FACTORIZED HIDDEN VARIABILITY LEARNING FOR ADAPTATION OF SHORT DURATION LANGUAGE IDENTIFICATION MODELS

Sarith Fernando[1,2], Vidhyasaharan Sethu[1], Eliathamby Ambikairajah[1,2]
[1]School of Electrical Engineering and Telecommunications, UNSW Sydney
[2]DATA61, CSIRO, Sydney, Australia
sarith.fernando@unsw.edu.au

ABSTRACT

Bidirectional long short term memory (BLSTM) recurrent neural networks (RNNs) have recently outperformed other state-of-the-art approaches, such as i-vector and deep neural networks (DNNs) in automatic language identification (LID), particularly when testing with very short utterances (~3s). Mismatches conditions between training and test data, e.g. speaker, channel, duration and environmental noise, are a major source of performance degradation for LID. A factorized hidden variability subspace (FHVS) learning technique is proposed for the adaptation of BLSTM RNNs to compensate for these types of mismatches in recording conditions. In the proposed approach, condition dependent parameters are estimated to adapt the hidden layer weights of the BLSTM in the FHVS. We evaluate FHVS on the AP17-OLR data set. Experimental results show that the FHVS method outperforms the standard BLSTM approach, achieving 27% relative improvements with utterance-level adaptation over the standard BLSTM for 1s duration utterances.

Index Terms— Language Identification, Bidirectional LSTM, Factorized Hidden Variability Learning, DNN Adaptation

1. INTRODUCTION

Mismatch between training and testing utterances has been a perennial problem in language identification (LID) [1, 2]. This mismatch may be compensated for in the feature domain by techniques such as shifted delta coefficients and eigenfeatures [3, 4]. Use of the total variability i-vector modelling approach [5] to acquire fixed length characterization of utterances is a common practice in modern LID systems. This structure features low intra-class variability, producing compact clusters provided that sufficient statistics are estimated accurately from an utterance. The major problem with this framework is performance degradation for short utterances. However, modern end-to-end automatic LID systems using bidirectional long short term memory (BLSTM) recurrent neural networks (RNNs) are proven to be effective for short duration LID tasks [6]. Realistically, all prevailing machine learning techniques using DNNs are vulnerable due to the mismatch conditions between the training and testing data and leads to performance degradations [7]. The variabilities arising from these mismatches can be normalized either by augmenting the features or by the model transforming to match testing conditions. Several adaptation techniques can be used to minimize the training and testing condition mismatches and overcome the problem of these variabilities. Although data exists with sufficient duration to train the system, test utterances can have very short durations in practice. These short duration utterances are the most affected by mismatched conditions in the testing phase.

Adaptation of DNNs is important to obtain significant reductions in error rates [8-11]. A popular approach is to combine Gaussian mixture model adaptation techniques with DNNs and train the tandem systems [12]. In tandem systems, a DNN extracts bottleneck features to train an i-vector system. Subspace methods are also used to perform DNN based acoustic model adaptation [13]. Mismatch is reduced in adaptation methods by changing a previously-trained model to match the test conditions, as opposed to adaptive training, which reduces the mismatch during training. Adaptation methods based on linear transformation augment a condition dependent linear layer with the original model [14]. The adaptation is performed to a subset of model parameters in subspace methods [15], which can help to avoid overfitting. The prominent feature of regularization based adaptation is to maintain similarity between training and testing conditions by introducing an additional error in the training process [16]. Methods using adaptive training can be grouped as cluster adaptive training [17], feature normalization techniques such as constrained maximum likelihood linear regression [18], and vocal tract length normalization [19].

In this paper, we propose the factorized hidden variability subspace (FHVS) technique which performs the adaptation in a subspace to capture the variability in training and testing conditions. Condition dependent parameters are estimated in FHVS and recombined to the BLSTM layer using new weights which are adaptively trained. During the training i-vectors used to initialize the FHVS and then during test-time, a test i-vector is transformed through the trained FHVS to utilize the condition information therein. The main intuition is to use i-vectors to learn the parameters of the FHVS system, since they are more withstanding to mismatch conditions. However, any other feature vector can also be used to learn this subspace. We propose a modified BLSTM structure which is utterance adaptively trained using i-vectors [5]. The i-vectors can be considered as low dimensional

