CROSS-LANGUAGE KNOWLEDGE TRANSFER USING MULTILINGUAL DEEP NEURAL NETWORK WITH SHARED HIDDEN LAYERS

Jui-Ting Huang[1], Jinyu Li[1], Dong Yu[2], Li Deng[2], and Yifan Gong[1]

[1]Online Services Division, Microsoft Corporation, Redmond, 98052, WA, USA
[2]Microsoft Research, Redmond, 98052, WA, USA
{jthuang, jinyli, dongyu, deng, ygong}@microsoft.com

ABSTRACT

In the deep neural network (DNN), the hidden layers can be considered as increasingly complex feature transformations and the final softmax layer as a log-linear classifier making use of the most abstract features computed in the hidden layers. While the log-linear classifier should be different for different languages, the feature transformations can be shared across languages. In this paper we propose a shared-hidden-layer multilingual DNN (SHL-MDNN), in which the hidden layers are made common across many languages while the softmax layers are made language dependent. We demonstrate that the SHL-MDNN can reduce errors by 3-5%, relatively, for all the languages decodable with the SHL-MDNN, over the monolingual DNNs trained using only the language specific data. Further, we show that the learned hidden layers sharing across languages can be transferred to improve recognition accuracy of new languages, with relative error reductions ranging from 6% to 28% against DNNs trained without exploiting the transferred hidden layers. It is particularly interesting that the error reduction can be achieved for the target language that is in different families of the languages used to learn the hidden layers.

Index Terms— deep neural network, CD-DNN-HMM, multilingual speech recognition, multitask learning, transfer learning

1. INTRODUCTION

The context-dependent deep neural network hidden Markov models (CD-DNN-HMMs) have outperformed the discriminatively trained conventional Gaussian mixture model (GMM) HMMs in many large vocabulary speech recognition (LVSR) tasks [1]-[11]. The DNN can be considered as a model that learns a complicated feature transformation (through many layers of nonlinearity in the hidden layers) and a log-linear classifier (through the softmax layer) jointly [4]. In most existing systems, the feature transformation determined by the hidden layers is learned from monolingual data.

In this paper we propose a shared-hidden-layer multilingual DNN (SHL-MDNN), in which the hidden layers are shared across many languages while the softmax layers are language dependent. The shared hidden layers (SHLs) and the separate softmax layers are jointly optimized using a multilingual training set. We can consider the SHLs as a universal feature transformation that works well for many languages.

The SHL-MDNN and its training procedure is an instance of the *multi-task learning* [12], with which multiple related tasks (LVSR systems for different languages) are trained simultaneously and benefit from each other. This implies that the SHL-MDNN can outperform the monolingual DNNs trained using only the language specific data, for all the languages decodable with the SHL-MDNN.

More interestingly, the universal feature transformation represented by the SHLs in the SHL-MDNN can be transferred to boost the performance of other (both resource-limited and resource-rich) languages not used in training the original SHL-MDNN. This is called *cross-lingual* model transfer and is a special case of the *transfer learning*. For a resource-limited language, we just reuse the SHLs from the SHL-MDNN and only tune the softmax layer, which can be stacked to the existing SHL-MDNN to allow it to recognize new language. For resource-rich languages, additional error reduction can be obtained by further adjusting the whole DNN. In either case, the training time is significantly reduced by initializing the model using the SHLs extracted from the SHL-MDNN.

We designed and conducted a series of experiments to evaluate the SHL-MDNN. We used four European languages to learn the SHL-MDNN and used English and Chinese as the target languages for cross-lingual model transfer. We demonstrate that the SHL-MDNN can reduce word error rate (WER) by 3-5% relatively, for all the four European languages used to train the SHL-MDNN, over the monolingual DNNs trained using only the language specific data. The SHL-MDNN also reduces WERs by 6%-28% relatively for English and Chinese LVSR systems through cross-lingual model transfer, with 3 to 100+ hours of target language training data, even though Chinese is very different from the European languages. While gains were only observed on resource-limited languages in most multilingual or cross-lingual research works, our results highlight the unique advantage of the SHL-MDNN that the performance of both resource-rich and resource-limited languages can be improved.

The rest of the paper is organized as follows. In Section 2 we describe the SHL-MDNN in detail and show that it can increase recognition accuracy for all languages used in training the SHL-MDNN. In Section 3 we illustrate the cross-lingual model transfer and its benefit on training LVSR systems for new languages. We discuss the related work in Section 4 and conclude the paper in Section 5.

