Cross-lingual phone mapping for large vocabulary speech recognition of under-resourced languages

Do, Van Hai; Xiao, Xiong; Chng, Eng Siong; Li, Haizhou

2014

DO, V. H., XIAO, X., CHNG, E. S., & LI, H. (2014). Cross-Lingual Phone Mapping for Large Vocabulary Speech Recognition of Under-Resourced Languages. IEICE Transactions on Information and Systems, E97.D(2), 285-295.

https://hdl.handle.net/10356/100818

https://doi.org/10.1587/transinf.E97.D.285

PAPER

Cross-Lingual Phone Mapping for Large Vocabulary Speech Recognition of Under-Resourced Languages

Van Hai DO[†,††a)], **Xiong XIAO**[††b)], ***Nonmembers***, **Eng Siong CHNG**[†,†††c)], ***and*** **Haizhou LI**[†,††,†††d)], ***Members***

SUMMARY This paper presents a novel acoustic modeling technique of large vocabulary automatic speech recognition for under-resourced languages by leveraging well-trained acoustic models of other languages (called source languages). The idea is to use source language acoustic model to score the acoustic features of the target language, and then map these scores to the posteriors of the target phones using a classifier. The target phone posteriors are then used for decoding in the usual way of hybrid acoustic modeling. The motivation of such a strategy is that human languages usually share similar phone sets and hence it may be easier to predict the target phone posteriors from the scores generated by source language acoustic models than to train from scratch an under-resourced language acoustic model. The proposed method is evaluated using on the Aurora-4 task with less than 1 hour of training data. Two types of source language acoustic models are considered, i.e. hybrid HMM/MLP and conventional HMM/GMM models. In addition, we also use triphone tied states in the mapping. Our experimental results show that by leveraging well trained Malay and Hungarian acoustic models, we achieved 9.0% word error rate (WER) given 55 minutes of English training data. This is close to the WER of 7.9% obtained by using the full 15 hours of training data and much better than the WER of 14.4% obtained by conventional acoustic modeling techniques with the same 55 minutes of training data.

key words: *speech recognition, under-resourced language, cross-lingual LVCSR, context-dependent, phone mapping*

1. Introduction

Automatic speech recognition (ASR) technology has made significant progress over the past few decades. Unfortunately, speech researchers have focused only on dozens out of thousands of spoken languages in the world [1]. One major obstacle to build an ASR system for a new language is that it is expensive to acquire a large amount of labeled speech data to train the acoustic model. To build a reasonable acoustic model for a large-vocabulary continuous speech recognition (LVCSR) system, usually tens to hundreds hours of training data are required, which makes a full fledged acoustic modeling process impractical, especially for under-resourced languages. This motivates studies to automatically transfer knowledge from acoustic models of resource-rich languages to under-resourced languages.

Various methods have been proposed to transfer acoustic knowledge from one language to another. An early study is called universal phone set [1], [2], which generates a common phoneme set by pooling the phoneme sets of all languages. A multilingual acoustic model is then trained from all languages by using the common phone set. To train the acoustic model for a new language, the acoustic model is bootstrapped from the multilingual acoustic model. This method can also work with unsupervised acoustic model training by using confident scores to select reliable sentences to train the model [3]. One drawback of the universal phone set approach is that it requires linguistic knowledge to build the common phoneme set of languages.

Another cross-lingual technique called cross-lingual tandem features operates on the feature level. In this approach, a multilayer perceptron neural network (MLP) trained on a source language is used to generate phone posterior probabilities on the target language [4]–[6]. These posteriors are then used as discriminative features for the conventional HMM/GMM (Hidden Markov Model / Gaussian Mixture Model) models of the target language. To improve performance, the source MLPs can be adapted to fit the target language better [5].

Besides improving the features, a novel acoustic model structure has also been proposed to reduce the amount of training data required for acoustic model training. In cross-lingual subspace GMM (SGMM) approach [7], [8], the parameters can be separated into two classes, i.e. the subspace parameters that are almost independent of languages, and phone state specific parameters that are language dependent. Hence, it is possible to borrow subspace parameters from well trained SGMM-based systems and train only the target language phone state specific parameters. As the phone state specific parameters only account for a small proportion of the overall parameters, they could be reliably trained with a small amount of training data.