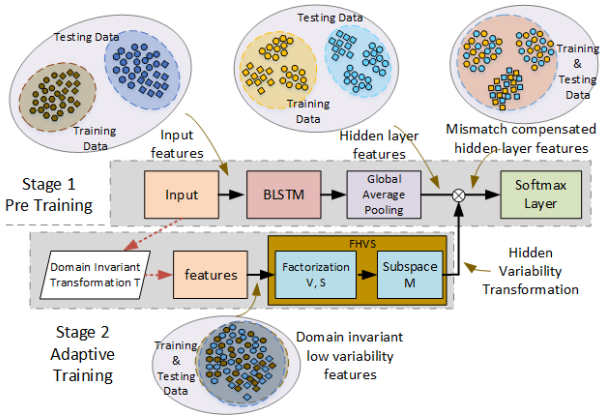

Fig 1: The marker shapes represent the instance labels and colors represent the original domains. Both training and testing domains are mapped to the hidden space using the unsupervised domain invariant transformation T (low variability subspace)**.** T is trained using i-vectors. V, S are the factorization matrices for the i-vector feature space. The metric M defined in the hidden variability space is learned to minimize the mismatch and to maximize the discriminative power between samples in the BLSTM network. Domain distributions are indicated by dashed ellipsoids. Our learning scheme non-linearly identifies the transformation M. This figure is best viewed in color.

representations of the utterance characteristics. Our method creates a new subspace that learns a feature transformation based on i-vectors.

2. PROPOSED FACTORIZED HIDDEN VARIABILITY SUBSPACE

A conceptual diagram of our proposal is as shown in Figure 1. Bidirectional LSTM systems are based on the idea that the output at time t may not only depend on the previous hidden elements in the sequence, but also on the future hidden elements. Stacking two LSTMs on top of each other forms a BLSTM. Then hidden states of both LSTMs are used to determine the output. The objective of LID is to accurately identify a given language from a pool of languages in a similar manner to a labeling task [20]. Therefore, labeling based on the past, present and future samples of the sequence may enhance the predictive capability of the embedded languages. BLSTMs process data in both directions and then the results of both directions are concatenated in the output of the BLSTM layer. This output $\boldsymbol{y_t}$ can be computed using the acoustic feature input to a layer $\boldsymbol{x_t}$, the forward sequence $\overrightarrow{\boldsymbol{h_t}}$ and backward sequence $\overleftarrow{\boldsymbol{h_t}}$ of the BLSTM hidden states at time t, as

$$\boldsymbol{y_t} = W_{\vec{h}y}\overrightarrow{\boldsymbol{h_t}} + W_{\overleftarrow{h}y}\overleftarrow{\boldsymbol{h_t}} \quad (1)$$

$$\overrightarrow{\boldsymbol{h_t}} = \mathcal{H}\left(W_{x\vec{h}}\boldsymbol{x_t} + W_{\vec{h}\vec{h}}\overrightarrow{\boldsymbol{h_{t-1}}} + \boldsymbol{b}_{\vec{h}}\right) \quad (2)$$

$$\overleftarrow{\boldsymbol{h_t}} = \mathcal{H}\left(W_{x\overleftarrow{h}}\boldsymbol{x_t} + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{\boldsymbol{h_{t+1}}} + \boldsymbol{b}_{\overleftarrow{h}}\right) \quad (3)$$

where W and $\boldsymbol{b}$ are the weights and biases respectively. For each LSTM memory block, the recurrent hidden layer function $\mathcal{H}$ is derived as in [21]. Finally, global average pooling is conducted for the complete sequence T of the BLSTM output, yielding $\boldsymbol{k}$ as,

$$\boldsymbol{k} = \sum_{\forall t \in T} \boldsymbol{y_t} \quad (4)$$

The main intuition for the factorization (V and S in Figure 1) begins with considering the case that an input feature $\boldsymbol{x_t}$ has been distorted by mismatched environment factors. This is passed through the system in equation (4) to get a distorted output $\boldsymbol{k}$. In this study, we propose compensating the hidden layer output vector by removing those unwanted parts in the network outputs caused by mismatched acoustic factors. This is achieved by adapting the existing DNN to overcome the mismatch between training and testing.

