Context-dependent Phone Mapping for Acoustic Modeling of Under-resourced Languages

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

[1]School of Computer Engineering, Nanyang Technological University
50 Nanyang Avenue, Singapore, 639798
[2]Temasek Laboratories@NTU, Nanyang Technological University, Singapore
[3]Institute for Infocomm Research, A*STAR, Singapore

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

Abstract

This paper presents the use of phone mapping for acoustic modeling of a language with limited training data. In this approach, we use well-trained acoustic models of a source language to generate acoustic scores for each feature vector of the target language. These scores are then mapped to the posteriors of context-dependent triphones of the target language using a limited amount of training data. In this paper, English is used as the source language while Malay is used as the target language. Experiments on a Malay large vocabulary continuous speech recognition (LVCSR) task show that with only a few minutes of training data we can achieve a low word error rate which significantly outperforms the best monolingual baseline acoustic model directly trained on the target language data. In addition, our study indicates that a consistent improvement is obtained when source acoustic scores are combined with speech attribute or bi-speech attribute posterior probabilities generated by the source attribute detectors to form the input for phone mapping.

Keywords
Speech recognition, LVCSR, under-resourced language, cross-lingual, phone mapping, speech attribute, multilayer perceptron.

1. Introduction

Although thousands of spoken languages are used today, few of them are focused in speech recognition community (Schultz and Kirchhoff, 2006). Typically, to build a good acoustic model for a large-vocabulary continuous speech recognition (LVCSR) system, hundreds to thousands hours of training data are needed. Obviously, it is very hard to meet this requirement for most languages. To overcome this problem, speech researchers try to transfer acoustic models which are well trained from another language to an under-resourced language (Schultz and Kirchhoff, 2006). This approach is called cross-lingual speech recognition which is illustrated in Fig. 1. The idea comes from the assumption that if

a person has a good ear (speech recognizer) for a source language, e.g. Vietnamese, he will be able to distinguish the sound units of another target language, e.g. English, although he does not understand that language. Hence, if we know the mapping rules from the sound units e.g. phones of Vietnamese to these of English, as illustrated in Fig. 1, we can recognize the English speech by using a Vietnamese ear and a mapping. In this paper, our study is about the mapping from source to target languages.

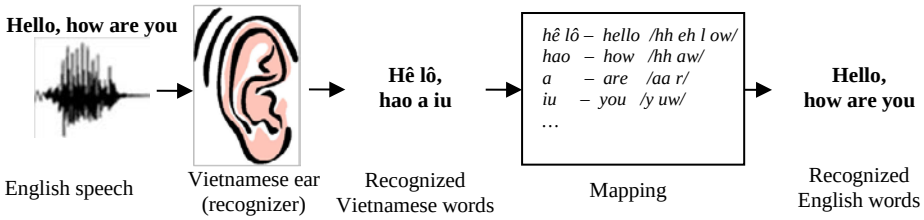

Fig 1. An illustration of cross-lingual speech recognition (Vietnamese to English case).

Schultz and Waibel, 2001 used a well-trained acoustic model of a source language to build the seed acoustic model for the target language. The mapping can be performed using either knowledge-based or data-driven method. This study showed the potential ability of using acoustic models of other languages to build an acoustic model of a target language. However, the mapping in (Schultz and Waibel, 2001) is a "hard mapping", that means a phone in the target language is mapped into a fixed phone of the source language.

Sim and Li, 2008 proposed a probabilistic phone mapping method to map a sequence of source phone symbols into a sequence of phones in the target language automatically using a maximum likelihood criterion, which is considered as a "soft mapping" technique. This method can work well even with a very limited amount of training data as the number of parameters in the mapping model is relatively small. The limitation of this method is that it uses the 1-best decoding results of the source recognizer as the input for mapping. That means we use only one recognized hypothesis and ignore other possibilities. Moreover, this method is also not suitable for large vocabulary tasks since in these cases the output of the source recognizer is not only affected by the source acoustic model but also by the source language model. Sim, 2009 used hybrid phone recognizers of source languages to produce source phone posterior probabilities instead of phone symbol sequences as suggested in (Sim and Li, 2008). After that these posteriors are mapped into monophone posteriors of the target language using a product-of-expert framework. This approach is more flexible than the phone symbol mapping since it uses phone posteriors instead of phone sequences as the input for the mapping. Experimental results in the NTIMIT corpus presented encouraging results for phone recognition.

