Context-dependent phone mapping for LVCSR of under-resourced languages

Van Hai Do[1,2], *Xiong Xiao*[2], *Eng Siong Chng*[1,2], *Haizhou Li*[1,2,3]

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]Temasek Laboratories@NTU, Nanyang Technological University, Singapore
[3]Institute for Infocomm Research, A*STAR, Singapore

{dova0001, xiaoxiong, aseschng}@ntu.edu.sg, hli@i2r.a-star.edu.sg

Abstract

This paper presents a context-dependent phone mapping approach for acoustic modeling of large vocabulary speech recognition for under-resourced languages by leveraging on well trained models of other languages. Generally speaking, phone mapping can be considered as a hybrid HMM/MLP (Hidden Markov Model / Multilayer Perceptron) model where the input of the MLP is phone acoustic scores, e.g. likelihood or posterior scores. In this paper, we use deep neural networks trained with a lot of Malay training data to generate bottleneck and posterior features for the target English acoustic models. We extend the concept of phone mapping by using not only posteriors but also bottleneck feature as the input for phone mapping. Experiments show that the phone mapping technique outperforms the cross-lingual tandem approach significantly. In addition, we also show that bottleneck and posterior features contain complementary information. A consistent improvement is obtained by combining these two feature streams to form the input for phone mapping.

Index Terms: context-dependent phone mapping, cross-lingual ASR, bottleneck feature, posterior feature, combination.

1. Introduction

Among thousands of spoken languages are used today, few of them are studied by the speech recognition community [1]. One of the major hurdles of ASR system deployment in new languages is that ASR system relies on a large amount of training data for acoustic modeling. Usually, to build a reasonable acoustic model for a large-vocabulary continuous speech recognition (LVCSR) system, tens to hundreds of hours of training data are required, which makes a full fledged acoustic modeling process impractical especially for under-resourced languages. This motivates us to investigate methods to automatically transfer well-trained acoustic models to under-resourced languages.

Various methods have been proposed for cross-lingual speech recognition such as universal phone set [2, 3], tandem approach [4–6], subspace GMMs (SGMMs) [7, 8]. Among these techniques, a group technique called cross-lingual phone mapping [9–13] is of our interest. In cross-lingual phone mapping, the target language speech data is first recognized into the phone sequences or phone posteriorgram of a source language, and then mapped to the phone sequences or posteriorgram of the target language. In [12, 13], we proposed a context-dependent phone mapping method for LVCSR and achieved significant improvement over monolingual models when only less than 1 hour of target language training data is available. In those studies, a multilayer perceptron (MLP) is used to map source acoustic scores (i.e. likelihood or posterior probabilities of phones or states) to the context-dependent states of the target language. In cases where insufficient training data are available for the target language, cross-lingual phone mapping may be more advantageous than the conventional acoustic model training method, due to the fact that it requires fewer data to train a phone-to-phone mapping system than to train a feature-to-phone mapping system from scratch. In cross-lingual phone mapping, the source acoustic model acts as a feature extractor that generates high-level and meaningful features for the mapping. This then allows the use of a simple mapping trained with very little data to map the source phones to the target phones.

Recently, bottleneck features [14] have been used widely in speech recognition and provide a consistent improvement over conventional features such as MFCCs, PLPs and posterior features as well. Bottleneck features are generated using an MLP with several hidden layers where the size of the middle hidden layer (bottleneck layer) is very small. With this structure, we can choose an arbitrary feature size without using dimensionality reduction step, independently on the MLP training targets. In [15, 16], bottleneck feature has been applied for cross-lingual ASR where a bottleneck MLP trained on other languages is used to generate bottleneck feature for an HMM/GMM model of the target language. They showed a potential ability of using cross-lingual bottleneck feature in speech recognition.

In this paper, we extend the concept of "phone mapping" by not only using cross-lingual phone based features (e.g. posteriors) but also bottleneck features as the input. In this case, the phone mapping framework can be considered as a cross-lingual hybrid HMM/MLP model [17] where the input feature is generated by models which are trained by another language (i.e. source language). We show in this paper that by using the hybrid HMM/MLP (mapping mode) for the target language we can achieve a significant improvement over the target HMM/GMM (tandem mode) in the case of limited training data even when the two models use the same input i.e. posteriors or bottleneck features. In addition, we also investigate the combination of bottleneck and posterior features in the cross-lingual phone mapping framework.