The modified $\boldsymbol{k}'$ can be found using the proposed factorized hidden variable subspace (FHVS) method, by adapting the pre-trained weight parameters of the BLSTM system and allowing to learn more abstract information throughout the learning scheme. Employing an utterance dependent (UD) feature transformation on the BLSTM weight parameters $W_{\vec{h}y}$ and $W_{\overleftarrow{h}y}$ in equation (1), equation (4) becomes

$$\boldsymbol{k}' = \left(\sum_{\forall t \in T} W_{\vec{h}y}\overrightarrow{\boldsymbol{h_t}} + W_{\overleftarrow{h}y}\overleftarrow{\boldsymbol{h_t}}\right) Q \quad (5)$$

where Q is the UD transformation matrix. Estimating the full matrix Q introduces a vast number of UD parameters, and so the utterance representation is reduced by applying a constraint on Q to be diagonal in the training process. We propose diagonal elements $\hat{\boldsymbol{p}}$ of Q to be learn using a low-dimensional representation of an utterance as,

$$\hat{\boldsymbol{p}} = \mathcal{H}(MVS\hat{\boldsymbol{\omega}} + \boldsymbol{\varphi}) \quad (6)$$

where $\hat{\boldsymbol{\omega}}$ is a vector in low dimensional space for a given utterance (in this paper, an i-vector), and M is the subspace and $\boldsymbol{\varphi}$ is the residual and UD training data used for learning. The V and S are the factorization matrices of feature vector $\hat{\boldsymbol{\omega}}$. The $\hat{\boldsymbol{\omega}}$ is extracted separately to the BLSTM training. Further, adding a nonlinear activation $\mathcal{H}$ in equation (6), specific to $\hat{\boldsymbol{\omega}}$ enables learning of an abstract representation during the training process. In addition, it enables learning $\hat{\boldsymbol{p}}$ more scalable to the original model.

The FHVS adaptation process produces a more stable Q when an additional constraint applied diagonally to the elements, restricting them in the span of $[0, r], r \in \mathbb{R}$. This scheme learns Q while keeping the initial model weights W fixed. Thus Q acts as a filter for the existing weights, improving them based on the condition/utterance dependent information. In the proposed method we are using the utterance-aware training (UaT) approach, in which utterance information is incorporated during training in addition to the standard acoustic features. The purpose of deploying UaT is that then BLSTM can make use of the additional information

about an utterance to adjust the model parameters for utterance normalization. The first step in UaT is the utterance information estimation where techniques like i-vectors [22] and bottleneck features [23] are commonly used. Further, it should be noted that here the adaptation conducted after the global average pooling layer where utterance level information can directly feed into the BLSTM network.

Typically, a matrix factorization problem finds matrices (S, V) such that the product SV^T closely approximates a given data matrix $\Omega \epsilon [\hat{\boldsymbol{\omega}}_1, \hat{\boldsymbol{\omega}}_2 \ldots \hat{\boldsymbol{\omega}}_N]$, while also requiring V and diagonal matrix S to satisfy certain properties such as non-negativity, sparseness, etc. This leads to an optimization problem of the form

$$\min_{S,V} l(\Omega, SV^T) + \Theta(S, V) \qquad (7)$$

where Θ is a regularization function to enforce the desired properties in S and V. Then l is some function that measures how closely Ω is approximated by SV^T. Improving the problem by exploiting a variety of contexts allows us to extract the different factors $\hat{\boldsymbol{\omega}}$, leading to superior generalization performance. Factorization can also be used to reduce the size of $\hat{\boldsymbol{\omega}}$ under a low rank assumption, which has the benefit of reducing the overall number of network parameters and improving training speed.

3. EXPERIMENTAL SETUP

The complete experimental setup is shown in Figure 2. The AP17-OLR training set with 10 languages for training [24] was used since it was developed specifically short duration language identification and allows for both testing under both matched and mismatched conditions. From this, 3 languages (Japanese, Russian and Korean) are recorded in two different environmental conditions and are designated 'mismatched', whereas all other languages have only one condition and are thus 'matched'. We have tested the system for 1s, 3s and 'all' duration development data which consists of 17964, 16404 and 17964 total utterances respectively. From these, we extract the 13-dimensional Mel-frequency cepstral coefficients (MFCC) features using a 25-ms window and a 10-ms frame-shift. After that bottleneck features (BNF) are extracted as in [24]. The BLSTM baseline is trained on the these BNF. Cepstral mean variance normalization is performed on the features locally, prior to presenting them to the BLSTM. The BLSTM has a single hidden layer with 1024 units, and 10 classes as the outputs. After the BLSTM layer, feature level averaging is conducted before feeding into softmax layer.

