Merging of Native and Non-native Speech for Low-resource Accented ASR

Sarah Samson Juan, Laurent Besacier, Benjamin Lecouteux, Tien-Ping Tan

Merging of Native and Non-native Speech for Low-resource Accented ASR

Sarah Samson Juan[1], Laurent Besacier[2], Benjamin Lecouteux[2], and Tien-Ping Tan[3]

[1]Faculty of Computer Science and Information Technology
Universiti Malaysia Sarawak, Sarawak, Malaysia
sjsflora@fit.unimas.my
[2]Grenoble Informatics Laboratory (LIG), Univ. Grenoble-Alpes, Grenoble, France
{laurent.besacier,benjamin.lecouteux}@imag.fr
[3]School of Computer Science, Universiti Sains Malaysia, Penang, Malaysia
tienping@cs.usm.my

Abstract. This paper presents our recent study on low-resource automatic speech recognition (ASR) system with accented speech. We propose multi-accent Subspace Gaussian Mixture Models (SGMM) and accent-specific Deep Neural Networks (DNN) for improving non-native ASR performance. In the SGMM framework, we present an original language weighting strategy to merge the globally shared parameters of two models based on native and non-native speech respectively. In the DNN framework, a native deep neural net is fine-tuned to non-native speech. Over the non-native baseline, we achieved relative improvement of 15% for multi-accent SGMM and 34% for accent-specific DNN with speaker adaptation.

Keywords: automatic speech recognition, cross-lingual acoustic modelling, non-native speech, low-resource system, multi-accent SGMM, accent-specific DNN

1 Introduction

Performance of non-native automatic speech recognition (ASR) is poor when few (or no) non-native speech is available for training / adaptation. Many approaches have been suggested for handling accented-speech in ASR, such as acoustic model merging ([16], [2], [22], [23]), applying maximum likelihood linear regression (MLLR) for adapting models to each non-native speaker [8], or adapting lexicon ([1], [4]).

Lately, Subspace Gaussian Mixture Models (SGMMs) ([17], [18]) have shown to be very promising for ASR in limited training conditions (see [13] and [11]). In SGMM modelling, the acoustic units are all derived from a common GMM called Universal Background Model (UBM). This UBM, which in some way represents the acoustic space of the training data, can be estimated on large amount of untranscribed data from one or several languages. The globally shared

parameters do not need the knowledge about the phone units used in the source language(s). Without this constraint of source-target mapping of acoustic units, the UBM can be well used in cross-lingual or multilingual (multi-accent) settings.

In the mean time, Deep Neural Networks (DNNs) have been increasingly employed for building efficient ASR systems. HMM/DNN hybrid systems clearly outperform HMM/(S)GMM systems for many ASR tasks [6] which include dealing with low-resource systems ([14], [25], [9]). Several studies have shown that multilingual DNNs can be achieved by utilizing multilingual data for conducting unsupervised pretraining [21] or training the whole network simultaneously ([9], [25], [5]).

In the above techniques, acoustic model merging can easily be conducted through sharing the UBMs (for SGMM) and hidden layers (for DNN) with other systems. But what is the optimal way to do so? Can we merge a large amount of native speech with a small quantity of non-native data? This paper tries to respond to these questions using both SGMM (less efficient than DNNs but more compact for embedded applications) and DNN (state-of-the-art) frameworks. We apply our methods to Malaysian English ASR, where a large amount of native (English) data is available (TED-LIUM corpus [20]), while only 2h of non-native speech is available. More precisely, we propose one strategy for each framework: (1) language weighting for multi-accent SGMMs and (2) accent-specific top layer for DNN. The first strategy is novel and involves manipulating the number of Gaussians of each native / non-native model for (multi-accent) UBM merging. In the second approach, we build accent-specific DNN similarly to last year's work of [10] but we make it work for a very low-resource setting and with speaker adaptation on top of it.

The rest of the paper is organized as follows. In Section 2 we describe the background of SGMM and DNN as well as their application to multilingual and multi-accent ASR. Section 3 presents the experimental setup for building native and non-native systems as well as the results of our baselines. In Section 4 and 5, we describe the proposed strategies and show their benefits to low-resource accented ASR. Last but not least, Section 6 concludes this paper.

