature and EN-BN feature) in Table 3 indicates i-vector based LID system can benefit from BN feature trained on different resources that both the BN features can provide complementary information.

4.3.2. Evaluation of Feature Combination

In this section we give a detailed analysis on the feature complementarity among original acoustic feature and deep BN features and in particular, how that can lead to an optimal combination strategy. We define three different groups of features and concatenate them to obtain refined feature representations

- Tandem1: this group combines MA-BN and EN-BN feature with acoustic PLP feature separately to create refined feature: MA-BN+PLP(Tandem1_1) and EN-BN+PLP(Tandem1_2). This combination strategy examines the complementarity between deep BN features and conventional acoustic feature. In the case of ASR, it is believed that the deep BN features can capture complementary information to traditional acoustic feature [17].
- Tandem2: this tandem feature is composed of two types of BN features: MA-BN and EN-BN feature. This tandem fashion allows us to evaluate the combination capability of the proposed deep bottleneck features.
- Tandem3: this group attaches the conventional acoustic feature to Tandem2 to explore a global combination strategy for all the developed features.

Table 4: Results EER (%) of tandem features

EER(%)	Tandem1_1	Tandem1_2	Tandem2	Tandem3	Baseline
3s	15.7	16.7	14.6	14.3	20.39
10s	6.4	6.4	5.4	5.4	9.77
30s	2.2	2.6	1.9	1.86	4.44

As observed from Table 4, both the Tandem1_1 and Tandem1_2 feature get a >18% relative EER reduction on 3s test condition with respect to baseline acoustic feature and larger improvement on longer duration test segments are observed. It highlights the degree of complementarity between conventional features and deep bottleneck features. The performance of Tandem2 shows that the tandem fashion of the BN features outperforms the individual bottleneck features and can obtain about 14% relative EER reduction when compared with the best individual BN feature on 3s test condition. It demonstrates that BN

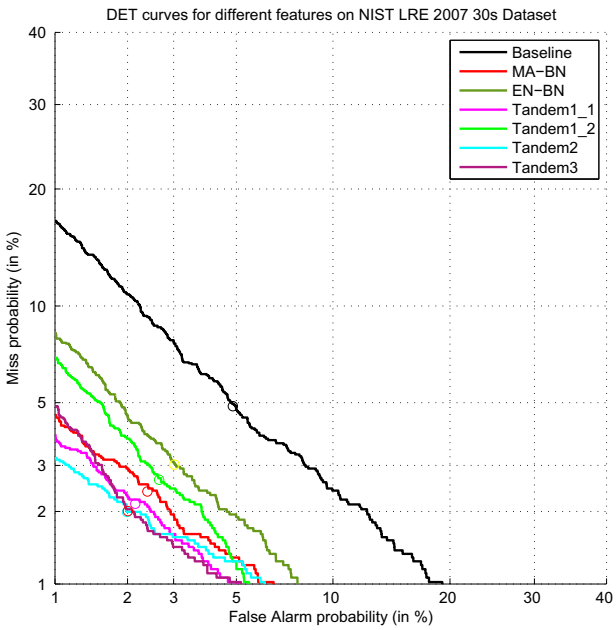

Figure 2: The DET curves for all the proposed approaches on NIST LRE 2007 30s condition

features extracted from BN-DNNs with different target labels capture different language specific information. As it shows in Table 4, the Tandem3 performs almost the same with Tandem2. Maybe the best concatenated deep feature can exploit enough language information to provide a better representation and some redundant information exists when fusing the best tandem BN feature with traditional acoustic feature.

Figure 2 shows the DET curves for all the proposed approaches investigated in this scenario on NIST LRE 2007 30s condition. Compared to the classical i-vector baseline system, the proposed deep features show significant performance and the optimal tandem feature achieves the best overall performance.

