i-vector representation based on bottleneck features for language identification

Yan Song, Bing Jiang, YeBo Bao, Si Wei and Li-Rong Dai

An i-vector representation based on bottleneck (BN) features is presented for language identification (LID). In the proposed system, the BN features are extracted from a deep neural network, which can effectively mine the contextual information embedded in speech frames. The i-vector representation of each utterance is then obtained by applying a total variability approach on the BN features. The resulting performance of LID has been significantly improved with the proposed BN feature based i-vector representation. Compared with the state-of-the-art techniques, the equal error rate is relatively reduced by about 40% on the National Institute of Standards and Technology (NIST) 2009 evaluation sets.

Introduction: Language identification (LID) is the task of automatically determining which natural language a given speech utterance belongs to. Generally, language information in speech is weak and latent, and largely dependent on the utterance content. Owing to diverse variations in speech utterances caused by different speakers, channels and background noise, it is difficult to extract the latent language information, especially for short-duration utterances with limited content. The key is to find effective representations of language information.

Existing LID systems generally represent the content with phonetic and acoustic features individually. The phonetic representations utilise the output of a specific phone recogniser (PR), and enjoy the power of robust language modelling techniques for contextual information, e.g. PR with language model (PRLM) and PR with support vector machine [1]. The acoustic representations are derived from a GMM model based on the shifted delta cepstra (SDC) feature, which is a linear expansion of the mel frequency cepstrum coefficient (MFCC) with fixed structure as in [2]. Compared with the phonetic representations, acoustic ones may take advantage of lower computational complexity, and do not need extra labelling for PR training. Recently, proposed acoustic modelling methods, such as GMM-SVM, factor analysis and total variability [1, 3], can achieve comparable or even better performance.

Generally, SDC features are believed to be effective, as they can exploit long-term contextual information between MFCCs. In this Letter, we ask the question 'whether the SDC feature is optimal for LID?'. The answer is often neglected. Despite the superior representative ability of SDC, one can reasonably question its optimality, as the fixed structure and linear operation may not be adaptive enough to fit the diverse variations of speech utterances. Recently, deep neural networks (DNNs) have achieved significant performance gains either as acoustic models or for front-end bottleneck (BN) feature extraction [4, 5] in many challenging speech recognition tasks. Motivated by the success of DNN, we utilise the BN feature as an alternative to SDC to train acoustic models for LID, i.e. *K*-nearest neighbours (KNN) classifier based on i-vector representation. To the best of our knowledge, it is the first instance of successfully applying DNN techniques for LID, and shows a significant performance improvement.

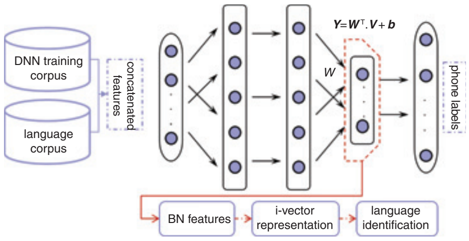

Fig. 1 *Proposed i-vector representation based on BN features for automatic LID*

BN feature extraction: Conventionally, BN features are generated by a specially structured DNN, which contains a narrow BN layer in internal position, the so-called BN DNN. Since the number of hidden nodes in the BN layer is normally much smaller than the other layers, DNN training will force the activation signals in the BN layer to form as a low-dimensional compact representation of the original inputs as long as the DNN is carefully trained to generate a low classification error rate in the output layer. Fig. 1 illustrates our BN feature extraction architecture.

The training of the BN DNN starts from Hinton and Salakhutdinov's unsupervised generative pre-training procedure [5] in which a stack of restricted Boltzmann machines will be trained to initialise weight matrices for better start points. After pre-training, the BN DNN is fine-tuned in a supervised manner with the standard error back-propagation procedure [6] to optimise a specified objective function. Given the training set $\boldsymbol{X}=\{\boldsymbol{x}_t,\ l_t\}_{t=1}^{\mathrm{T}}$, where l_i is the phone label of $\boldsymbol{x}_i$, the objective function can be defined as total negative log posterior probability of training data:

$$D=-\sum_{t=1}^{T}\log P(l_t|x_t) \qquad (1)$$

After the BN DNN is well trained, the linear outputs of the hidden units in the BN layer are directly used as BN features. Suppose $\boldsymbol{V}$ denotes the sigmoid outputs from the hidden layer just before the BN layer, then the BN features $\boldsymbol{Y}$ can be derived by using $\boldsymbol{Y}=\boldsymbol{W}^{\mathrm{T}}\cdot\boldsymbol{V}+\boldsymbol{b}$, where $\boldsymbol{W}$ is the BN layer weight matrix and $\boldsymbol{b}$ is a bias vector.

i-vector extraction: By using the trained BN DNN, the universal background model (UBM) can be obtained from the BN features extracted from a language corpus. We used data provided by the National Institute of Standards and Technology (NIST), including the Callhome corpus, Callfriend corpus, LRE03, LRE05 and LRE07 corpora and narrow-band VOA corpora. With the UBM model, each utterance is represented by using i-vector via low-dimensional transformation from the original feature sequence. Specifically, the i-vector representation can typically be formulated as

$$\boldsymbol{M}=\boldsymbol{m}+\boldsymbol{T}\boldsymbol{\omega} \qquad (2)$$

where $\boldsymbol{M}$ is the super-vector created by stacking all the mean vectors from the GMM for a given utterance. Typically, the GMM is adapted from the UBM model by using maximum *a posteriori*. $\boldsymbol{m}$ is the super-vector from the UBM model, $\boldsymbol{T}$ is the low-rank loading matrix with each column representing the basis vector of a so-called 'total variability' space. $\boldsymbol{\omega}$ refers to the i-vector that we need to represent the speech utterance.