2 Background of acoustic modelling for cross-lingual or accented ASR

2.1 Subspace Gaussian Mixture Models

The GMM and SGMM acoustic models are similar since each emission probability of each HMM state is modelled with a Gaussian mixture model. However, in the SGMM approach, the Gaussian means and mixture component weights are generated from the phonetic and speaker subspaces along with a set of weight projections. For SGMM, the state probabilities are defined following the equations below [18]:

$$p(\mathbf{x}|j) = \sum_{m=1}^{M_j} c_{jm} \sum_{i=1}^{I} w_{jmi} \mathcal{N}(\mathbf{x}; \mu_{jmi}, \mathbf{\Sigma}_i) \quad (1)$$

$$\mu_{jmi} = \mathbf{M}_i \mathbf{v}_{jm}, \tag{2}$$

$$w_{jmi} = \frac{\exp \mathbf{w}_i^T \mathbf{v}_{jm}}{\sum_{i'=1}^{I} \exp \mathbf{w}_{i'}^T \mathbf{v}_{jm}} \tag{3}$$

where $\mathbf{x} \in \mathbb{R}^D$ denotes the D-dimensional feature vector, $j \in \{1..J\}$ is the HMM state, i is the Gaussian index, m is the substate and c_{jm} is the substate weight. Each state j is associated to a vector $\mathbf{v}_{jm} \in \mathbb{R}^S$ (S is the phonetic subspace dimension) which derives the means, μ_{jmi} and mixture weights, w_{jmi}, I is the number of Gaussians for each state. The phonetic subspace $\mathbf{M}_i$, weight projections $\mathbf{w}_i^T$ and covariance matrices $\mathbf{\Sigma}_i$, i.e., the globally shared parameters $\mathbf{\Phi}_i = \{\mathbf{M}_i, \mathbf{w}_i^T, \mathbf{\Sigma}_i\}$ are common across all states.

These parameters can be shared and estimated over multiple language data. [13] and [11] presented cross-lingual and multilingual work using SGMM for improving ASR with very limited training data. In both studies, the cross-lingual approach was carried out by porting the UBM which was trained using source language data, to SGMM training of target language. Basically, the SGMM model was derived from the UBM of source language. For the second approach, the strategy involved training UBM using more than one language data and then employed the multilingual UBM for SGMM training of a specific language. Applying both methods improved ASR performance of monolingual system.

This idea motivates us to investigate a multi-accent approach using this framework. We propose to build SGMM models which are derived from *merged* UBMs, rather than carrying out the SGMM training in a multilingual fashion (see studies on non-native SGMM in [15], [24]). The method is particularly appealing if one wishes to consider borrowing UBMs of other systems. Our strategy and experiments are described in Section 4.

2.2 Deep Neural Networks

Deep Neural Network (DNN) for ASR is a feedforward neural network with hidden layers. Mathematically, each output (node) of the l-th layer of a DNN can be defined as

$$\mathbf{x}_l = \sigma(\mathbf{b}_l + \mathbf{W}_l \mathbf{x}_{l-1}),\ for\ 1 \leq l < L \tag{4}$$

where $\mathbf{W}_l$ is the connection weight from $\mathbf{x}_l$ and $\mathbf{x}_{l-1}$, the output of the $(l-1)$-th layer, while $\mathbf{b}_l$ is the bias. The hidden output $\mathbf{x}_l$ is a sigmoid function defined as $\sigma(x) = (1 + \exp(-x))^{-1}$. The last ($L$-th) layer of the DNN uses a softmax function to obtain the posterior probability of each HMM state j given the acoustic observation $\mathbf{o}_t$ at time t:

$$p(j|\mathbf{o}_t) = \frac{\exp(x_L)}{\Sigma_{j'} \exp(x_L)}\,. \tag{5}$$

Optimizing hidden layers can be done by pretraining the network using Restricted Boltzmann Machines (RBM) [7]. The generative pretraining strategy builds stacks of RBMs corresponding to the number of desired hidden layers and provides better starting point (weights) for DNN fine-tuning through

backpropagation algorithm. Pretraining a DNN can be carried out in a unsupervised manner because it does not involve specific knowledge (labels, phone set) of a target language[1]. Only the softmax layer is sensitive to the target language. It is added on top of the hidden layers during fine-tuning and its ouput corresponds to the HMM states of the target language.