In this paper, we propose a new "soft-mapping" method for large vocabulary tasks. In this method, conventional source Hidden Markov Model / Gaussian Mixture Model (HMM/GMM) and hybrid Hidden Markov Model / Multilayer Perceptron (HMM/MLP) (Bourlard and Morgan, 1993) acoustic models are used to recognize speech of the target language. Specifically, source acoustic models are used to generate acoustic scores, e.g. likelihoods of target language feature vectors given the source language phone class or source phone posteriors given target language features vectors. These scores are then mapped to the posteriors of context-dependent triphone states in the target language using multilayer perceptron neural networks (MLPs). In addition, in this paper we also investigate

the usefulness of speech attributes such as vowel, nasal, fricative in our phone mapping framework. The motivation is that speech attributes are more language independent than phones, so they are expected to be more robust in cross-lingual speech recognition. We also conduct a detailed comparison between different phone mapping topologies.

In this study, we use English as the source language and Malay, an Asian language as the target language. Experiments are conducted using only 8, 16 or 64 minutes of Malay training data which are randomly selected from a large vocabulary Malay corpus (Tan et al., 2009). In this study, we concentrate on 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.

The rest of the paper is organized as follows. Section 2 describes our proposed method for cross-lingual acoustic modeling. In this section, we also investigate application of speech attributes to improve our cross-lingual phone mapping recognition. Section 3 reports the experiments for the proposed phone mapping. Section 4 is the conclusion.

2. Acoustic Modeling with Cross-lingual Phone Mapping

2.1. Overview of Cross-lingual Acoustic Model

Speech units such as phones in different languages can be similar, they are however unlikely to be identical. Hence, it is not appropriate if we try to find a "hard mapping" between them. One speech unit in a language can have an overlapped distribution with several speech units of another language. Hence, "soft mapping" is a better option and can be interpreted as an attempt to interpolate a number of source sounds to a sound in the target language. Specifically, the distribution of the target HMM state model is a combination of HMM state distributions in the source language.

In this study, we use two types of source acoustic models to generate acoustic scores $\boldsymbol{v}_t^S$ for each speech feature vector $\boldsymbol{x}_t^T$ of the target language. The conventional source HMM/GMM model produces likelihood score $p(\boldsymbol{x}_t^T|s_i^S)$ while the hybrid source HMM/MLP model generates posterior score $p(s_i^S|\boldsymbol{x}_t^T)$ for each HMM state s_i^S in the source acoustic model. Note that we use the superscripts ^{T} and ^{S} to denote *T*arget and *S*ource languages, respectively. Since the source acoustic models are well-trained with a lot of training data, they can model well the distribution of speech units in the source language. If the distribution of the target and source speech units has a large overlap, we can use these scores, $\boldsymbol{v}_t^S$ to estimate speech units of the target language. In this paper, multilayer perceptron neural networks (MLPs) are used to map from the source acoustic scores to speech units of the target language. We use an MLP with the softmax activation function at the output layer so after training, it can model the state posterior probabilities of the target language $p(s_k^T|\boldsymbol{v}_t^S)$.

2.2. Context-dependent Modeling for Source and Target Acoustic Models

We extend Sim, 2009's use of source monophone acoustic models to source triphone acoustic models to generate acoustic scores. As the number of triphone states is much higher than the number of monophone states, triphone modeling will obtain more detailed input for mapping as compared to Sim, 2009.

In this study, we not only map source scores to monophone states in the target language, we also map to triphone states. Since the amount of training data in the target language is very limited, we used state tying strategy to reduce the number of states in the triphone target acoustic model.