2. SHARED-HIDDEN-LAYER MULTILINGUAL DNN

Figure 1 depicts the architecture of the proposed SHL-MDNN. In this architecture, the input and hidden layers are shared across all the languages the SHL-MDNN can recognize, and can be considered as a universal feature transformation (or front-end). The

softmax layers, however, are not shared. Instead, each language has its own softmax layer to estimate the posterior probabilities of the senones (tied triphone states) specific to that language.

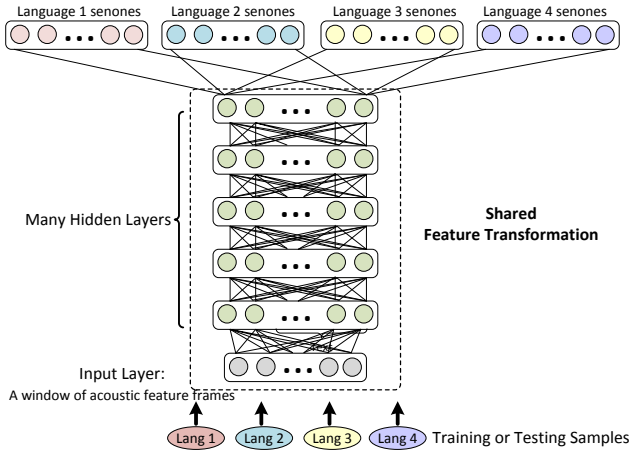

Figure 1: Architecture of the shared-hidden-layer multilingual DNN

As usual, the input layer covers a long contextual window of the acoustic feature (e.g., MFCC or log filter bank) frames. Since the shared hidden layers are to be used by many languages, language specific transformations such as HLDA cannot be applied. This requirement will not limit the performance of the CD-DNN-HMM, though, because any linear transformation can be subsumed by the DNN as indicated in [4].

The key to the successful learning of the SHL-MDNN is to train the model for all the languages simultaneously. When batch training algorithms, such as L-BFGS or the Hessian free algorithm [8], are used, this is trivial since all the data will be used in each update of the model. However, if mini-batch training algorithms, such as the mini-batch stochastic gradient ascent (SGA), are used, it means each mini-batch should be drawn from all the training data available. This can be efficiently accomplished by randomizing the training utterance list across the languages before feeding it into our DNN training tool.

The SHL-MDNN can be pretrained in either supervised or unsupervised way. In this study we have adopted the unsupervised pre-training procedure used in our previous study [1]. This is because the unsupervised pretraining does not involve the language-specific softmax layer and so can be carried out easily without any modification of our existing tool.

The fine-tuning of the SHL-MDNN can be carried out using the conventional backpropagation (BP) algorithm. However, since a different softmax layer is used for each different language, the algorithm needs to be adjusted slightly. When a training sample is presented to the SHL-MDNN trainer, only the shared hidden layers and the language-specific softmax layer are updated. Other softmax layers are kept intact. The SHLs serve as a structural regularization to the model and the entire SHL-MDNN and its training procedure can be considered as an example of multi-task learning.

After being trained, the SHL-MDNN can be used to recognize speech of any language used in the training process. By sharing the hidden layers in the SHL-MDNN and by using the joint training strategy, we can improve the recognition accuracy of all the languages decodable by the SHL-MDNN over the monolingual DNNs trained using data from individual languages only.

We evaluated the SHL-MDNN on a Microsoft internal speech recognition task. The training set contains 138-hour (hr) French (FRA), 195-hr German (DEU), 63-hr Spanish (ESP), and 63-hr Italian (ITA) speech data. The SHL-MDNN used in the experiment has 5 hidden layers, each with 2048 nodes. The input to the DNN is 11 (5-1-5) frames of the 13-dim MFCC feature with its derivatives and accelerations. For each language, the output layer has 1.8k senones determined by the GMM-HMM system trained with the maximum likelihood estimation (MLE) on the same training set. The SHL-MDNN was initialized using the unsupervised DBN-pretraining procedure, and then refined with BP using senone labels derived from the MLE model alignment. The trained DNNs are plugged in the CD-DNN-HMM framework designed for LVSR [1].