As shown in [21], using untranscribed data for RBM pretraining as a multilingual strategy has little effect on improving monolingual ASR performance. The *transfer learning* [5] approach has shown large recognition accuracy improvements. The method involves removing the top layer of a multilingual DNN and fine-tuning the hidden layers to a specific language.

Recently, a multi-accent DNN with accent specific softmax layer has been proposed for improving decoding performance of English ASR for British and Indian accents [10]. The accent adaptation approach yielded better decoding results compared to non-adapted DNNs. Another attempt to improve ASR performance on non-native task was done by [3] for Mandarin language. They also proved the interest of adapting non-native accents over the baseline DNN model.

In this paper, we investigate a method similar to [10] to build the accent-specific network models, but we apply it in a very low-resource setting. Previously, the method has been tested with larger amount of non-native speech (x10 or x100 compared to our experimental conditions). Hence, we try to measure the effectiveness of the approach when the ratio between non-native data and native data is largely unbalanced. In addition, we develop a strategy to handle cross-lingual DNNs with different feature transforms for speaker adaptation.

3 Experimental Setup

The ASR experiments were conducted on Kaldi speech recognition toolkit [19]. This section reports non-native and native speech databases used in our investigation. Besides that, we present the baseline results for non-native ASR based on GMM, SGMM and DNN.

3.1 Data

The non-native speech corpus contains 15h of English speech spoken by 24 Malaysians (of Malay, Chinese and Indian origin). The data were collected by *Universiti Sains Malaysia* for conducting research on acoustic model merging for ASR (see [23] for more details). Table 1 shows the amount of data used to train and evaluate the non-native ASR. We employed 2h of transcribed data for training the system and evaluate its performance on 4h of transcribed speech. For SGMM training, 9h of untranscribed data were added to the 2h of transcribed speech to build the UBM. Our system used the CMU pronunciation dictionary (no non-native adaptation of the lexicon) which has more than 100k words.

[1] In that sense, RBM pretraining (for DNN) and UBM training (for SGMM) are both unsupervised methods to get an initial representation of the acoustic space before modelling the speech units

Furthermore, we used a trigram language model for decoding. The model was trained on news data, taken from a local English news website[2]. After evaluating the LM on the test transcription data, the LM perplexity is 189 while the OOV rate is 2.5%.

Table 1. *Statistics of the non-native speech data for ASR.*

Train		Test
Untranscribed	Transcribed	
9h	2h	4h

To obtain a baseline for native ASR, we used the first release of TED-LIUM [20] corpus[3]. The transcriptions of this corpus were generated by the *Laboratoire d'Informatique* at *Université du Maine* (LIUM) for the International Workshop on Spoken Language Translation (IWSLT) evaluation campaign in 2011. The corpus contains speeches that were excerpted from video talks of the TED website. We used 118h to train the system and 4h for evaluation. Besides that, we used a pronunciation dictionary which was included in the package. For decoding, we used a trigram language model which was built on TED and WMT11 (Workshop on Machine Translation 2011) data. The model perplexity is 220 after estimation on the test data.

3.2 Baseline systems

For the non-native ASR system, we trained a triphone acoustic model (39 MFCC with deltas and deltas deltas) using 776 states and 10K Gaussians. Then, we trained SGMM using the same decision trees as in the previous system. The SGMM was derived from a UBM with 500 Gaussians and phonetic subspace dimension was $S = 40$. The UBM was trained on 11h data. We built a DNN based on state-level minimum Bayes risk [12] (sMBR) and the network had 7 layers, each of the 6 hidden layers had 1024 hidden units. The network was trained from 11 consecutive frames (5 preceding and 5 following frames) of the same MFFCs as in the GMM system. Besides that, the same HMM states were used as targets of the DNN. The initial weights for the network were obtained using Restricted Boltzmann Machines (RBMs) that resulted in a deep belief network with 6 stacks of RBMs. Fine tuning was done using Stochastic Gradient Descent with per-utterance updates, and learning rate 0.00001 which was kept constant for 4 epochs. To run our DNN experiments, we utilized a GPU machine and CUDA toolkit to speed up the computations.