The training process of our cross-lingual phone mapping is illustrated in Fig. 2 and summarized in the following steps:

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.
2. Evaluate feature vector $\boldsymbol{x}_t^T$ of the target language training data on the source model to generate cross-lingual feature vector $\boldsymbol{v}_t^S$ which can be likelihood scores or posterior probabilities.
3. Train the MLP mapping. Use $\boldsymbol{v}_t^S$ as the input of the mapping and the triphone state label generated in Step 1 as the target of the mapping.

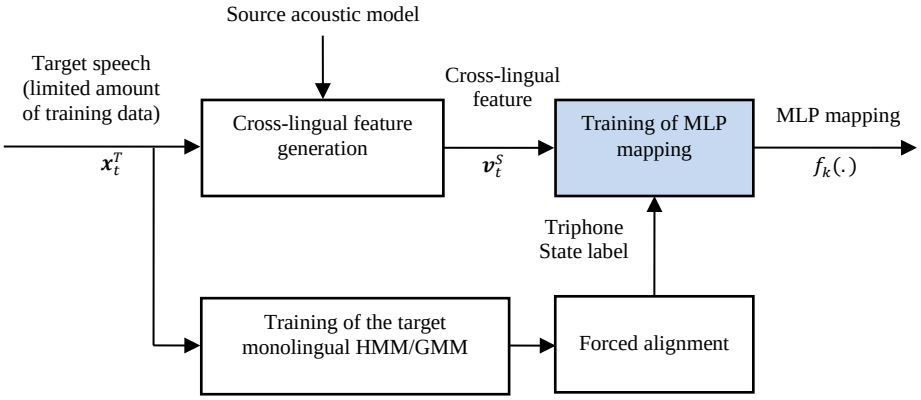

Figure 2. A diagram of the training process for context-dependent phone mapping.

The decoding process with a cross-lingual phone mapping acoustic model for LVCSR can be summarized as follows and illustrated in Fig. 3.

1. Generate cross-lingual feature vector $\boldsymbol{v}_t^S$ for the test data in the same way as in Step 2 of the training procedure.
2. Use the trained phone mapping to map $\boldsymbol{v}_t^S$ to the target language tied-state posterior: $p(s_k^T|\boldsymbol{v}_t^S) = f_k(\boldsymbol{v}_t^S)$.
3. Convert target tied-states posterior $p(s_k^T|\boldsymbol{v}_t^S)$ to likelihood $p(\boldsymbol{v}_t^S|s_k^T)$ by normalizing them with their corresponding prior $p(s_k^T)$. The priors are obtained from the target training label.
4. Use the state likelihoods, together with target language model and lexicon for Viterbi decoding.

2.3. Speech Attributes for Cross-lingual ASR

In the two previous subsections, we discussed about using source acoustic models which are used to generate phone based cross-lingual feature for mapping to the target language.

In this subsection, we extend the concept of phone mapping by using a special type of speech units called speech attributes.

Speech attributes (SAs) such as voicing, nasal, dental, describe how speech is articulated. They are also called phonological features, articulatory features, or linguistic features. SAs have been shown to be complementary to traditional features such as MFCCs or PLPs in automatic speech recognition (Kirchhoff, 1998; Metze and Waibel, 2002; Li et al., 2005; Do et al., 2011a). In addition, SAs are more language universal than phones and hence they can adapt to a new language easily (Lyu et al., 2008; Siniscalchi et al., 2012). In this paper, we also conduct experiments to demonstrate the usefulness of SAs in cross-lingual ASR which can help to improve the performance of our phone mapping framework. Specifically, SA detectors are trained on English and then applied for Malay speech.