The i-vectors are extracted on top of the previously mentioned MFCCs. The universal background model consists of 2048 Gaussians. We extract i-vectors that are of 400 dimensions. Training is done using truncated back propagation through time (BPTT) with sequences of 100 frames. The initial BLSTM base model in Stage 1 (shown in Figure 1) is trained first. Then factorization matrices V and S are estimated using training i-vectors. In Stage 2, M and $\boldsymbol{\varphi}$

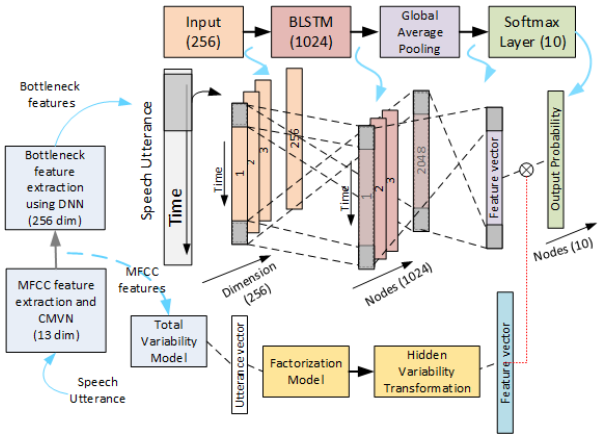

Fig 2: Bidirectional factorized hidden variability subspace framework for language identification.

are learned using the factorized i-vectors for training utterances, while keeping the initial model weights fixed. The final classification is performed using BNF and extracted factorized test i-vectors as input to the BLSTM.

In the results Section 4, we discuss two factorization methods, single value decomposition (SVD) [25] and linear discriminant analysis (LDA) [26]. With regards to the regularization term from equation (7), the SVD method aims to find components S, V that account for maximum variance in the training data (including error and intra-variable variance) whereas LDA maximizes the class separation (i.e. inter-class variance) when classes are known.

4. RESULTS AND ANALYSIS

In this analysis the focus is to test the hypothesized FHVS methods ability to compensate for mismatch conditions on a LID task, and to test the effectiveness of the above factorization methods on the i-vector (low variability) space. Herein, we refer to FHVS using SVD and LDA as SVD_HVS and LDA_HVS respectively.

4.1 Analysis of factorized hidden variability subspace

Before evaluating the performance of the proposed LID system, the feature space of each system (BLSTM and SVD_HVS) was investigated. Histograms of the lengths (magnitude) of BLSTM output vectors were generated for the Korean language before and after the proposed transformation to ascertain if there were differences in feature distributions. These are shown in Figure 3. Comparing Figures 3.a and 3.b, we can observe that the overlap between distributions has increased after FHVS transformation. This demonstrates that FHVS helps to overcome mismatch between feature vectors from training and testing utterances. Distribution mismatch can also be quantified using Kullback–Leibler divergence (KL). KL divergence is calculated between two distributions (assuming the distributions are Gaussian) for BLSTM output feature vectors

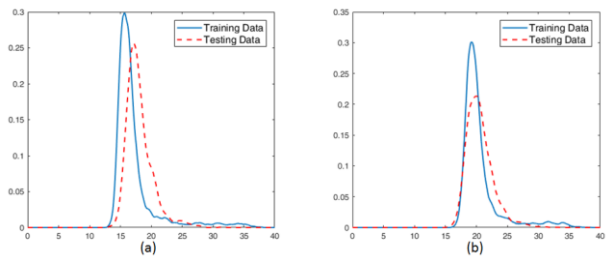

Fig 3: Comparison of training and testing data (a) before and (b) after SVD_HVS transformation (showing greater overlap) for Korean, a 'mismatched' language.

from training and testing utterances. These are modelled separately for Korean language. The resulted values of KL divergence became to be 0.7084 and 0.2232 before and after the transformation respectively, which demonstrate that there is a lower mismatch in the transformed space. Likewise, the KL divergence for all other languages is analyzed. Improvements are seen in KL divergence for each individual language, but Japanese, Russian and Korean had the highest improvement compared to all other languages. This suggests that FHVS transformation is more effective when there is a mismatch between training and testing data. Figure 4 shows a similar trend in system performance for individual languages.