For the native ASR system, we built a triphone acoustic model with 3304 states and 40K Gaussians. Subsequently, we built SGMM system using the same decision trees and 500 UBM Gaussians. Lastly, we trained a DNN with 7 layers

[2] http://www.thestar.com.my/

[3] We are aware that TED-LIUM is not a truly native English corpus (non-native speakers of multiple origins) but we consider here that the corpus permit to build an efficient system to decode native English ASR. Thus, in this paper we call it "excessively" a native corpus.

using the same setting for building non-native DNN. The three systems were evaluated on native speech (TED task) and we achieved the following WER results: 30.55% for GMM, 28.05% for SGMM and 19.10% for DNN.

Table 2. *Word error rates (WER %) of ASR with non-native (2h) and native (118h) acoustic models on the non-native evaluation data (4h test) - same pronunciation dictionary and language model for both system.*

Acoustic Models	Non-native	Native
GMM	41.47	57.09
SGMM	40.41	45.84
DNN	32.52	40.70

Table 2 presents the baseline results of systems that used non-native and native acoustic models, evaluated on accented speech. For non-native acoustic modelling, SGMM and DNN systems outperformed the GMM system. The systems gave 3% and 22% relative improvement, respectively. Using these non-native models (trained on 2h only!) to decode non-native speech resulted lower word error rate (WER) compared to the pure native ASR systems (trained on 118h). In the following sections, we try to take advantage of both corpora (*large* native and *small* non-native) by merging acoustic models (or data) efficiently.

4 Language weighting for multi-accent Subspace Gaussian Mixture Models

4.1 Proposed Method

In SGMM, the system is initialized by a Universal Background Model (UBM) which is a mixture of full-covariance Gaussians. This single GMM is trained on all speech classes that are pooled together. The advantage of this model is that it can be trained on large amount of untranscribed data or multiple languages, as shown in [13] for cross-lingual SGMM in low-resource conditions. The authors showed that the SGMM global parameters are transferable between languages, especially when the parameters are trained in multilingual fashion. Thus, this gives an opportunity for low-resource systems to borrow UBM trained from other sources.

Figure 1 illustrates the process of UBM merging through language weighting. The first step is to choose a language weight, α to L_1 in order to determine the number of Gaussians to be kept for merging ($(1-\alpha)$ is given to L_2). Intuitively, a larger α should be given to the less represented source data. Then, we use data that are representative of the ASR task in order to find the top αN Gaussians in L_1 UBM using maximum likelihood criterion. The same process is done for the L_2 UBM but only (1-α)N Gaussians are selected. The final step applies weight normalization before merging all the Gaussians in a single GMM. The final UBM should have the same number of Gaussians if both initial UBMs are the same size.

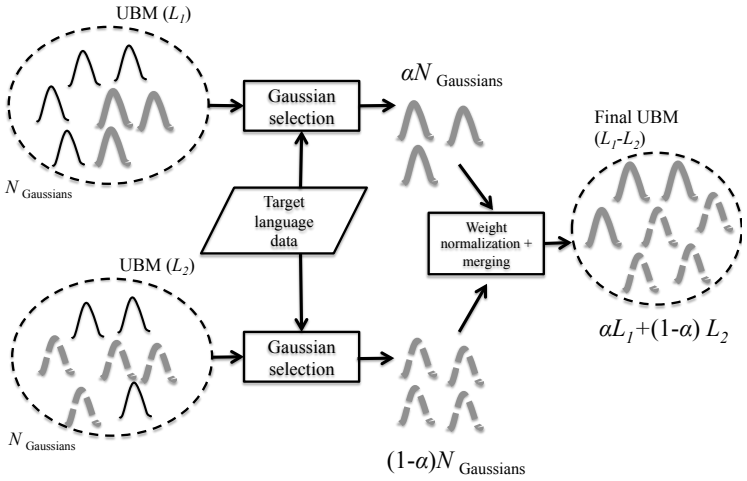

Fig. 1. *An Illustration of UBM merging through language weighting*

For experiments, we built a multi-accent UBM by merging native and non-native models using our language weighting strategy. To implement this, we used UBM of native speech (trained on 118h) and UBM of non-native speech (trained on 11h). Each of the UBMs has 500 Gaussians. Using the two models, we employed the language weighting approach for obtaining several multi-accent UBMs. Thereafter, these UBMs were used to estimate the parameters of non-native SGMM systems. Subsequently, we trained multi-accent SGMMs with different numbers of substates, ranging from 800 to 8750. By doing this, we obtained several SGMMs for each multi-accent UBM applied. We summarize our results by reporting only the highest (maximum), average and best (minimum) SGMM results, as shown in Figure 2.