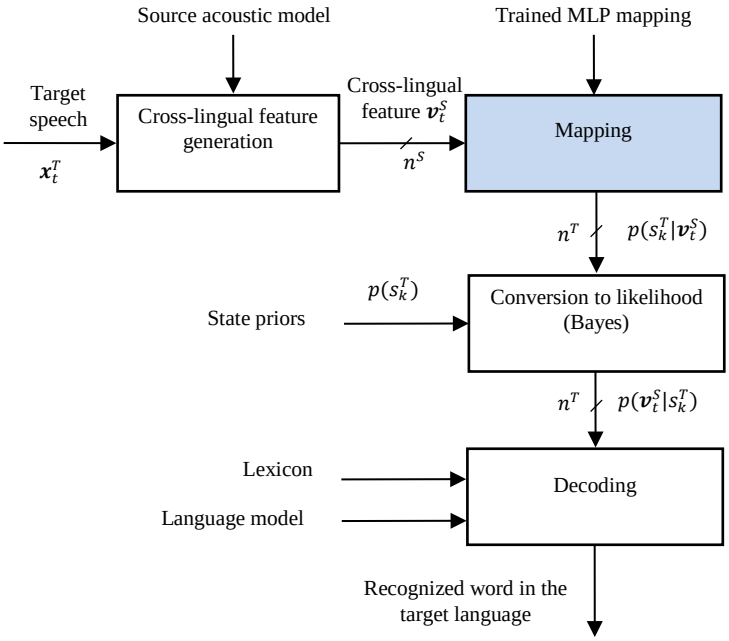

Figure 3. A diagram of the decoding process using cross-lingual phone mapping.

In this paper, we focus on classifying two groups of SAs. They are manners and places of articulation (Li et al., 2005) which include:

- Manners: vowel, stop, fricative, approximant, nasal, silence.
- Places: low, mid, high, dental, labial, coronal, retroflex, velar, glottal, silence.

Two SA detectors which realized by 3-layer-MLPs with 1000 hidden units are used to recognize manners and places. In our experiments, each SA is divided into three states which are similar to the 3-state-phone topology in ASR. To train the attribute detectors with MLPs, each speech frame in the training data needs a label, i.e. the ground truth of which manner and place classes the frame belongs to. These labels are obtained using the mapping from English phone set to SAs (Siniscalchi and Lee, 2009).

Similar to our cross-lingual phone mapping in Fig. 2 and Fig. 3, to apply SA recognition in cross-lingual tasks, we use SA detectors trained with English (i.e. source language) training data to recognize Malay (i.e. target language) speech to provide SA

posterior probabilities. These posteriors are then mapped to context-dependent Malay states as in our context-dependent phone mapping approach.

3. Experiments

3.1. Experimental Setup

Source acoustic models: Two popular types of English acoustic models are used which are trained with 16 hours of training data from the Aurora 4 corpus (Parihar and Picone, 2002). In the source HMM/GMM acoustic models, each state is modeled by a 16 mixture GMM while in the source hybrid HMM/MLP acoustic models, conventional 3-layer-MLPs with 1000 hidden units are used.

Target language corpus: In this study, Malay is used as the target language. A small amount of clean speech is randomly selected from the large vocabulary Malay corpus (Tan et al., 2009) and used as the training data. The 64-minute clean test set consists of 800 sentences.

Features: The features used in this study are the conventional 12^{th} order Mel frequency cepstral coefficients (MFCCs) including the C0, along with their first and second temporal derivatives. The frame length is 25ms and the frame shift is 10ms. To reduce acoustic mismatches across different recording environments, features are normalized to zero mean and unit variance over each utterance. The hybrid HMM/MLP systems use a 9-frame-window to build the input feature.

Language model and dictionary: The Malay tri-gram language model (Xiao et al., 2010) is used for word recognition. The test set contains a vocabulary of 50k words.

MLP training: To train the mapping MLP to generate the state posterior probabilities (Fig. 2), the training set is separated into two parts randomly. The first part around 90% is used to train the network weights. The rest part is used as the development set to prevent the network from over-fitting. The network weight set which produces the lowest frame error rate on the development set is selected. In all experiments for cross-lingual and hybrid baseline acoustic models, 3-layer MLPs with 500 hidden units are used. Better performance may be obtained if the MLPs structure is chosen carefully.

Transition probabilities for HMM model: In the cross-lingual and hybrid models, we simply set all transition probabilities which include the self-loop probability and the probability from the current state to the next state equal to 0.5. Further improvements can be obtained if they are borrowed from an existing acoustic model and tuned carefully.