Table 1: Compare Monolingual DNN and Shared-Hidden-Layer Multilingual DNN in WER (%)

	FRA	DEU	ESP	ITA
Test Set Size (Words)	40k	37k	18k	31k
Monolingual DNN (%)	28.1	24.0	30.6	24.3
SHL-MDNN (%)	27.1	22.7	29.4	23.5
Relative WER Reduction (%)	3.6	5.4	3.9	3.3

Table 1 compares the word error rate (WER) obtained on the language specific test sets using the monolingual DNN (trained using only the data from that language) and the SHL-MDNN (whose hidden layers are trained using data from all four languages). From the table we can observe that the SHL-MDNN outperforms the monolingual DNN with a 3-5% relative WER reduction across all the languages. Note that when training monolingual DNNs, we shuffled the training utterances as well and adopted the same epoch numbers per language as in SHL-MDNN. Therefore, we ascribe the gain of SHL-MDNN to cross-language knowledge. It is encouraging that even for FRA and DEU, which have more than 100 hours of training data, SHL-MDNN can still provide improvement. This is not the only advantage of the SHL-MDNN. For example, since multiple languages are simultaneously decodable with its unified DNN structure, the SHL-MDNN makes multilingual LVSR easy and efficient.

3. CROSS-LINGUAL MODEL TRANSFER

The shared hidden layers (SHLs) extracted from the multilingual DNN can be considered as an intelligent feature extraction module jointly trained with data from multiple source languages. As such they carry rich information to distinguish phonetic classes in multiple languages and can be carried over to distinguish phones in new languages.

The procedure of cross-lingual model transfer is simple. We extract the SHLs from the SHL-MDNN and add a new softmax layer on top of it. The softmax layer's output nodes correspond to the senones in the target language. We then fix the hidden layers and only train the softmax layer using training data from the target language. If enough training data is available, additional gains may be achieved by further tuning the entire network.

To evaluate the effectiveness of cross-lingual model transfer, we used American English (ENU) (phonetically close to the

European languages used to train the SHL-MDNN) and Mandarin Chinese (CHN) (far away from the European languages) as the target languages and ran a series of experiments. The ENU test set consists of 2286 utterances (or 18k words) and the CHN test set has 10510 utterances (or 40k characters).

3.1. Hidden Layers Are Transferable

The first question is whether the hidden layers are transferable to other languages. To answer this question, we assume we have access to 9 hours of ENU training data (55737 utterances). We have several choices in building the ENU CD-DNN-HMM system. As shown in Table 2 the baseline DNN is trained solely using the 9-hr ENU training set. With this approach we only achieved a WER of 30.9% on the ENU test set. An alternative approach is to leverage the hidden layers (feature transformation) learned from other languages. In this experiment we chose to use 138 hours of FRA training data to train a monolingual DNN. We then extracted the hidden layers of this DNN to be used in the ENU DNN. If we fix the hidden layers and only train the ENU specific softmax layer using the 9-hr ENU training data we obtain absolute 2.6% WER reduction (30.9% → 27.3%) from the baseline DNN. If we retrain the whole FRA DNN using the 9-hr ENU data, we got a WER of 30.6%, which is only slightly better than the 30.9% baseline WER. These results indicate that the feature transformation represented by the hidden layers in the FRA DNN can be effectively transferred to recognize the ENU speech.

Table 2: Compare ENU WER with and without Using Hidden Layers (HLs) Transferred from the FRA DNN.

	WER (%)
Baseline (9-hr ENU)	30.9
FRA HLs + Train All Layers	30.6
FRA HLs + Train Softmax Layer	27.3
SHL-MDNN + Train Softmax Layer	25.3

We further transferred the shared hidden layers (SHLs) extracted from the SHL-MDNN described in Section 2 to train the ENU DNN. The last row in Table 2 indicates that the HLs extracted from the SHL-MDNN are more effective than that extracted from the FRA DNN when transferred to build the ENU DNN. In fact we got additional absolute 2.0% WER reduction (27.3% → 25.3%) by doing so. Overall, by using the cross-lingual model transfer we got 4.6% absolute (or 18.1% relative) WER reduction from the baseline ENU DNN.