Table 1 gives the performance of the BLSTM system and the gain that can be achieved by performing additional FHVS transformation. It is clear that FHVS (SVD_HVS) has significant improvements (73.2 to 79.4) for 1s duration utterances. This improvement is highly significant (**15.66**) in 'mismatched' condition languages (Japanese, Russian, and Korean) compared to 'matched' languages.

Table 1. Performance of the proposed system (SVD_HVS) compared to a BLSTM system for AP17-OLR 1s duration for matched and mismatched conditions.

Condition		Accuracy [%]		Improvement [%]
		BLSTM	SVD_HVS	
1	Matched	77.43	81.72	5.25
2	Mismatched	63.66	75.48	**15.66**
Overall		73.2	**79.4**	7.81

4.2 Reliability and effectiveness

Here, we compare the effectiveness of FHVS systems with respect to the original BLSTM framework as well as the hidden variability subspace (HVS) without factorization [27]. It should be noted that we have not applied any dimension reduction in the factorization methods in order to find the optimum number of factors. Therefore, for the purpose of direct comparison with HVS and FHVS we continued to employ 400 dimensional i-vectors without any fine tuning. We have tested only the effectiveness of orthogonalization

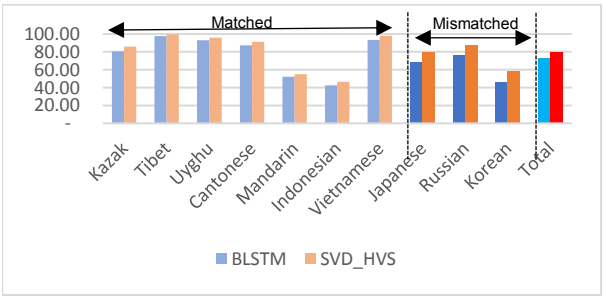

Fig 4: System performance comparison of BLSTM and SVD_HVS systems for AP17-OLR 1s condition in terms of accuracy, for each language.

into factors to learn the HVS parameters. Table 2 shows that the LDA_HVS system achieves the highest relative improvement of 27.32% in terms of Cavg, and 21.57% relative equal error rate (EER) reduction compared with the reference BLSTM system, confirming the effectiveness of the FHVS method. Moreover, it is clear that SVD_HVS and LDA_HVS systems have similar but significant performance gains. However, it is noticeable that HVS gives the majority of performance gain compared to baseline BLSTM and factorization helped to further reduce the error rates only slightly. Finally, we can see similar performance gains across the different durations of the utterances.

Table 2. Performance of the proposed FHVS system compared to the BLSTM system for AP17-OLR dataset.

System	Performance [%]					
	1s		3s		all	
	Cavg	EER	Cavg	EER	Cavg	EER
BLSTM	12.14	10.8	6.68	6.12	5.89	5.24
+HVS	9.14	8.55	4.11	3.98	3.64	3.38
+SVD_HVS	8.86	8.54	3.89	3.90	3.42	3.30
+LDA_HVS	**8.82**	**8.47**	**3.77**	**3.79**	**3.30**	**3.19**

5. CONCLUSION

In this paper, we have proposed a factorized hidden variability subspace (FHVS) method for mismatch adaptation to normalize multiple variabilities of speech signals for language identification. Our FHVS analysis shows that the orthogonality between different attribute subspaces is increased, further improving the performance over that of the hidden variability subspace (HVS) method. The proposed FHVS method estimates utterance dependent parameters in a FHVS and connects this to BLSTM layer using new weights which are adaptively trained. We evaluated the FHVS system on the AP17-OLR database. Experimental results showed that FHVS outperforms both the standard BLSTM system and a HVS approach.

6. REFERENCES

[1] R. Travadi, M. Van Segbroeck, and S. S. Narayanan, "Modified-prior i-vector estimation for language identification of short duration utterances," in *INTERSPEECH*, 2014, pp. 3037-3041.

[2] M.-G. Wang, Y. Song, B. Jiang, L.-R. Dai, and I. McLoughlin, "Exemplar based language recognition method for short-duration speech segments," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7354-7358.