3.2. The Baseline Monolingual Acoustic Models

Experiments for two baselines: conventional HMM/GMM and hybrid HMM/MLP (Bourlard and Morgan, 1993) with both monophone and triphone acoustic modeling are conducted with 8, 16 and 64 minutes of Malay training data. In the case of the monophone models, there are 102 states (i.e. 34 phones x 3 states/phone). To make a fair comparison in the triphone cases, with different amounts of training data we try to build the triphone models with the similar number of tied-states. We have three triphone models with 248, 249 and 250 tied states for the cases 8, 16 and 64 minutes of training data, respectively. The reason for using a relative small number of tied-states in the triphone models is that the

amount of training data is quite small. Table 1 shows the performance in word error rate (WER) of the conventional HMM/GMM and hybrid HMM/MLP models for monophone and triphone cases. Note that the number of Gaussian mixtures in the HMM/GMM models has been optimized to get the best performance. It can be seen that the performance of the two systems drops significantly when less training data is used. In addition, the triphone models outperform the corresponding monophone models significantly. It demonstrates the importance of context modeling in speech recognition. From the table, we also see that the hybrid HMM/MLP models perform slightly better than the corresponding HMM/GMM models for all cases.

No	Method	Amount of training data			
		8 minutes	16 minutes	64 minutes	100 hours
Monophone model					
1	HMM/GMM	26.4	22.7	17.1	-
2	Hybrid HMM/MLP	25.0	21.3	15.0	-
Triphone model					
3	HMM/GMM	21.6	18.1	13.6	7.4
4	Hybrid HMM/MLP	21.4	18.0	13.4	-

Table 1. The WER (%) of different monolingual acoustic models with different amounts of Malay training data.

Note that all results above are significantly worse than the result 7.4% obtained by the monolingual triphone HMM/GMM model which is trained with all 100 hours of Malay training data.

3.3. The Cross-lingual Acoustic Models

As shown in Fig. 3, in our cross-lingual acoustic models, feature vectors of the target language are passed through the source acoustic model which can be a conventional HMM/GMM or a hybrid HMM/MLP model to obtain n^S acoustic scores from n^S source HMM states. n^S can be 120 states for English monophone models or 1003 tied-states for English triphone models. These n^S scores are mapped to n^T states of the target language. n^T can be 102 states for the Malay monophone model or 249 tied-states for the Malay triphone model (for the case of 16 minutes of Malay training data).

Table 2 shows the results for word recognition with 16 minutes of Malay training data. The first two rows are the WER for the baseline monolingual models. The next two rows represent the result for the proposed cross-lingual models which use the conventional source HMM/GMM, and the last two rows are the WER of the cross-lingual models which use the hybrid source HMM/MLP to generate the acoustic scores. It can be seen that all cross-lingual models outperform the baseline models significantly with the same 16 minutes of training data. In addition, using triphone models in the target language provides a better performance than using monophone models.

No	Method	Target model	
		Monophone (n^T= 102)	Triphone (n^T= 249)
Baseline monolingual model			
1	HMM/GMM	22.7	18.1
2	Hybrid HMM/MLP	21.3	18.0
Proposed cross-lingual model (Source HMM/GMM)			
3	Source monophone (n^S = 120)	18.7	15.6
4	Source triphone (n^S = 1003)	17.4	15.3
Proposed cross-lingual model (Source HMM/MLP)			
5	Source monophone (n^S = 120)	17.8	15.3
6	Source triphone (n^S = 1003)	17.0	15.1

Table 2. The WER (%) of the different monolingual and cross-lingual acoustic models with 16 minutes of Malay training data.

From the table, we can also see that if we use source triphone acoustic model instead of using the source monophone acoustic model, we obtained an additional gain. This demonstrates that using context modeling in the source language can improve the performance of the cross-lingual model. In this experiment, we do not observe much difference in the acoustic scores between those generated by the source HMM/GMM and those by the source HMM/MLP acoustic model.