3.2. Size of Target Language Training Data Matters

In this section we examine the effect of multilingual DNN cross-lingual model transfer when different sizes of target language training data (ENU, 3, 9 and 36 hours) are available. Table 3 summarizes the results. From the table, we can observe that by using the transferred SHLs, we can consistently outperform the baseline DNNs that do not use cross-lingual model transfer. We can also observe that when different sizes of target languages are available, the best learning strategy is different. In this experiment, we can observe that when less than 10 hours of target language training data are available, the best strategy is to only train a new softmax layer. By doing so we got 28.0% and 18.1% relative WER reduction over the baseline DNNs, when 3 and 9 hours of ENU speech data are available, respectively. However, when the amount of training data is large enough, further adapting the whole DNN can provide additional error reduction. For example, when 36 hours of ENU speech data are available, we got additional absolute 0.8% WER reduction (22.4% → 21.6%) by adapting all layers.

Table 3: Compare the Effect of Target Language Training Set Size in WER (%) when SHLs Are Transferred from the SHL-MDNN

ENU training data (#. Hours)	3	9	36
Baseline DNN (no Transfer)	38.9	30.9	23.0
SHL-MDNN + Train Softmax Layer	**28.0**	**25.3**	22.4
SHL-MDNN + Train All Layers	33.4	28.9	**21.6**
Best Case Relative WER Reduction (%)	28.0	18.1	6.1

3.3. Transferring to Mandarin Chinese Is Effective

To understand whether the effectiveness of the cross-lingual model transfer approach is sensitive to the language similarities between the source and the target languages, we used Mandarin Chinese (CHN) to simulate the second target language and applied the cross-lingual model transfer technique. Table 4 lists the character error rates (CERs) for both the baseline and the Multilingual-boosted DNN when the size of Chinese training data varies. We can see that in all cases CER reduction is observed by using the transferred SHLs. Even if we have 139 hours Of CHN data we can still benefit from the SHL-MDNN with 8.3% relative CER reduction. Moreover, using only 36 hours of CHN data we can achieve 28.4% CER on the test set by transferring the SHLs from the SHL-MDNN. This is better than the 29.0% CER obtained with the baseline DNN trained using the 139 hours of CHN training data, a save of over 100 hours of CHN transcription effort. To achieve the results reported in this table, we only trained the softmax layers when less than 9 hours of CHN data are available and further retrained all layers when more 10 hours of CHN data are available.

Table 4: Effectiveness of Cross-Lingual Model Transfer on CHN Measured in CER Reduction (%).

CHN Training Set (Hrs)	3	9	36	139
Baseline - CHN only	45.1	40.3	31.7	29.0
SHL-MDNN Model Transfer	35.6	33.9	28.4	26.6
Relative CER Reduction	21.1	15.9	10.4	8.3

3.4. Using Label Information Is Important

There is some evidence [13] in the computer vision community to suggest that features extracted using the unsupervised approach from a large amount of data are able to do classification tasks very well. This triggered some interests in the speech recognition community as it is much easier to obtain untranscribed speech data than transcribed ones for model training. Therefore, a related question is whether the label information is important for effectively learning the shared representation from the multilingual data. To answer this question, we compared the systems with and without using the label information when training the shared hidden layers. More specifically for the case without using the label information we used the multilingual DNN right after the pre-training stage. We see from Table 5 that while there is a small gain by using pre-trained only multilingual DNN and adapting the

whole network with ENU data (30.9% → 30.2%), the gain is significantly smaller than that obtained when label information is used (30.9% → 25.3%). These results clearly indicate that labeled data are much more valuable than unlabeled data and using label information is critical in learning effective features from multilingual data. Note that the gain we got from using unsupervised knowledge transfer is significantly smaller than that reported in [14]. We believe this is partly because we used much larger data sets than that used in [14] for both the source and target languages.