2. Context-dependent cross-lingual phone mapping

In our context-dependent cross-lingual phone mapping, a target language feature frame $\mathbf{o}_t$ is encoded by a vector of source acoustic scores $\mathbf{v}_t$ generated by the model trained on a source language. $\mathbf{v}_t$ can be source posterior probabilities given by a source HMM/MLP [13] or source likelihood scores given by a source HMM/GMM [12, 13]. In this paper, the concept "phone mapping" is extended by using bottleneck feature as the cross-

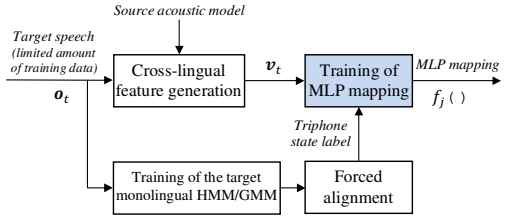

Figure 1: *A diagram of the training process for context-dependent phone mapping.*

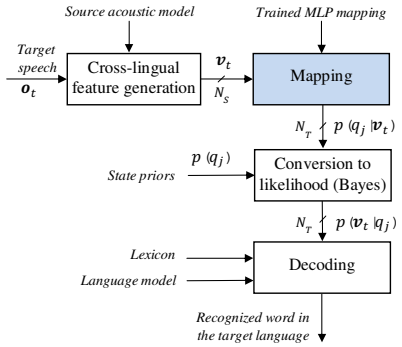

Figure 2: *A diagram of the decoding process using cross-lingual phone mapping.*

lingual input feature for phone mapping. $\mathbf{v}_t$ is then mapped into the tied states of context-dependent triphones in the target language:

$$p(q_j|\mathbf{v}_t) = f_j(\mathbf{v}_t), j = 1, ..., N_T \quad (1)$$

where N_T is the number of tied-states in the target language acoustic model. In our paper, the mapping function f is implemented by an MLP.

The training process of our cross-lingual phone mapping is illustrated in Fig. 1 and summarized in the following steps:

Step 1: Build the monolingual HMM/GMM target language acoustic model from the limited training data. Use decision tree to tie the triphone states to a predefined number. Generate the triphone state label for the training data using forced alignment.

Step 2: Evaluate the feature vectors $\mathbf{o}_t$ of the target language training data on the source model to generate cross-lingual feature vector $\mathbf{v}_t$. In this paper, we use deep neural networks (DNNs) [18] which are well trained by the source language data to generate posterior and bottleneck features, $\mathbf{v}_t$.

Step 3: Train the MLP mapping. Use $\mathbf{v}_t$ as the input of the mapping and the triphone state label generated in Step 1 as the target of the mapping.

The decoding process with a cross-lingual phone mapping acoustic model for LVCSR can be summarized as follows and illustrated in Fig. 2.

Step 1: Generate cross-lingual feature vector $\mathbf{v}_t$ for the test data in the same way as in Step 2 of the training procedure.

Step 2: Use the trained phone mapping to map $\mathbf{v}_t$ to the target language tied-state posterior $p(q_j|\mathbf{v}_t) = f_j(\mathbf{v}_t)$.

Step 3: Convert the target tied-states posterior to likelihood $p(\mathbf{v}_t|q_j)$ by normalizing them with their corresponding priors $p(q_j)$. The priors are obtained from the target training label.

Step 4: Use the state likelihoods, together with target language model and lexicon for Viterbi decoding.

3. Task and Systems

3.1. Task

To examine the effect of the proposed method, we use Malay - an Asian language as the source language and English (Aurora-4 task [19]) as the presumed under-resourced language. The Aurora-4 task has been chosen as the target under-resourced language as the effect of sufficient training data for it is well known, and we can hence clearly demonstrate the effect of reduced resources and our proposed work.

The source DNNs are trained with more than 100 hours of Malay read speech data [20]. While the target acoustic models are trained from a limited amount of English training data which is randomly selected from the clean training data of the Aurora-4 task. Aurora-4 is adapted from the Wall Street Journal (WSJ0) corpus. There are 7138 clean training sentences, or roughly 15 hours of speech data. We randomly select sentences from the 7138 sentences to generate the training sets of sizes 7 minutes, 16 minutes, and 55 minutes. For testing, we used the small clean test set of Aurora-4, which consists of 166 sentences, or 20 minutes of speech.