In the techniques discussed so far, we still need to train the whole or part of the acoustic model from limited target language training data. However, in cross-lingual phone mapping approach [9]–[11], the acoustic model of a source language is used directly and only the mapping from source language phones to target language phones is learned. In other words, to recognize an under-resourced target language, speech data of the target language is first recognized into the phone sequences or phone posteriorgram of a source language using the well-trained source acous-

Manuscript received March 18, 2013.
Manuscript revised August 12, 2013.

[†]The authors are with the School of Computer Engineering, Nanyang Technological University, Singapore.
[††]The authors are with the Temasek Laboratories@NTU, Nanyang Technological University, Singapore.
[†††]The author is with the Institute for Infocomm Research, Singapore.
a) E-mail: dova0001@ntu.edu.sg
b) E-mail: xiaoxiong@ntu.edu.sg
c) E-mail: aseschng@ntu.edu.sg
d) E-mail: hli@i2r.a-star.edu.sg

DOI: 10.1587/transinf.E97.D.285

tic model. These phone sequences or posteriorgram produced by the source acoustic model are then mapped to the phone sequences or posteriorgram of the target language using a knowledge-based [9] or a data driven approach [10], [11]. Cross-lingual phone mapping is motivated by the fact that all human languages share similar acoustic space, i.e. most sound units e.g. phones are shared by different languages. Hence, a target language speech can be represented by a phone sequence/posteriorgram produced by another language for speech recognition purpose, if the acoustic spaces of the two languages are overlapping. In cases when there are insufficient training data for the target language, cross-lingual phone mapping may be more advantageous than the conventional acoustic model training method as it requires fewer data to train a phone-to-phone mapping system than to train a feature-to-phone mapping system from scratch. This can be explained as follows: for the cross-lingual phone mapping, the source acoustic model acts as a feature extractor to generate high-level and meaningful features for the mapping. This then allows the use of a simple mapping trained with little data to map the source phones to the target phones.

In this paper, we aim to build an LVCSR system for a language with very few training data (less than 1 hour). We extend the phone mapping framework in [11] with two major improvements. First, to retain sharp resolution in the acoustic space, we map source triphone states to target triphone states as opposed to mapping from source monophone states to target monophone states. We call this context-dependent cross-lingual phone mapping. Second, we also examine the use of source language's likelihood scores generated by a conventional HMM/GMM model for the mapping. This makes our approach more easily applicable to various types of available source acoustic models than [11] which uses only posterior probabilities generated by a hybrid HMM/MLP model. In addition, we also study the use of multiple source acoustic models in the cross-lingual phone mapping framework. In one scenario, two source models trained from the same source language training data are used together to generate scores for cross-lingual phone mapping. In another scenario, two source models from two different source languages are used. In both cases, significant improvements are achieved.

The rest of the paper is organized as follows. In Sect. 2, our proposed context-dependent phone mapping is described in details. In Sect. 3, we introduce the experimental setup, results, and discussions. Finally, we conclude in Sect. 4.

2. Cross-Lingual Phone Mapping

2.1 Prior Work on Cross-Lingual Phone Mapping

Several cross-lingual phone mapping methods have been studied in the past. In [2], Schultz and Waibel used a hard-decision phone mapping to build the seed model of the target language from a well-trained model of the source language. Le and Besacier [9] created a phone mapping based on expert knowledge. Each target language phone is mapped to a fixed source language phone. In [10], Sim and Li proposed a probabilistic phone mapping to map a source phone sequence to a target language phone sequence using a maximum likelihood criterion. This method works well with a limited amount of training data due to the small number of parameters. As a one-to-one mapping between phones of two languages is rare, it is desirable to have "soft-mapping" of phones rather than "hard-mapping". In [11], a soft-mapping method is proposed which maps source phone posteriorgram to the target phone posteriorgram. The use of phone posteriorgram avoids the loss of information due to the "quantization effect" of phone recognition as in [10]. The mapping of the source phone posteriors to the target phone posteriors is implemented using a product-of-expert method realized by an MLP.

2.2 Context Independent Cross-Lingual Phone Mapping

In this section, we first describe the context independent cross-lingual phone mapping proposed in [11]. In the next section, we will describe our proposed context-dependent mapping.