4.2 Results

Our findings show that using the proposed strategy resulted in significant improvement from the SGMM baseline. We reach the lowest WER when the SGMM system was obtained from a multi-accent UBM with 250 Gaussians from native and 250 Gaussians from non-native ($\alpha = 0.5$, WER=37.71%). This result proves that carefully controlling the contribution of two (unbalanced) data as sources for UBM training is a way to optimize ASR performance. In this experiment, the optimal α obtained tells us that non-native (Malaysian) data (in small quantity but very representative of the ASR task) and native (TED-LIUM) data (bigger corpus with speaker diversity) contribute equally to the acoustic space representation.

Furthermore, we did not gain WER improvements when the amount of substates increased. The minimum WERs shown in the figure are results for SGMMs with 800 substates. We extended our investigation to evaluate ASR performance for very compact (smaller number of Gaussians) UBM. The UBM

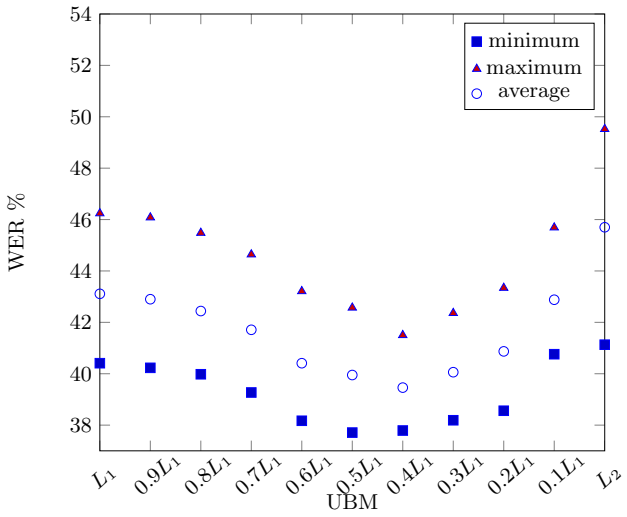

Fig. 2. *Min, max and average performance (WER (%)) of multi-accent SGMM based on language weighting strategy for non-native ASR (4h test). Note: non-native (L_1) native (L_2) and $\alpha = 0.1, ..., 0.9$.*

was built with only 50 Gaussians using native/non-native data and then we applied the same language weighting strategy to obtain multi-accent UBMs.

Table 3. *A summary of results from the SGMM experiments on the non-native ASR (4h test). Different UBMs were employed for building SGMM with 2h of non-native training data.*

SGMM	WER (%)
Non-native UBM500	40.41 (baseline)
Native UBM500	41.13
For $\alpha = 0.5$,	
a. Multi-accent UBM500	37.71
b. Multi-accent UBM50	34.24

The non-native system significantly improved after applying this method. Table 3 shows the comparison between multi-accent SGMM with UBM=500 and UBM=50. For $\alpha = 0.5$, the new multi-accent SGMM outperformed the one with more UBM Gaussians by 9% relative improvement on the WER. The result shows that deriving SGMM from a compact UBM gives better performance in very low-resource conditions. We also tried even smaller UBM but the WERs started to go back up (39.85% for UBM with 5 Gaussians!).

5 Accent-specific top layer for DNN

5.1 Proposed Method

Figure 3 illustrates the training process for obtaining an accent-specific DNN. We began with a network that was fine-tuned on native speech (last line and last column in Table 2). Then, we removed the softmax layer of native (source) DNN. Subsequently, a new softmax layer was added through fine-tuning the whole network on the non-native (target) training data. For this condition, we built the DNN on the GMM baseline for non-native.