[3] S. Fernando, V. Sethu, and E. Ambikairajah, "Eigenfeatures: An alternative to Shifted Delta Coefficients for Language Identification," presented at the SST2016, Parramatta, Australia, 2016.

[4] S. Fernando, V. Sethu, and E. Ambikairajah, "A Feature Normalisation Technique for PLLR Based Language Identification Systems," in *INTERSPEECH*, 2016, pp. 2925-2929.

[5] N. Dehak, P. A. Torres-Carrasquillo, D. A. Reynolds, and R. Dehak, "Language Recognition via i-vectors and Dimensionality Reduction," in *INTERSPEECH*, 2011, pp. 857-860.

[6] S. Fernando, V. Sethu, E. Ambikairajah, and J. Epps, "Bidirectional Modelling for Short Duration Language Identification," presented at the Interspeech 2017, Sweden, 2017.

[7] Joachim Fainberg, Steve Renals, and P. Bell, "Factorised representations for neural network adaptation to diverse acoustic environments," presented at the Interspeech 2017, 2017.

[8] L. Samarakoon and K. C. Sim, "On combining i-vectors and discriminative adaptation methods for unsupervised speaker normalization in dnn acoustic models," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, 2016, pp. 5275-5279.

[9] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based speaker adaptation of deep neural networks for french broadcast audio transcription," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 6334-6338.

[10] S. Xue, H. Jiang, L. Dai, and Q. Liu, "Speaker adaptation of hybrid NN/HMM model for speech recognition based on singular value decomposition," *Journal of Signal Processing Systems,* vol. 82, pp. 175-185, 2016.

[11] L. Samarakoon and K. C. Sim, "Learning factorized feature transforms for speaker normalization," in *Automatic Speech Recognition and Understanding (ASRU), 2015 IEEE Workshop on*, 2015, pp. 145-152.

[12] F. Richardson, D. Reynolds, and N. Dehak, "A Unified Deep Neural Network for Speaker and Language Recognition," *arXiv preprint arXiv:1504.00923,* 2015.

[13] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," 2014.

[14] B. Li and K. C. Sim, "Comparison of discriminative input and output transformations for speaker adaptation in the hybrid NN/HMM systems," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[15] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, 2014, pp. 6359-6363.

[16] D. Yu, K. Yao, H. Su, G. Li, and F. Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7893-7897.

[17] T. Tan, Y. Qian, M. Yin, Y. Zhuang, and K. Yu, "Cluster adaptive training for deep neural network," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*, 2015, pp. 4325-4329.

[18] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP),* vol. 23, pp. 1469-1477, 2015.

[19] L. Lee and R. C. Rose, "Speaker normalization using efficient frequency warping procedures," in *Acoustics, Speech, and Signal Processing, 1996. ICASSP-96. Conference Proceedings., 1996 IEEE International Conference on*, 1996, pp. 353-356.

[20] X. Ma and E. Hovy, "End-to-end sequence labeling via bi-directional lstm-cnns-crf," *arXiv preprint arXiv:1603.01354,* 2016.

[21] A. Zeyer, P. Doetsch, P. Voigtlaender, R. Schlüter, and H. Ney, "A comprehensive study of deep bidirectional LSTM RNNs for acoustic modeling in speech recognition," *arXiv preprint arXiv:1606.06871,* 2016.

[22] 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, pp. 1569-1570, 2013.

[23] P. Matejka, L. Zhang, T. Ng, H. S. Mallidi, O. Glembek, J. Ma, *et al.*, "Neural network bottleneck features for language identification," *Proc. IEEE Odyssey,* pp. 299-304, 2014.

[24] Zhiyuan Tang, Dong Wang, Yixiang Chen, and Q. Chen, "AP17-OLR Challenge: Data, Plan, and Baseline," *arXiv:1706.09742,* 2017.

[25] G. H. Golub and C. Reinsch, "Singular value decomposition and least squares solutions," *Numerische mathematik,* vol. 14, pp. 403-420, 1970.

[26] A. J. Izenman, "Linear discriminant analysis," in *Modern multivariate statistical techniques*, ed: Springer, 2013, pp. 237-280.

[27] S. Fernando, V. Sethu, and E. Ambikairajah. 2018, Hidden variability subspace learning for adaptation of deep neural networks. *Electronics Letters 54(3),* 173-175. Available: http://digital-library.theiet.org/content/journals/10.1049/el.2017.4027