In cross-lingual phone mapping, the first step is to convert the speech data of the target language to either phone sequences [10] or phone posteriors [11] of the source language. Take monophone state posteriors [11] as an example, for the t^{th} frame of the target language speech, $\mathbf{o}_t$, a source posterior vector is generated in which each element represents the posterior probability of a source phone state given the speech frame, i.e. $p(s_i|\mathbf{o}_t)$, where s_i is the i^{th} phone state of the source language acoustic model. The source posterior vector is denoted as $\mathbf{u}_t = [p(s_1|\mathbf{o}_t), \ldots, p(s_{N_S}|\mathbf{o}_t)]^T$, where N_S is the number of states in the source language. Figure 1 illustrates a posteriorgram of an English sentence generated by a Malay MLP monophone recognizer. The x-axis is time while the y-axis represents phone states-ID of the Malay MLP recognizer. The intensity of each point in the posteriorgram illustrates posterior probability $p(s_i|\mathbf{o}_t)$.

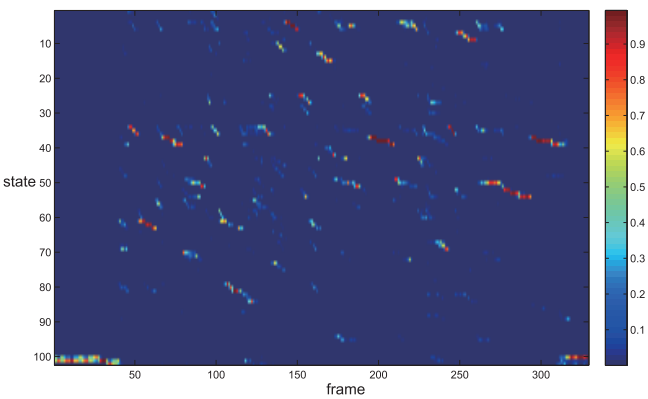

Fig. 1 Posteriorgram of an English utterance generated by a Malay MLP phone recognizer where x-axis is time, y-axis is Malay phone states-ID, intensity of each point represents posterior probability.

The representation $\mathbf{u}_t$ is then mapped to the phone state posteriors of the target language states:

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

where $f_j(\cdot)$ is a mapping function to be learned, q_j is the j^{th} state of the target language acoustic model, and N_T is the number of target phone states. The target phone posteriors are then converted to likelihoods using the Bayes formula and used for decoding.

In the previous study of cross-lingual phone mapping [11], monophone states are used as the class units in both the source and target languages. As a result, the acoustic resolution of the system is low and this may affect the performance of the cross-lingual system. In this paper, we adopt the above soft-mapping framework and made two improvements. One is to use triphone states rather than monophone states as the acoustic units in both the source and target languages. In this way, the resolution of the system is increased. Second, we not only use posteriors generated hybrid models, but also propose to use likelihood scores generated by a conventional source HMM/GMM model as the input for phone mapping. To the best of our knowledge, this is the first work where likelihood scores from the HMM/GMM acoustic model of a foreign language are successfully used to cross-lingual speech recognition. We will explain these two improvements in the following sections.

2.3 Context-dependent Cross-Lingual Phone Mapping

To build a high resolution acoustic model for the target language, the input representation of the acoustic space should be as detailed as possible. We know that monophone states are just a coarse representation of the acoustic space. A triphone acoustic system that takes phone context into consideration has a sharper acoustic resolution and has been widely used in conventional HMM/GMM-based LVCSR systems. Therefore, we propose to extend monophone mapping to triphone mapping between source and target languages. There are several advantages of using triphone states. One obvious advantage is that triphone-based acoustic models are easy to obtain as the mainstream acoustic model technology for LVCSR is based on triphone modeling. Well-trained triphone-based acoustic models of many popular languages can easily be obtained and used for cross-lingual mapping. Another advantage of using HMM/GMM model in the source language to generate acoustic scores is that, many acoustic modeling techniques, such as model adaptation, can be more easily applied to conventional HMM/GMM systems than to hybrid systems. Hence, cross-lingual phone mapping may potentially benefit from these existing techniques. For example, when we move from clean to noisy environments, we can adapt the source acoustic model to the noisy environments to reduce the acoustic mismatch.