3.4. Cross-lingual Acoustic Models with Different Amounts of Training Data

We now further examine the effect of amounts of training data to the performance of different acoustic models. 8, 16 and 64 minutes of Malay training data randomly selected from the full training set are used for these experiments.

No	Method	Amount of training data		
		8 minutes	16 minutes	64 minutes
Baseline monolingual model				
1	HMM/GMM	21.6	18.1	13.6
2	Hybrid HMM/MLP	21.4	18.0	13.4
Proposed cross-lingual model (Source HMM/GMM)				
3	Source monophone (n^S = 120)	18.2	15.6	12.3
4	Source triphone (n^S = 1003)	17.3	15.3	11.4
Proposed cross-lingual model (Source HMM/MLP)				
5	Source monophone (n^S = 120)	18.0	15.3	12.3
6	Source triphone (n^S = 1003)	17.9	15.1	12.0

Table 3. The WER (%) of different acoustic models with different amounts of target training data (the target acoustic model is triphone).

In the previous two subsections, we have shown that the context-dependent modeling for the target language gives a significantly better performance than context independent

modeling for both the baseline monolingual models and the proposed cross-lingual models. In this subsection, we just focus on the context-dependent triphone systems in the target language.

Table 3 shows the performance of different acoustic models with three different amounts of training data. In all experiments, context-dependent modeling is used for the target language. The first two rows are the two baseline monolingual models. Row 3 and 4 represent the performance of the cross-lingual models where the conventional source HMM/GMM is used. The last two rows represent the WER of the cross-lingual models where the hybrid source HMM/MLP is used.

We can see that with the same amount of Malay training data, all proposed cross-lingual models outperform the baseline models significantly. Although the improvement of the proposed cross-lingual models over the monolingual models reduces when more target training data is available, even with using 64 minutes of the target training data, we can get a significant benefit from the well trained source acoustic model.

3.5. Speech Attributes and bi-Speech Attributes for Cross-lingual ASR

In this subsection, we conduct experiments to show the usefulness of SAs in our cross-lingual ASR framework. Similar to our phone mapping illustrated in Fig. 2 and Fig. 3, SA posterior probabilities estimated by the SA detectors are mapped to context dependent tied states of the target language by an MLP. The third row in Table 4 shows the results using SAs for cross-lingual ASR with different amounts of target Malay training data. Note that all experiments in this table are conducted using context-dependent triphones for the target language. It can be seen that generally using SAs alone provides lower performance over phone based approach i.e. the first and second row in Table 4. However, in row 4, and 5 where SA posteriors are concatenated with likelihood scores generated by the phone based source HMM/GMM model, consistent improvements are observed over the both individual systems. This demonstrates that SAs provide complementary information than phone based scores in cross-lingual tasks.

No	Input	Amount of training data		
		8 minutes	16 minutes	64 minutes
Cross-lingual model (Source HMM/GMM)				
1	Source monophone likelihood	18.2	15.6	12.3
2	Source triphone likelihood	17.3	15.3	11.4
Context independent speech attributes (mono-SAs)				
3	Mono-SAs	18.7	16.2	13.1
4	Mono-SAs+monophone likelihood (1)	17.7	15.0	11.8
5	Mono-SAs+triphone likelihood (2)	16.8	14.5	10.9
Context-dependent speech attributes (bi-SAs)				
6	Bi-SAs	18.1	15.9	11.5
7	Bi-SAs+monophone likelihood (1)	17.4	14.6	11.2
8	Bi-SAs+triphone likelihood (2)	16.6	14.3	10.6

Table 4. Speech attributes and bi-speech attributes in cross-lingual phone mapping.