Table 5: Compare Features Learned from Multilingual Data with and without Using Label Information on ENU Data

	WER (%)
Baseline (9-hr ENU)	30.9
SHL-MDNN + Train Softmax Layers (no label)	38.7
SHL-MDNN + Train All Layers (no label)	30.2
SHL-MDNN + Train Softmax Layers (use label)	25.3

4. RELATION TO OTHER WORK

The motivation of this study is to use knowledge learned from multiple languages to improve the performance of each individual language. In the conventional GMM-HMM ASR framework, data sharing across languages has to occur at a certain level of acoustic units such as monophones [15][16] or states [15][17][18]. The cross-language mapping needs to be established either by manual rules (e.g., the IPA universal phone set [15][16]) or by data-driven clustering [15][16][17][18]. Instead of enforcing the hard mapping, which could inaccurately represent the phonetic structure, the SHLs in the SHL-MDNN provides a natural structure to be shared across languages. Among the GMM-HMM-based approaches subspace GMMs [19] might be the closest to DNN. Since majority of the subspace GMM parameters are shared across the states, they can be naturally trained by data from other resources or languages. Only those state-specific parameters are trained for language-specific models. However, the modeling power of Gaussian mixtures is significantly less than that of DNNs and so the features learned from the subspace GMMs would be much less selective and invariant than that from the DNNs.

The value of cross-lingual transfer has also been investigated in multilayer perceptrons (MLPs) under the tandem framework [20][21]. It has been shown that to build a resource-limited tandem ASR system, one can use the MLP trained with a resource-rich language to improve the recognition accuracy. Recently, people have found out that resource-rich languages can also benefit from cross-lingual MLP features. Plahl et al [22] changed the network topology to the bottleneck structure (BN) and showed that the English or Chinese cross-lingual MLP features (trained with more than 2000 hours of English or Chinese data) outperform the French-only BN features on the French test set, even for the case where 230 hours of French training data are available. This is consistent with our findings in the CD-DNN-HMM framework: the benefit of out-of-language data is not limited to low-resource languages, and the degree of kinship between the source and the target language becomes unimportant if the neural network is powerful enough. While only a single source language is used to build the cross-lingual features in [22], a similar work [23] combines multiple languages to train a single BN-MLP using IPA universal phone symbols. Alternatively in [24], the entire BN-MLP is trained in a sequential way: first training with a source language, then the top layer will be replaced with the phoneme set for another source language for another weight retraining. Our multilingual DNN is similar to [24] in that we also keep the distinct language-specific top layers. But we adopt a parallel training strategy where all utterances from multiple languages are shuffled so that all the languages are trained simultaneously. Our experiments show that the parallel training strategy is better than the sequential training strategy. Another difference between ours and that in [24] is that we used senones as the DNN outputs and used DNNs instead of shallow MLPs.

The idea of DNN based multitask learning has been applied to the field of natural language processing (NLP). A DNN that incorporate several related NLP tasks, such as part-of-speech tagging, chunking, semantic role labeling (SRL), and named entity tagging, was proposed and trained jointly on these tasks [25]. They show that with multitask learning, SRL achieved state-of-the-art performance with only word vectors as the input to the DNN, while the current NLP community considers syntax as a mandatory feature for the SRL task.

5. CONCLUSIONS

We proposed a shared-hidden-layer multilingual DNN architecture in which the hidden layers are shared across multiple languages and serve as universal feature transformation. In the context of the CD-DNN-HMM LVSR framework, we verified the effectiveness of the proposed SHL-MDNN by the improved WERs in all of the four European languages used in training the SHL-MDNN. We also demonstrated that the hidden layers in the SHL-MDNN can be effectively transferred for use by and benefit for other languages, even if large volumes of training data are available for the target language or the target language is phonetically far from the source languages used to train the SHL-MDNN.

The implication of this work is significant and far reaching. It suggests the possibility to quickly build a high-performance CD-DNN-HMM system for a new language from an existing multilingual DNN. This huge benefit would require a small amount of training data from the target language, although having more data would further improve the performance, can completely eliminate the unsupervised pre-training stage, and can train the DNN with much fewer epochs. Our work also indicates the possibility to build a universal ASR system efficiently under the CD-DNN-HMM framework. Such a system can not only recognize many languages and improve the accuracy for each individual language, but also expand the languages supported by simply stacking softmax layers for new languages.

In our current study, we only used four European languages to build the multilingual DNN. We believe further performance improvement can be achieved by using additional and more diversified languages to cover a wider range of phonetic variations. In this study we showed that when 36+ hours of target language training data are available we can obtain additional gain by further adjusting the full DNN. We believe this is an indication that the model size of the multilingual DNN, which is the same as that of the monolingual DNN in our study, should be expanded to model the greater variability observed in multiple languages. We plan to investigate all these in our future work.