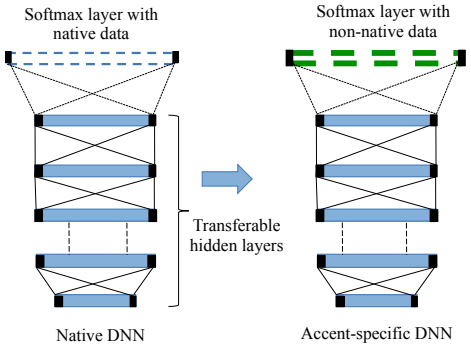

Fig. 3. *Process of obtaining accent-specific DNN (right) using hidden layers of native DNN*

We also built a second system for evaluating this approach. The system was speaker adapted and built upon new HMM/GMM acoustic models. First, we trained new native and non-native triphone models on new feature vectors using linear discriminant analysis (LDA) and maximum likelihood linear transform (MLLT), as well as speaker adaptive training using feature-space maximum likelihood linear regression (fMLLR). One important trick is to use feature transforms that were acquired from the native corpus (with large number of speakers), during LDA+MLLT training of non-native system. If not done this way, we observed no improvement with speaker adaptation (merging non-native and native DNNs with different feature transforms is not good). Then, we trained DNN for native and later we removed the top layer of the model. Subsequently, we fine-tuned the remaining DNN layers on the non-native data.

5.2 Results

We tested the DNNs on the same non-native evaluation data (4h test). Table 4 presents our findings. Both results are significantly better than the pure non-native DNN baseline (last line in Table 2). For example, we achieved 24% and 34% relative improvement respectively over the non-native DNN baseline (32.52%). Thus, the hidden layers of the native DNN proved to be useful for

Table 4. *WERs of accent-specific DNN on the non-native ASR task (4h test).*

DNN with accent-specific top layer	WER (%)
a. No speaker adaptation	24.89
b. Speaker adaptation	21.48

improving the low-resource non-native ASR. Besides that, our approach for building DNN with speaker adaptation and accent-specific top layer provided the best result. We obtained 14% relative improvement over the accent-specific DNN without speaker adaptation.

6 Conclusions

We have proposed two approaches for optimal merging of native and non-native data in order to improve accented ASR with limited training data. The first approach introduced a language weighting strategy for constructing multi-accent compact SGMM acoustic models. In this approach, we used language weights to control the number of Gaussians of each UBM involved in the merging process. Improvement of the ASR performance was observed with language weighting. The second approach involved fine-tuning the hidden layers of native DNN on the non-native training data. We applied this approach for obtaining accent-specific DNN with and without speaker adaptation. For the former, we trained the DNN on HMM/GMMs that had feature transforms of the native speech data. Both DNNs outperformed the DNN baseline. Overall, the approaches used in this study resulted in encouraging improvement in WER. Over the non-native baseline, we achieved relative improvement of 15% for SGMM (multi-accent UBM50) and 34% for DNN (accent-specific with speaker adaptation).

References

1. Arslan, M.J., Hansen, J.L.: A Study of the Temporal Features and Frequency Characteristics in American English Foreign Accent. Journal of the Acoustic Society (1996)
2. Bouselmi, G., Fohr, D., Haton, J.P.: Fully Automated Non-native Speech Recognition using Confusion-based Acoustic Model Intergration. In: Proceedings of Eurospeech. pp. 1369–1372. Lisboa (2005)
3. Chen, X., Cheng, J.: Deep Neural Network Acoustic Modeling for Native and Non-native Mandarin Speech Recognition. In: Proceedings of International Symposium on Chinese Spoken Language Processing (2014)
4. Goronzy, S.: Robust Adaptation to Non-native Accents in Automatic Speech Recognition. Springer (2002)
5. Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., Dean, J.: Multilingual Acoustic Models using Distributed Deep Neural Networks. In: Proceedings of ICASSP (2013)
6. Hinton, G., Deng, L., Yu, D., Mohamed, A.R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Dahl, T.S.G., Kingsbury, B.: Deep Neural Networks for Acoustic