In our proposed context-dependent cross-lingual phone mapping, a target language feature frame $\mathbf{o}_t$ is encoded into a vector of source acoustic scores, $\mathbf{v}_t$, which can be source likelihoods:

$$\mathbf{v}_t = [p(\mathbf{o}_t|s_1), \ldots, p(\mathbf{o}_t|s_{N_S})]^T \quad (2)$$

or source posteriors:

$$\mathbf{v}_t = [p(s_1|\mathbf{o}_t), \ldots, p(s_{N_S}|\mathbf{o}_t)]^T \quad (3)$$

where N_S is the number of tied-states in the source acoustic model, s_i in (2), (3) is the i^{th} tied-state in the source acoustic model. Similar to the monophone state mapping, the source triphone acoustic scores $\mathbf{v}_t$ is mapped to the target triphone tied-states by

$$p(q_j|\mathbf{v}_t) = f_j(\mathbf{v}_t), j = 1, \ldots, N_T \quad (4)$$

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 a 3-layer MLP. We will explain why 3-layer MLPs are used as the phone mapping in the next section.

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

Step 1 Build the conventional HMM/GMM baseline acoustic model from the limited training data of the target language. 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 vector $\mathbf{o}_t$ of the target language training data on the source acoustic model to generate the acoustic score vector $\mathbf{v}_t$ as in (2) or (3).

Step 3 Train the mapping MLP, $f_j(.)$ of Eq. (4). 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. 3.

Step 1 Generate source acoustic score vector sequences $\mathbf{v}_t$ 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 posteriors $p(q_j|\mathbf{v}_t)$.

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

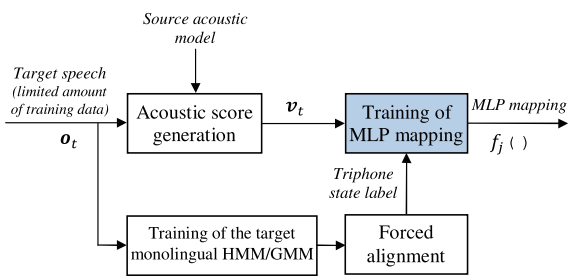

Fig. 2 A diagram of the training process for cross-lingual phone-mapping.

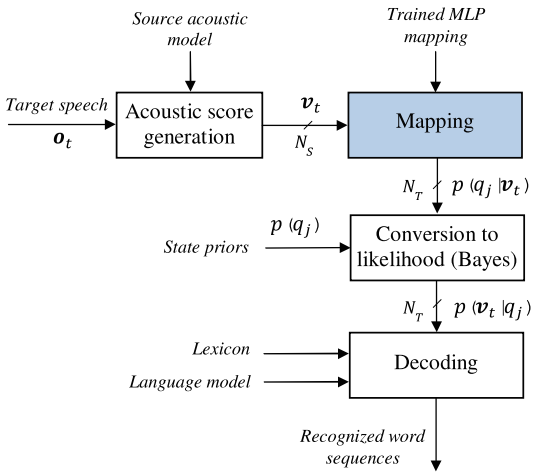

Fig. 3 A diagram of the decoding process using cross-lingual phone mapping.

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

2.4 Combination of Different Types of Inputs for Phone Mapping

Since our proposed method can handle different types of source acoustic models and different source languages as well to generate acoustic scores for phone mapping, an improvement can be obtained by combining these input streams if they provide complementary information. In this study, we investigate two levels of combination: feature combination and probability combination [18], [24]–[26].

The feature combination scheme is a simple and straightforward approach. In this approach, different features are concatenated to form a single input feature vector. For example, in the case when we use two types of source acoustic models i.e. HMM/GMM and HMM/MLP, likelihood scores generated by the source HMM/GMM and posterior scores generated by the source HMM/MLP models are concatenated to form the input of the mapping.