6. REFERENCES

[1] G. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large vocabulary speech recognition," *IEEE Trans. Speech and Audio Proc.*, vol. 20, no. 1, pp. 30 – 42, 2012

[2] D. Yu, L. Deng, and G. Dahl, "Roles of pretraining and fine-tuning in context-dependent DNN-HMMs for real-world speech recognition," in *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, Dec. 2010.

[3] F. Seide, G. Li, and D. Yu, "Conversational speech transcription using context-dependent deep neural networks," in *Proc. Interspeech*, pp. 437-440, 2011.

[4] F. Seide, G. Li, X. Chen, D. Yu, "Feature engineering in context-dependent deep neural networks for conversational speech transcription," in *Proc. ASRU*, pp. 24-29, 2011.

[5] A. Mohamed, G. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22, 2012.

[6] N. Jaitly, P. Nguyen, and V. Vanhoucke, "application of pretrained deep neural networks to large vocabulary speech recognition," in *Proc. Interspeech*, 2012.

[7] T. N. Sainath, B. Kingsbury, B. Ramabhadran, P. Fousek, P. Novak, and A.-r. Mohamed, "Making deep belief networks effective for large vocabulary continuous speech recognition," in *Proc. ASRU*, pp. 30-35, 2011.

[8] B. Kingsbury, T. N. Sainath, and H. Soltau, "Scalable minimum bayes risk training of deep neural network acoustic models using distributed hessian-free optimization," in *Proc. Interspeech*, 2012.

[9] H. Su, G. Li, D. Yu, F. Seide, "Error back propagation for sequence training of context-dependent deep networks for conversational speech transcription", in *Proc. ICASSP* 2013.

[10] M. Seltzer, D. Yu, Y. Wang, "An investigation of deep neural networks for noise robust speech recognition", in *Proc. ICASSP* 2013.

[11] G. Hinton, L. Deng, D. Yu, G. Dahl, A.-R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. Sainath, and B. Kings- bury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, 2012.

[12] R. Caruana, "Multitask Learning," *Machine Learning*, Vol. 28, pp. 41-75, Kluwer Academic Publishers, 1997

[13] Q. Le, M. Ranzato, R. Monga, M. Devin, K. Chen, G. Corrado, J. Dean, A. Ng, "Building high-level features using large scale unsupervised learning," *International Conference in Machine Learning*, 2012

[14] P. Swietojanski, A. Ghoshal, S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proc. SLT* 2012.

[15] T. Schultz and A. Waibel, "Language independent and language adaptive acoustic modeling for speech recognition," in *Speech Communication*, August 2001, Volume 35, Issue 1-2, pp. 31-51

[16] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C-H Lee, "A study on multilingual acoustic modeling for large vocabulary ASR," in *Proc. ICASSP*, pp. 4333–4336, 2009

[17] T. Niesler, "Language-dependent state clustering for multilingual acoustic modeling," *Speech Communication*, vol. 49, 2007

[18] D. Yu, L. Deng, P. Liu, J. Wu, Y. Gong, A. Acero, "cross-lingual speech recognition under runtime resource constraints," in *Proc. ICASSP*, pp. 4193-4196, 2009

[19] L. Burget et al, "Multilingual Acoustic Modeling for Speech Recognition Based on Subspace Gaussian Mixture Models," in *Proc. ICASSP*, Dallas, 2010

[20] A. Stolcke, F. Grzl, M-Y Hwang, X. Lei, N. Morgan, D. Vergyri, "Cross-domain and cross-lingual portability of acoustic features estimated by multilayer perceptrons," in *Proc. ICASSP*, 2006

[21] S. Thomas, S. Ganapathy and H. Hermansky, "Cross-lingual and Multi-stream Posterior Features for Low Resource LVCSR Systems," in Proc. *Interspeech*, 2010

[22] C. Plahl, R. Schlueter and H. Ney, "Cross-lingual portability of Chinese and English neural network features for French and German LVCSR," in *Proc. ASRU*, USA, 2011

[23] N. Vu, W. Breiter, F. Metze, T. Schultz, "An investigation on initialization schemes for multilayer perceptron training using multilingual data and their effect on ASR performance," in *Proc. Interspeech*, 2012

[24] S. Thomas, S. Ganapathy and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proc. ICASSP*, 2012

[25] R. Collobert and J. Weston, "A unified architecture for natural language processing: deep neural networks with multitask learning," in *International Conference in Machine Learning*, 2008.