In our previous study (Do et al., 2011a), we showed that SA and phone recognition can be improved by considering the left or right context of SAs called bi-SAs. In that paper, SA and bi-SA detectors were trained and tested in the same language. In this paper, we apply bi-SAs in cross-lingual ASR tasks. Specifically, SAs are expanded into bi-SAs by considering the left or right context of SAs. Four bi-SA detectors have been created which are: left-bi-manner, right-bi-manner, left-bi-place and right-bi-place. Give a Malay speech frame, each detector will generate posterior probabilities for each class of bi-SAs. These posteriors are combined and mapped into Malay tied states. The WER of using cross-lingual bi-SAs are presented in row 6 of Table 4. It is similar to Do et al., 2011a where bi-SAs can help to improve the performance of monolingual ASR tasks, bi-SAs also performs better over mono-SAs (row 3) in cross-lingual tasks. In addition, as shown in the last two rows in Table 4, a consistent improvement is achieved when bi-SA posteriors are combined with monophone or triphone likelihood scores.

3.6. Deep or Shallow Structures for Phone Mapping?

In all of our previous experiments, MLPs with 3 layers are used as the phone mapping. In this subsection, we will conduct a comparison between different mapping topologies for the case of under-resourced languages.

As shown in Fig. 2 and Fig. 3, in our cross-lingual phone mapping, the source acoustic model acts as a feature extractor to generate high-level and meaningful features (i.e. posteriors, likelihoods) for the mapping. This raises the question of whether we can use a simple mapping to perform this task, for example linear combination. On the other hand, recently, neural networks with many hidden layers have been applied successfully for speech recognition (Do et al., 2011b; Mohamed et al., 2012; Dahl et al., 2012). They show significant improvements over 3-layer NNs. This motivates us to investigate the ability of deep neural networks in phone mapping.

In this subsection, we conduct experiments which use three different types of neural network topologies:

- 2-layer neural networks i.e. without hidden layer. Linear activation function is used at the output layer.
- 3-layer neural networks i.e. with one hidden layer. 500 hidden units are used at the hidden layer.
- 4-layer neural networks i.e. with two hidden layers where each consists of 500 hidden units.

We examine three types of input for mapping which are:

- 351 MFCCs i.e. 9 frames x 39 dimensions.
- 120 English likelihood scores generated by the English monophone HMM/GMM model.
- 1003 English likelihood scores generated by the English triphone HMM/GMM model.

While context-dependent triphone states are used as the target of the mapping for all experiments.

Table 5 shows the phone mapping results for three different amounts of target Malay training data i.e. 8, 16 and 64 minutes. Note that in the case of MFCC input i.e. hybrid model, neural network is also considered as a mapping from input cepstral feature, i.e. MFCC, to HMM state posterior probabilities.

To compare experiment 1, 2, 3 with experiment 4, 5, 6 in Table 5, it can be seen that 3-layer NN outperforms the corresponding 2-layer MLP significantly. Especially, in the case MFCCs are used as the input. This can be explained that mapping from low-level feature such as MFCCs to state posteriors is more complicated than from source acoustic scores, it requires more powerful classifiers to accurately map MFCC to states of the target language. Although the mapping from source acoustic scores i.e. likelihoods to the target states is simpler, it benefits to use 3-layer NN even in the case of very limited training data.

Now, we examine whether deeper NNs can help to improve performance of phone mapping. The last three rows in Table 5 represents results for phone mapping using 4-layer NN. We can see that generally no improvement is achieved using deeper networks over conventional 3-layer NNs. One possible reason is that deep NNs can suffer from over-fitting in the case of under-resourced languages where a small amount of training samples may not be able to train a deep network with many hidden layers. We can see this reason in the last column of Table 5 when 64 minutes of training data is available, performance of 4-layer NNs approaches 3-layer NNs' performance while with fewer training data, deep NNs provide a consistently worse result.

No	Input	Amount of training data		
		8 minutes	16 minutes	64 minutes
2-layer neural network				
1	MFCC (351)	29.1	22.8	20.1
2	Source monophone likelihood (120)	19.6	17.7	14.6
3	Source triphone likelihood (1003)	19.5	16.3	13.4
3-layer neural network				
4	MFCC (351)	21.4	18.0	13.4
5	Source monophone likelihood (120)	18.2	15.6	12.3
6	Source triphone likelihood (1003)	17.3	15.3	11.4
4-layer neural network				
7	MFCC (351)	23.3	18.8	14.1
8	Source monophone likelihood (120)	18.5	16.6	12.2
9	Source triphone likelihood (1003)	18.3	15.8	11.8

Table 5. Comparison of different phone mapping topologies with 8, 16 and 64 minutes of Malay training data.