In this study, we concentrate on fast acoustic model training with a limited amount of speech data. We assume that the language model and pronunciation dictionary of the target language are available.

3.2. Cross-lingual feature generation systems

As shown in Fig. 1, 2, source acoustic models are used to generate cross-lingual feature vector $\mathbf{v}_t$ for target speech $\mathbf{o}_t$. In this section, we describe two pretrained deep neural networks (DNNs) trained from the source Malay corpus [20] and used to generate cross-lingual features for the target English language. DNNs with restricted Boltzmann machine (RBM) pretraining [18] are used in this study since it has been found that using pretrained DNNs is clearly superior to conventional MLPs in the tandem system [21], bottleneck feature [22] and the hybrid model [23–25]. The frame alignments are obtained from the monophone Malay HMM/GMM model after applying forced alignment. From the two DNN models, we generate three types of features for English.

3.2.1. Cross-lingual bottleneck feature

To generate bottleneck feature, a 5-layer deep neural network (BN-DNN) is used as shown in Fig. 3(a). 9 frames of 39-dimensional MFCCs are used to form the input for the BN-DNN. 102 outputs represent 102 monophone states in the Malay source acoustic model i.e. 34 phones x 3 states/phone. 39 hidden units are used at the bottleneck layer while the two other hidden layers contain 2000 units.

3.2.2. Cross-lingual posterior feature from BN-DNN

To compare with the bottleneck feature, we also use posterior probabilities generated by the same BN-DNN (Fig. 3(a)). These posteriors are used directly as the input feature for the target HMM/MLP (mapping mode) or taken log and applied principal component analysis (PCA) to decorrelate the feature when using for the target HMM/GMM (tandem mode).

3.2.3. Cross-lingual posterior feature from POS-DNN

In [26], the authors claimed that placing a bottleneck layer before the output can reduce the frame classification accuracy at the output layer. This may cause a performance degradation

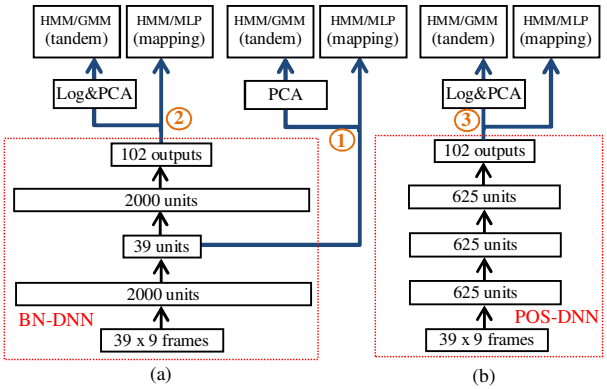

Figure 3: *Three different cross-lingual features for acoustic models of the target language.*

in cross-lingual models. In this paper, we build a DNN with 3 equal-sized hidden layers as shown in Fig. 3(b) to generate posterior probabilities for cross-lingual acoustic models. We call this DNN as POS-DNN which has 625 units for each hidden layer. With this structure, the number of parameters in the POS-DNN and the BN-DNN is similar, around 1 million.

4. Experiments

4.1. Experimental setup

Features: The features to train source neural networks and build monolingual acoustic models for the target language are the conventional 12^{th}-order MFCCs and C0, along with their first and second temporal derivatives. The frame length is 25ms and the frame shift is 10ms. To reduce recording mismatch between the source and target corpora, utterance-based mean and variance normalization (MVN) is applied to both training features of Malay and training and testing features of English.

Language model and dictionary: The standard Wall Street Journal English bigram language model is used in word recognition experiments. The test set contains a vocabulary of 5k words. The CMU dictionary is used and there are 40 monophones, including the silence phone.

Neural network training: Two 5-layer-DNNs are trained with the source language to generate bottleneck and posterior features. These DNNs are first initialized by RBMs [18]. For the first RBM (i.e. Gaussian-Bernoulli), we use a learning rate of 0.005 with 10 pretraining epochs. The rest RBMs (i.e. Bernoulli-Bernoulli), we use a learning rate of 0.05 with 5 pretraining epochs. After that the DNNs are fine tuned by the back-propagation algorithm. The "newbob"[1] procedure is used with the initial learning rate of 0.008.