5. Conclusion and Future Work

This scenario presents a detailed research on applying various types of deep BN features on language identification task. Two types of BN-DNNs are proposed and experiment results show that phone-discriminant BN-DNN is more effective for language identification. Considering the available resources, different training procedure on utilizing multilingual resources are evaluated. Experiments show UBM-based i-vector LID system benefits significantly from the tandem fashion of BN features extracted respectively from BN-DNNs. Finally, to take advantage of the potential complementarity among different features, we combine deep features internally and with conventional acoustic features. Observed from the experiment, the best tandem feature obtains 42% relative EER reduction over the traditional acoustic feature system on average on NIST LRE 2007 testset.

In the future we intend to continue further research on deep BN features such as: the idea of training language-discriminant deep feature extractor like in [11] and combination different features to obtain refined feature representation.

6. References

[1] Y. Song, B. Jiang, Y. Bao, S. Wei, and L.-R. Dai, "I-vector representation based on bottleneck features for language identification," *Electronics Letters*, November 2013.

[2] P. A. Torres-carrasquillo, E. Singer, M. A. Kohler, and J. R. Deller, "Approaches to language identification using gaussian mixture models and shifted delta cepstral features," in *Proc. ICSLP 2002*, pp. 89–92.

[3] F. Allen, E. Ambikairajah, and J. Epps, "Language identification using warping and the shifted delta cepstrum," in *Multimedia Signal Processing, 2005 IEEE 7th Workshop*, Oct 2005.

[4] M. A. Kohler and M. Kennedy, "Language identification using shifted delta cepstra," in *Circuits and Systems. The 45th Midwest Symposium on*, Aug 2002.

[5] H. Wang, C.-C. Leung, T. Lee, B. Ma, and H. Li, "Shifted-delta mlp features for spoken language recognition." *IEEE Signal Process. Lett.*, 2013.

[6] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language Recognition via I-Vectors and Dimensionality Reduction," in *INTERSPEECH 2011*, Florence, Italy, Aug. 2011, pp. 857–860.

[7] B. Jiang, Y. Song, S. Wei, I. M.-L, and L.-R. Dai, "Task-aware deep bottleneck features for spoken language identification." in *Fifteenth Annual Conference of the International Speech Communication Association. 2014.*

[8] G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *Signal Processing Magazine, IEEE*, Nov 2012.

[9] Y.-B. Bao, H. Jiang, H. Liu, and L.-R. Dai, "Investigation on dimensionality reduction of concatenated features with deep neural network for lvcsr systems." in *(Signal Processing (ICSP), 2012 IEEE 11th International Conference*.

[10] M. Paulik, "Lattice-based training of bottleneck feature extraction neural networks." in *Fourteenth Annual Conference of the International Speech Communication Association. 2013*, 2013.

[11] T.-F. Fu, Y.-M. Qian, Y. Liu, and K. Yu, "Tandem deep features for text-dependent speaker verification." in *Fifteenth Annual Conference of the International Speech Communication Association. 2014*.

[12] G. Heigold, V. Vanhoucke, A. W. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks." in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2013.

[13] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, May 2013.

[14] Y. Liu, T.-F. Fu, Y.-C. Fan, Y.-M. Qian, and K. Yu, "Speaker verification with deep features." in *Neural Networks (IJCNN), International Joint Conference on*, 2014.

[15] P. Bell, P. Swietojanski, and S. Renals, "Multi-level adaptive networks in tandem and hybrid asr systems." in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013.

[16] D.Martinez, O.Plchot, L. Burget, O. Glembek, and P. Matejka., "Language recognition in ivectors space." in *Conference of the International Speech Communication Association. 2011.*, pp. 861–864.

[17] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks." in *Conference of the International Speech Communication Association. 2011*. ISCA, pp. 237–240.

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

[19] J. Li, R. Zheng, and B. Xu, "Investigation of cross-lingual bottleneck features in hybrid asr systems." in *Fifteenth Annual Conference of the International Speech Communication Association. 2014.*