4. Conclusion and Discussion

In this paper, we proposed a new technique for cross-lingual acoustic modeling. Our method uses the acoustic scores generated by either a source conventional HMM/GMM or a source hybrid HMM/MLP acoustic models as the input to map to the context-dependent triphones in the target language. Experimental results showed that using a few minutes of training data with the proposed cross-lingual model can produce a low WER result which outperforms significantly the baseline monolingual models. In our experimental results, not much difference between using source HMM/GMM or HMM/MLP acoustic models was observed. That means we can use source conventional HMM/GMM models for cross-lingual speech recognition effectively. It is noted that HMM/GMM architecture is more

straightforward than HMM/MLP in terms of implementation. Moreover, we can gain from advanced techniques which have been developed for HMM/GMM model for a long time such as adaptation.

In this paper, we also investigate the usefulness of speech attributes as well and proposed context-dependent bi-speech attributes in cross-lingual ASR. The results showed that using SAs and bi-SAs can help to improve phone mapping consistently. We also examine phone mapping using different of neural network topologies. The result indicated that in the case of under-resourced ASR, using 3-layer NN topology provides better performance over both shallower topology i.e. 2-layer NNs and deeper structure i.e. 4-layer NNs.

REFERENCES

Bourlard, H., and Morgan, N., 1993, Continuous Speech Recognition by Connectionist Statistical Methods, in *IEEE Transactions on Neural Networks*, vol. 4, pp. 893–909.

Dahl, G. E., et al., 2012, Context-Dependent Pre-Trained Deep Neural Networks for Large-Vocabulary Speech Recognition, in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, pp. 30–42.

Do, V. H., Xiao, X., and Chng, E. S., 2011a, Speech Attribute Recognition using Context-Dependent Modeling, in Proc. *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*.

Do, V. H., Xiao, X., and Chng, E. S., 2011b, 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).*

Kirchhoff, K., 1998, Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments, in Proc. *International Conference on Spoken Language Processing (ICSLP)*, pp. 891–894.

Li, J., Tsao, Y., and Lee, C.-H., 2005, A study on knowledge source integration for candidate rescoring in automatic speech recognition, in Proc. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 837–840.

Lyu, D.-C., et al., 2008, Continuous phone recognition without target language training data, in *Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH)*, pp. 2687–2690.

Metze, F., and Waibel, A., 2002, A flexible stream architecture for ASR using articulatory features, in Proc. *International Conference on Spoken Language Processing (ICSLP).*

Mohamed, A., Dahl, G.E., and Hinton, G., 2012, Acoustic modeling using deep belief networks, in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 1, pp. 14–22.

Parihar, N., and Picone, J., 2002, Aurora Working Group: DSR Front End LVCSR Evaluation AU/384/02, in *Mississippi State Univ., Tech. Rep.*

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

Schultz, T., and Waibel, A., 2001, Experiments on Cross-Language Acoustic Modeling, in Proc. *International Conference on Spoken Language Processing (ICSLP)*, pp. 2721–2724.

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

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

Siniscalchi, S. M, et al., 2012, Experiments on Cross-Language Attribute Detection and Phone Recognition With Minimal Target-Specific Training Data, in *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 20, no. 3, pp. 875–887.

Siniscalchi, S., and Lee, C.-H. 2009, A study on integrating acoustic phonetic information into lattice rescoring for automatic speech recognition, in *Speech Communication*, vol. 51, no. 11, pp. 1139–1153.

Tan, T. P., Xiao, X., Tang, E. K., Chng, E. S., and Li, H., 2009, MASS: A Malay Language LVCSR Corpus Resource, in Proc. *O-COCOSDA*, pp. 25–30.

Xiao, X., Chng, E. S., Tan, T. P., and Li, H., 2010, Development of a Malay LVCSR System, in Proc. *O-COCOSDA*.