We use 3-layer MLPs with 500 units to build the baseline monolingual hybrid target acoustic models and cross-lingual phone mapping. Our preliminary experiments show that using deeper structures does not bring any improvement over 3-layer MLPs. This may due to a limited amount of training data is not enough to train a big structure with many hidden layers.

Transition probabilities in HMM model: In the cross-lingual phone mapping and monolingual hybrid baseline acoustic models, for each HMM state, the probability of jumping to the next state is simply set to 0.5. The probability of remaining in the state is hence also 0.5.

[1] http://www1.icsi.berkeley.edu/Speech/faq/nn-train.html

4.2. Baseline monolingual acoustic models

In this subsection, we describe two baseline acoustic models for English, i.e. HMM/GMM and hybrid HMM/MLP models [17]. The experiments will show how much the performance of conventional acoustic modeling approaches will degrade if the training data is insufficient. As we indicated in our previous studies [12, 13], using context-dependent triphone acoustic models in the target language brings a significant improvement over the monophone models even in the case of very limited training data. In this paper, all of the target acoustic models are context-dependent triphone with 243 tied-states. The reason for using a relative small number of tied-states is that only less than one hour of training data is available for building the state-tying decision tree and for training the resulting triphone models.

The first row of Table 1 shows performance of the monolingual HMM/GMM models in word error rate (WER) with three different amounts of English training data i.e. 7, 16 and 55 minutes. We also build an HMM/GMM acoustic model using all 15 hours of training data in Aurora 4 and get a WER of 7.9%. It can be seen that the performance of the HMM/GMM model drops significantly when less training data is available.

To train the MLP in the hybrid HMM/MLP, MLP uses frame labels which are obtained from forced alignment of the HMM/GMM model above. We first use one frame of 39-dimensional MFCCs as the input of the MLP to compare with the HMM/GMM model when they use exactly the same input feature. As shown in the second row of Table 1, the hybrid HMM/MLP outperforms the corresponding HMM/GMM significantly especially when a small amount of training data is available. Now, we use more frames as the input for MLP as recent studies in hybrid systems [23–25]. The last row of Table 1 presents WER of the hybrid model when 9 frames of MFCCs are used. Surprisingly, using more frames as the input causes a performance degradation in the hybrid model. It may be due to that using more input frames can cause over-fitting for the MLP when an extremely small training size is used. When 55 minutes of training data is used, using 1 or 9 frames provides a similar performance for the hybrid model.

Table 1: The WER (%) of the monolingual acoustic models with three different amounts of English training data.

Acoustic model	Amount of training data		
	7 minutes	16 minutes	55 minutes
HMM/GMM	30.9	23.1	14.8
HMM/MLP (1 frame)	26.9	20.5	14.3
HMM/MLP (9 frame)	29.9	22.5	14.4

4.3. Cross-lingual acoustic models

In this subsection, we will investigate performance of the cross-lingual HMM/GMM (tandem mode) and HMM/MLP (mapping mode) in the case of only 7, 16 or 55 minutes of English training data is available.

The WER of the monolingual and cross-lingual models is shown in Table 2. The first row is the result of the monolingual models which use 39-dimensional MFCC feature as the input. The second row represents the WER of the target models which uses cross-lingual bottleneck feature generated by the Malay BN-DNN. We can see that using cross-lingual bottleneck feature achieves a significant improvement over MFCCs. Moreover, using bottleneck features with HMM/MLP (mapping

Table 2: The WER (%) of an English ASR system with monolingual and cross-lingual acoustic models trained on three different amounts of English training data.

No	Input feature		7 minutes		16 minutes		55 minutes	
			HMM/GMM (Tandem)	HMM/MLP (Mapping)	HMM/GMM (Tandem)	HMM/MLP (Mapping)	HMM/GMM (Tandem)	HMM/MLP (Mapping)
1	MFCCs(39)		30.9	26.9	23.1	20.5	15.2	14.3
2	Bottleneck(39)		24.6	17.5	18.5	15.3	12.6	11.5
3	POS from BN-DNN	PCA(39)	24.8	16.5	19.3	14.6	13.9	11.4
4		PCA(102)	27.9		21.2		16.1	
5	POS from POS-DNN	PCA(39)	23.9	16.1	17.3	13.8	12.2	10.4
6		PCA(102)	25.5		19.2		14.1	
7	BN + POS from BN-DNN		-	16.0	-	14.3	-	10.8
8	BN + POS from POS-DNN		-	15.5	-	13.4	-	9.9

mode) produces lower WER than using bottleneck features with HMM/GMM (tandem model) in all training sizes.

