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

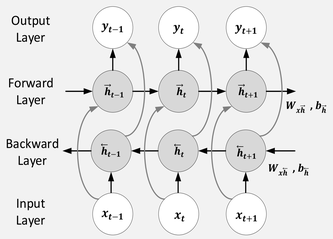
Sarith Fernando[1,2], *Vidhyasaharan Sethu*[1], *Eliathamby Ambikairajah*[1,2],
[1] The School of Electrical Engineering and Telecommunications, The University of New South Wales, Sydney NSW 2052, Australia
[2] DATA61, CSIRO, Australia, Sydney NSW 2015, Australia
sarith.fernando@unsw.edu.au

1. Introduction

- Mismatched 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
  - Compensate for these types of mismatches in recording conditions

2. Bidirectional LSTM recurrent neural networks

- The memory blocks of the LSTM hidden layers store the temporal state of the input at each time step, taking into account previous frames
- Bidirectional LSTMs are instead based on the idea that the output at time 't' may depend on both previous elements in the sequence as well as future elements



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

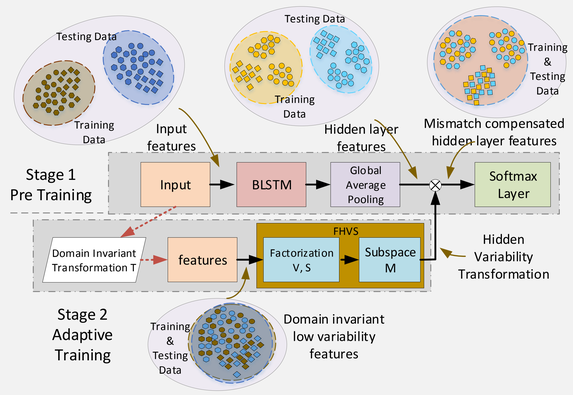
$$\vec{h}_t = \mathcal{H}\left(W_{x\vec{h}}x_t + W_{\vec{h}\vec{h}}\vec{h}_{t-1} + b_{\vec{h}}\right) \quad (2)$$

$$\overleftarrow{h}_t = \mathcal{H}\left(W_{x\overleftarrow{h}}x_t + W_{\overleftarrow{h}\overleftarrow{h}}\overleftarrow{h}_{t+1} + b_{\overleftarrow{h}}\right) \quad (3)$$

- Cell weights may help by capturing high level phoneme information
- Forget weights may help to reduce other common variations between languages
- Global average pooling is conducted for the complete sequence T of the BLSTM output, yielding k as,

$$k = \sum_{\forall t \in T} W_{\vec{h}y}\vec{h}_t + W_{\overleftarrow{h}y}\overleftarrow{h}_t \quad (4)$$

3. Proposed Factorized Hidden Variability Subspace



- DNNs are vulnerable due to the mismatch conditions between the training and testing data and leads to performance degradations
- To compensating the hidden layer output vector by removing mismatched acoustic factors in the network outputs
- This is achieved by adapting the existing BLSTM to overcome the mismatch between training and testing
- The modified k' can be found using the proposed factorized hidden variable subspace (FHVS) method, by adapting the pre-trained weight parameters of the BLSTM system
- Employing an utterance dependent (UD) feature transformation Q on the BLSTM weight parameters

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

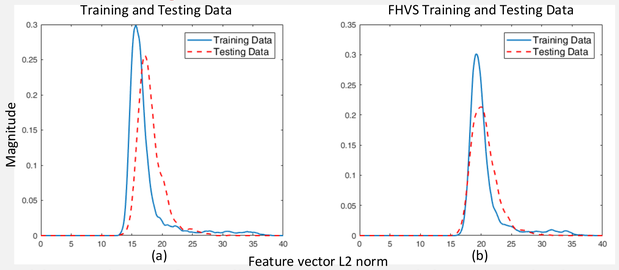
- Diagonal elements $\hat{p}$ of Q to be learn using a low-dimensional representation of an utterance as

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

where $\hat{\omega}$ is a vector in low dimensional space for a given utterance (in this paper, an i-vector), and M is the subspace and φ is the residual

- The V and S are the factorization matrices of feature vector $\hat{\omega}$

5. Results And Analysis



- The resulting values of KL divergence were (a) 0.7084 before and (b) 0.2232 after the FHVS (SVD_HVS) transformation
- Demonstrate that there is a lower mismatch in the transformed space

Table 1: Performance of the proposed system FHVS (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

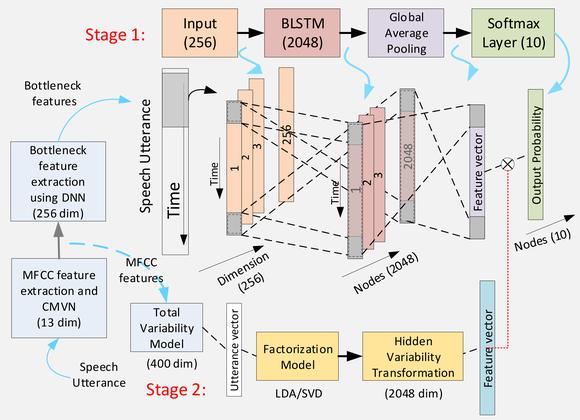
- FHVS (SVD_HVS) has significant improvements for 1s duration utterances and improvement is highly significant in 'mismatched' condition compared to 'matched'

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
FHVS	+SVD_HVS	8.86	8.54	3.89	3.90	3.42	3.30
FHVS	+LDA_HVS	8.82	8.47	3.77	3.79	3.30	3.19

4. System Description

- Test data
  - AP17-OLR > 10 target languages
  - Developed specifically for short duration language identification
  - 3 languages (Japanese, Russian and Korean) are recorded in two different environmental conditions and are designated 'mismatched'
  - All other languages have only one condition and are thus 'matched'
  - Tested the system for 1s, 3s and 'all' duration development data which consists of 17964, 16404 and 17964 total utterances respectively
- Features
  - Stage 1 > BNF of 256 dimensions
  - Stage 2 > i-vectors: 400 dimensions (UBM : 2048 component GMM)



6. Conclusion

- Proposed a factorized hidden variability subspace (FHVS) method for mismatch adaptation to compensate for multiple variabilities of speech signals for language identification
- 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
- FHVS method estimates utterance dependent parameters in a FHVS and connects this to BLSTM layer using new weights which are adaptively trained
- Experimental results showed that FHVS outperforms both the standard BLSTM system and a HVS approach