Modeling in Speech recognition. IEEE Signal Processing Magazine 29(6), 82–97 (2012)
7. Hinton, G.E.: A Practical Guide to Training Restricted Boltzmann Machines. Utml tr 2010-003, Dept. Computer Science, University of Toronto (2010)
8. Huang, C., Chang, E., Zhou, J., Lee, K.F.: Accent Modeling based on Pronunciation Dictionary Adaptation for Large Vocabulary Mandarin Speech Recognition. In: Proceedings of ICLSP. vol. 2, pp. 818–821 (2000)
9. Huang, J.T., Li, J., Yu, D., Deng, L., Gong, Y.: Cross-language Knowledge Transfer using Multilingual Deep Neural Network with Shared Hidden Layers. In: Proceedings of ICASSP (2013)
10. Huang, Y., Yu, D., Liu, C., Gong, Y.: Multi-accent Deep Neural Network Acoustic Model with Accent-specific Top Layer using the KLD-regularized Model Adaptation. In: Proceedings of INTERSPEECH (2014)
11. Imseng, D., Motlicek, P., Bourlard, H., Garner, P.N.: Using Out-of-language Data to Improve Under-resourced Speech recognizer. Speech Communication 56(0), 142–151 (2014)
12. Kingsbury, B.: Lattice-based Optimization of Sequence Classification Criteria for Neural Network Acoustic Modeling. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 3761–3764 (April 2009)
13. Lu, L., Ghoshal, A., Renals, S.: Cross-lingual Subspace Gaussian Mixture Models for Low-resource Speech recognition. In: IEEE/ACM Transactions on Audio, Speech and Language Processing. vol. 22, pp. 17–27 (January 2014)
14. Miao, Y., Metze, F.: Improving Low-resource CD-DNN-HMM using Dropout and Multilingual DNN Training. In: Proceedings of INTERSPEECH. pp. 2237–2241 (2013)
15. Mohan, A., Ghalehjegh, S.H., Rose, R.C.: Dealing with Acoustic Mismatch for Training Multlingual Subspace Gaussian Mixture Models for Speech Recognition. In: Proceedings of ICASSP. pp. 4893–4896. IEEE, Kyoto (March 2012)
16. Morgan, J.J.: Making a Speech Recognizer Tolerate Non-native Speech through Gaussian Mixture Merging. In: Proceedings of ICALL'04. Venice (2004)
17. Povey, D., Burget, L., Agarwal, M., Akyazi, P., Feng, K., Ghoshal, A., Glembek, O., Goel, N., Karafiat, M., Rastrow, A., Rose, R.C., Schwarz, P., Thomas, S.: Subspace Gaussian Mixture Models for Speech Recognition. In: Proceedings of ICASSP (2010)
18. Povey, D., Burget, L., Agarwal, M., Akyazi, P., Kai, F., Ghoshal, A., Glembek, O., Karafiàt, N.G.M., Rastrow, A., Rose, R.C., Schwartz, P., Thomas, S.: The Subspace Gaussian Mixture Model - a Structured Model for Speech recognition. Computer Speech and Language 25, 404–439 (2011)
19. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Schwarz, P., Silovskỳ, J., Stemmer, G., Veselỳ, K.: The Kaldi Speech Recognition Toolkit. In: Society, I.S.P. (ed.) Proceedings of Workshop on Automatic Speech Recognition and Understanding. vol. IEEE Catalog No. : CFP11SRW-USB (December 2011)
20. Rousseau, A., Deléglise, P., Estève, Y.: TED-LIUM: An Automatic Speech Recognition Dedicated Corpus. In: Proceedings of LREC. pp. 125–129. European Language Resources Association (ELRA) (2012)
21. Swietojanski, P., Ghoshal, A., Renals, S.: Unsupervised Cross-lingual Knowledge Transfer in DNN-based LVCSR. In: Proceedings of ICASSP (2013)
22. Tan, T.P., Besacier, L.: Acoustic Model Interpolation for Non-native Speech Recognition. In: Proceedings of ICASSP (2007)

23. Tan, T.P., Besacier, L., Lecouteux, B.: Acoustic Model Merging using Acoustic Models from Multilingual Speakers for Automatic Speech Recognition. In: Proceedings of International Conference on Asian Language Processing (IALP) (2014)
24. Tong, R., Lim, B.P., Chen, N.F., Ma, B., Li, H.: Subspace Gaussian Mixture Models for Computer-assisted Language Learning. In: Proceedings of ICASSP. pp. 5347–5351. IEEE (2014)
25. Vu, N.T., Imseng, D., Povey, D., Motlíček, P., Schultz, T., Bourlard, H.: Multilingual Deep Neural Network based Acoustic Modeling for Rapid Language Adaptation. In: Proceedings of ICASSP (2014)