We also conduct cross-lingual experiments where the input feature is posterior probabilities (POS) generated by the same Malay BN-DNN. Before using this feature for the target HMM/GMM, PCA transform is apply to decorrelate the feature. We examine two cases: (i) keep all 102 dimensions after PCA, (ii) only 39 highest variance dimensions are kept after applying PCA. We can see the result in Table 2 that by keeping all 102 components after PCA (row 4), performance of the target HMM/GMM is worse than the model which uses only 39 dimensions (row 3). It proves that HMM/GMM model is hard to handle high dimensionality input even in the case of more training data is available (i.e. 55 minutes). However, in both cases, performance of POS feature is still lower than BN feature even they are generated by the same BN-DNN. It may be due to information loss during the dimensionality reduction step in the posterior feature.

On the other hand, by using the phone mapping approach i.e. hybrid HMM/MLP as the target acoustic model to map Malay posterior probabilities to English context-dependent states, we can use 102 source posteriors directly without using PCA. The WER of the phone mapping is listed in row (3-4) of Table 2. We can see a big improvement is observed over the corresponding cross-lingual tandem approach for all three different amounts of training data. This reason can come from finding the mapping between two phone sets is more straightforward than modeling source posteriors using Gaussian distributions in the case of limited training data. We also see that although using BN feature can get better performance than using POS feature in the tandem mode, in the mapping mode, POS outperforms BN feature. It shows that mapping from phones to phones may be easier than from BN feature to phones.

In the previous experiments, POS feature is generated from the BN-DNN. Now, we use the POS-DNN i.e. the 3 equal-sized hidden layer DNN (Fig. 3(b)) to generate POS feature for cross-lingual acoustic models. As shown in Table 3, the frame classification accuracy after back-propagation training of the Malay POS-DNN is consistently higher than the Malay BN-DNN for both the training and development sets although both the networks contain a similar number of parameters. Row 5 and 6 of Table 2 show the WER of cross-lingual acoustic models which use posterior probabilities generated by the POS-DNN as the input feature. We can see that better frame accuracy in the DNN of the source language results in a significant WER reduction for the cross-lingual acoustic models in both the tandem and mapping modes. For instance, with 16 minutes of English training data, the phone mapping model can achieve 13.8% WER which is 3.5% absolute better than the best cross-lingual tandem model.

Table 3: Frame classification accuracy of the Malay BN-DNN and the Malay POS-DNN.

Neural network topology	Frame accuracy (%)	
	Train set	Dev set
BN-DNN (351:2000:39:2000:102)	76.0	75.6
POS-DNN (351:625:625:625:102)	78.2	77.6

4.4. Combination of different cross-lingual feature streams

Bottleneck and posterior features represent different types of information of speech signal. While posterior feature represents speech signal at the phone state level, bottleneck feature represents data in a compact way by compressing speech data using its small bottleneck layer. Hence, it may be beneficial to combine these features in cross-lingual models. In this subsection, we combine bottleneck and posterior features by concatenating them to form the input for the target hybrid HMM/MLP (mapping mode). The result of the combined systems is shown in the last two rows of Table 2. We can see that a consistent improvement up to 0.5% absolute WER reduction is obtained by the combined models over the individual models.

5. Conclusion

In this paper, we applied the context-dependent phone mapping framework for a cross-lingual LVCSR task in the case of very limited training data. Experiments have been conducted to show that using the phone mapping framework can provide a significant improvement over the cross-lingual tandem approach even they use the same input feature. We also showed that bottleneck and posterior features contain complementary information even when they are generated by the same neural network. A consistent improvement is obtained by combining these two feature streams.

6. Acknowledgment

The authors would like to thank Prof. Nelson Morgan, Dr. Steven Wegmann, Dr. Adam Janin, Arlo Faria at the International Computer Science Institute, Berkeley, USA for allowing to use ICSI's computing resources and fruitful discussions on tandem, hybrid models and neural network implementation.