In the probability combination method, target state probabilities are combined using linear or nonlinear functions. In this paper, we simply combine multiple phone mapping models at the target probability level using the unweighted sum rule. Combined posterior probability $p_C(q_j|\mathbf{v}_t)$ of target state q_j given by input vector $\mathbf{v}_t$ is computed by taking the average value of all individual target state posterior probability $p_k(q_j|\mathbf{v}_t)$ estimated by the k^{th} phone mapping as illustrated in Eq. (5),

$$p_C(q_j|\mathbf{v}_t) = \frac{1}{N}\sum_{k=1}^{N} p_k(q_j|\mathbf{v}_t) \tag{5}$$

where N is the number of individual phone mappings.

3. Experiments

3.1 Tasks and Databases

To verify the performance the proposed cross-lingual phone mapping method, we use Malay - an Asian language as the source language and English (Aurora-4 task [12]) as the presumed under-resourced language. The Aurora-4 task adapted from the Wall Street Journal (WSJ0) corpus 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. In the Aurora-4 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. Malay-to-English phone mappings are trained from a limited amount of English training data. In addition, we also use Hungarian as a source language to investigate the effect of multiple source languages in phone mapping.

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 Experimental Setup

Source acoustic models: We evaluate two different Malay source acoustic models: conventional HMM/GMM model and hybrid HMM/MLP model. Both models are trained from more than 100 hours of Malay read speech data [13]. In the HMM/GMM model, triphone HMMs are used and the triphone states are clustered to 1592 tied states by using decision tree based clustering. The emission probability distribution of each tied state is represented by a GMM with 32 Gaussian mixtures. The hybrid HMM/MLP model uses the same HMM structure as the HMM/GMM model, and state posterior probabilities are estimated by a 3-layer-MLP with 2000 hidden units. Both HMM/GMM and HMM/MLP source acoustic models have about 4 million free parameters. We also trained two monophone based HMM/GMM and HMM/MLP source acoustic models with 102 monophone states for comparison purpose.

Besides Malay source models, we also experiment with Hungarian HMM/MLP monophone model.

Features: The features used in this study are the conventional 12^{th}-order Mel frequency cepstral coefficients (MFCCs) and C0 energy, along with their first and second temporal derivatives. The frame length is 25 ms and the frame shift is 10 ms. 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. The hybrid HMM/MLP model uses input feature vectors concatenated from 9 frames of MFCC features.

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 5 k words. The CMU dictionary are used which consists of 40 phones, including the silence phone.

MLP network training: To train MLP neural networks for both phone mapping and the monolingual hybrid baseline models to generate state posterior probabilities, the limited training set is separated into two parts randomly. The first part is used as the training data to update network weights and contains around 90% of the training set. The rest is used as the development set to prevent the network from over-fitting. The network weight set which produces the lowest frame error rate in the development set is selected (early stopping). In all experiments, 3-layer MLPs with 500 hidden units are used. Our study shows that the performance of the phone mapping is quite stable when 500 or more hidden units are used. Although the amount of parameters in the phone mapping neural network is quite large, the use of early stopping criterion prevents overtraining effectively.

Transition probabilities in HMM model: In the cross-lingual and 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.

3.3 Baseline Acoustic Models

In this section, we describe two baseline monolingual acoustic models for English, i.e. the HMM/GMM model and the HMM/MLP model. The experiments are carried to examine how the performance of conventional acoustic modeling is affected by insufficient training data. In addition, we conduct a cross-lingual tandem acoustic model for comparison purpose.

3.3.1 Monolingual HMM/GMM Acoustic Models

We build two baseline HMM/GMM acoustic models using 16 minutes of English training data, one is a monophone model and the other is a tied-states triphone model. In the monophone model, there are 120 states (i.e. 40 phones x 3 states/phone); while in the triphone model, there are 243 tied-states. The reason for using a relative small number of tied-states in the triphone model is that only 16 minutes of training data is available for building the state-tying decision tree and for training the resulting triphone models. The number of tied-states in the triphone model is chosen to be about twice the number of monophone states to evaluate the effect of context-dependent acoustic modeling.

Table 1 shows the performance of the monophone and triphone models with different model complexities. It is observed that the best triphone model (4 Gaussian mixtures) outperforms the best monophone model (8 Gaussian mixtures), although the two acoustic models contains comparable total number of parameters. The results show that triphone model is more robust than monophone model even when only a very limited amount of training data is available. The best WER obtained by triphone model is 23.1%, which is much higher than the 7.9% obtained by the triphone model with the full 15 hours of training data. This shows that conventional HMM/GMM system does not perform well under very small training size scenario.