Classification: A linear classifier can then be applied with given language labels for classification or identification. Generally, SVM can be easily applied for classifier training and testing. Considering the computational complexity and conceptual simplicity, we applied the linear discriminative analysis (LDA) and within-class normalisation (WCCN) techniques [3] to find the most discriminative i-vector representation. Finally, each language was represented by the average-pooling of the training speech utterances of each language. For each test utterance, a KNN classifier was used with the confidence scores estimated by using the cosine distance for each language.

Experiment: To evaluate the effectiveness of the proposed BN feature based LID method, we conducted experiments by using the 2009 NIST language evaluation dataset. The data include 30 813 speech utterances with durations of 30, 10 and 3 s. These are from 23 natural languages, including some highly confusing language pairs, such as Russian and Ukrainian, Hindi and Urdu, Bosnian and Croatian etc. A detailed description of the dataset can be found in [7].

In addition to the LID training data provided by NIST, we use approximately 1000 h of Mandarin with phone-level labels for training the DNN. As illustrated in [8], each speech frame is represented by 43-dimensional (43D) features, including 39D MFCC and 4D pitch feature. The input to the BN DNN is the concatenate of 10 frames, and thus has a dimension of 430. The 5-hidden-layer BN DNN is trained for extracting 43D BN features.

For the LID acoustic modelling, a UBM was trained from the language corpus with 2048 mixtures. A 600D i-vector is extracted from each utterance. After LDA and WCCN, each language is finally represented by a 22D vector.

In the experiment, performance was evaluated by using equal error rate (EER). We compared our proposed i-vector based on BN features (BN i-vector) with several 'state-of-the-art' LID techniques, including

(1) i-vector based on SDC features (SDC i-vector).
(2) PRLM based on phonetic representation using Hungarian phone-recogniser trained based on NN/HMM (Hungarian PRLM).
(3) The PRLM based on phonetic representation using Russian phone-recogniser trained based on NN/HMM (Russian PRLM).

The results are shown in Table 1. Note that slight differences may exist in detailed implementation, such as the dataset, model training etc. For comparison, we re-implemented the above-mentioned techniques with the same experimental setting as for the BN i-vector method. The performance is comparable with the best reported results from NIST LRE 2009 as [3].

Table 1: Performance comparisons of the BN i-vector with other state-of-the-art techniques on NIST LRE 2009 test set

Methods	30 (s) (%)	10 (s) (%)	3 (s) (%)
SDC i-vector [3]	2.40	4.80	14.20
SDC i-vector	2.50	5.34	19.75
Hungarian PRLM	2.62	6.65	18.97
Russian PRLM	2.42	6.52	18.06
BN i-vector	1.98	3.47	9.71

The performance was evaluated on test utterances with 30, 10 and 3 s durations, in terms of EER (%).

Conclusion: This Letter presents an i-vector representation based on BN features extracted from the DNN. It benefits from the excellent capability of a DNN to mine information embedded in the high-dimensionality data. LID evaluation results show significantly improved performance on the 30, 10 and 3 s test conditions of NIST LRE 2009, compared with the state-of-the-art. This is especially evident for the short-duration test condition.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (grant 62173264), National 973 Program of China (grant 2012CB326405) and the Chinese Universities Scientific Fund (grant Wk2100060008).

20 May 2013
doi: 10.1049/el.2013.1721
One or more of the Figures in this Letter are available in colour online.

Yan Song, Bing Jiang, YeBo Bao, Si Wei and Li-Rong Dai (*National Engineering Laboratory for Speech and Language Information Processing, University of Science and Technology of China, People's Republic of China*)

E-mail: songy@ustc.edu.cn

References

1 Torres-Carasquillo, P.A., Singer, E., Gleason, T., *et al.*: 'The MITLL NIST LRE 2009 language recognition system'. Proc. ICASSP, Dallas, TX, USA, March 2010, pp. 4994–4997
2 Torres-Carasquillo, P.A., Singer, E., Kohler, M.A., *et al.*: 'Approaches to language identification using Gaussian Mixture models and shifted delta cepstral features'. Proc. Interspeech, Denver, CO, USA, September 2002
3 Dehak, N., Torres-Carasquillo, P.A., Reynolds, D., *et al.*: 'Language recognition via i-vectors and dimensionality reduction'. Proc. Interspeech, Florence, Italy, August 2011, pp. 857–860
4 Yu, D., and Seltzer, M.L.: 'Improved bottleneck features using pre-trained deep neural networks'. Proc. Interspeech, Florence, Italy, August 2011, pp. 237–240
5 Hinton, G.E., and Salakhutdinov, R.R.: 'Reducing the dimensionality of data with neural networks', *Science*, 2006, **313**, (5786), pp. 504–507
6 Bishop, C.M.: 'Neural Networks for Pattern Recognition' (Oxford University Press, Oxford, UK, 1995)
7 Information Access Division, National Institute of Standards and Technology (NIST): 'NIST LRE 2009 evaluation plan' http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf, accessed May 2009
8 Bao, Y.B., Jiang, H., Dai, L.R., and Liu, C.: 'Incoherent training of deep neural networks to de-correlated bottleneck features for speech recognition'. ICASSP, Vancouver, BC, Canada, March 2013