7. References

[1] T. Schultz and K. Kirchhoff, "Multilingual Speech Processing", 1st edition, Elsevier, Academic Press, 2006.

[2] T. Schultz and A. Waibel, "Experiments On Cross-Language Acoustic Modeling," in Proc. International Conference on Spoken Language Processing (ICSLP), 2001, pp. 2721-2724.

[3] N. T. Vu, F. Kraus, and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 5000–5003.

[4] A. Stolcke, F. Grezl, M. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2006, pp. 321-324.

[5] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multistream posterior features for low resource LVCSR systems," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), 2010, pp. 877-880.

[6] P. Lal, "Cross-lingual Automatic Speech Recognition using Tandem Features," Ph.D. thesis, The University of Edinburgh, 2011.

[7] L. Burget, et al. "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2010, pp. 4334-4337.

[8] L. Liang, A. Ghoshal, and S. Renals, "Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition," in Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2012, pp. 4877–4880.

[9] K. C. Sim and H. Li, "Context Sensitive Probabilistic Phone Mapping Model for Cross-lingual Speech Recognition," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), 2008, pp. 2715-2718.

[10] K. C. Sim and H. Li, "Stream-based Context-sensitive Phone Mapping for Cross-lingual Speech Recognition," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), 2009, pp. 3019-3022.

[11] K. C. Sim, "Discriminative Product-of-expert Acoustic Mapping for Crosslingual Phone Recognition," in Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), 2009, pp. 546-551.

[12] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "Context dependant phone mapping for cross-lingual acoustic modeling," in Proc. International Symposium on Chinese Spoken Language Processing (ISCSLP), 2012, pp. 16-20.

[13] V. H. Do, X. Xiao, E. S. Chng, and H. Li, "A Phone Mapping Technique for Acoustic Modeling of Under-resourced Languages," in Proc. International Conference on Asian Language Processing (IALP), 2012, pp. 233-236.

[14] F. Grezl, M. Karafiat, S. Kontar, and J. Cernock, Probabilistic and bottleneck features for LVCSR of meetings," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2007, vol. 4 pp. 757-760.

[15] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The Language-Independent Bottleneck Features," in Proc. IEEE Workshop on Spoken Language Technology (SLT), 2012, pp. 336-341.

[16] N. T. Vu, F. Metze, and T. Schultz, "Multilingual Bottle-Neck Features and its Application for Under-resourced Languages," in Proc. International Workshop on Spoken Languages Technologies for Under-resourced Languages (SLTU), 2012.

[17] H. Bourlard and N. Morgan, "Continuous Speech Recognition by Connectionist Statistical Methods," in IEEE Transactions on Neural Networks, vol 4, 1993, pp. 893-909.

[18] G. E. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," Neural Computation 18, 2006, pp. 1527-1554

[19] N. Parihar and J. Picone, "Aurora Working Group: DSR Front End LVCSR Evaluation AU/384/02," in Inst. for Signal and Infomation Process., Mississippi State Univ., Mississippi, Tech. Rep., 2002.

[20] X. Xiao, E. S. Chng, T. P. Tan, and H. Li, "Development of a Malay LVCSR System," in Proc. Oriental COCOSDA, 2010.

[21] O. Vinyals and S. V. Ravuri. "Comparing multilayer perceptron to Deep Belief Network Tandem features for robust ASR," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2011, pp. 4596-4599.

[22] D. Yu and M. L. Seltzer, "Improved bottleneck features using pretrained deep neural networks," in Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), 2011, pp. 237-240.

[23] V. H. Do, X. Xiao, and E. S. Chng, "Comparison and Combination of Multilayer Perceptrons and Deep Belief Networks in Hybrid Automatic Speech Recognition Systems," in Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2011.

[24] A-r. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," in IEEE Transactions on Audio, Speech, and Language Processing, Vol. 20, No.1, 2012, pp. 14-22.

[25] G. E. Dahl et al, "Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition." in IEEE Transactions on Audio, Speech, and Language Processing Vol.20, No.1, 2012, pp. 30-42.

[26] T. N. Sainath, B. Kingsbury, and B. Ramabhadran, "Auto-encoder bottleneck features using deep belief networks," in Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012, pp. 4153-4156.