Table 1 Word error rate (WER) (%) of the monolingual monophone and triphone baseline HMM/GMM models for 16 minutes of training data with different model complexities.

% Number of Gaussian mixtures	Monophone Model	Triphone Model
2	36.2	26.6
4	29.9	23.1
8	24.9	24.2
16	25.0	-

3.3.2 Monolingual Hybrid HMM/MLP Acoustic Models

We have also trained two English monolingual hybrid HMM/MLP models [14] using the same 16 minutes of training data to compare against the two models reported in Sect. 3.3.1. Hybrid HMM/MLP acoustic models offer several advantages over the HMM/GMM approach such as: MLPs are discriminative as compared to GMMs. Furthermore, HMM/MLP does not make parametric assumption about the distribution of inputs. The HMM/MLP approach has been applied successfully for phone recognition [15] and recently for word recognition [16].

In this experiment, MLPs are used to predict the posterior probabilities of the monophone states and triphone tied-states. The frame-level state labels used for MLP training are obtained from the HMM/GMM baseline models above. The WER for the hybrid monophone and triphone models are 24.6% and 22.5% (the second row of Table 2), respectively. These results show a slight improvement over the best corresponding HMM/GMM models.

3.3.3 Cross-Lingual Tandem Baseline

In this paper, we also build cross-lingual tandem systems which were proposed for resource-limited acoustic modeling. In the cross-lingual tandem approach, the source MLP recognizer is used to generate phone or state posterior scores for the target speech. These scores are then used as the feature for the target HMM/GMM in the tandem approach [4]–[6].

In our experiments, the Malay MLP is used to generate state posterior probabilities. The natural logarithm is applied on the posteriors to make them closer to the Gaussian distribution. As the dimensionality of posterior is usually high, principal component analysis (PCA) is used to project the log posterior vectors to 39-dimensional feature vectors. After that the posterior feature vectors are augmented with 39-dimensional MFCCs to form 78-dimensional vectors which are used as the input feature for an HMM/GMM

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

	Target model	
	Monophone (N_T = 120)	Triphone (N_T = 243)
Baseline monolingual acoustic model		
Monolingual HMM/GMM	24.9	23.1
Monolingual HMM/MLP	24.6	22.5
Baseline cross-lingual tandem acoustic model		
Source monophone	22.2	19.4
Source triphone	22.9	20.1
Proposed cross-lingual acoustic model (source HMM/GMM)		
Source monophone (N_S = 102)	20.6	18.3
Source triphone (N_S = 1592)	19.3	16.7
Proposed cross-lingual acoustic model (source HMM/MLP)		
Source monophone (N_S = 102)	19.6	17.6
Source triphone (N_S = 1592)	18.3	16.4

model.

In this paper, two types of source MLPs are used, i.e. monophone and triphone networks which have 102 and 1592 outputs, respectively. The result of the tandem approach with the two different types of source MLPs is shown in the third and forth rows of Table 2. It can be seen that both the two cross-lingual tandem models outperform the monolingual HMM/GMM and HMM/MLP models (the first two rows) significantly. These results demonstrate the benefit of using acoustic scores of a well-trained source acoustic model. However, using the source triphone MLP to generate tandem feature performs slightly worse than using the source monophone MLP. This can be explained as follows: although feature generated by the triphone MLP may contain richer information with higher resolution, it loses more information after the dimensionality reduction step from 1592 to 39 dimensions. In the case of under-resourced language, it is hard to increase the number of preserved dimensions because of the curse of dimensionality in the target HMM/GMM model with a limited amount of training data.

3.4 Cross-Lingual Phone Mapping Acoustic Models

Now we report the experiments of the proposed cross-lingual acoustic model trained on the same amount of training data as that in the baseline models on the Aurora-4 task. As shown in Fig. 3, 39-dimensional MFCC feature vector, $\boldsymbol{o}_t$ is passed through the source acoustic model to obtain N_S likelihood scores or N_S posteriors from the source HMM/GMM and HMM/MLP models, respectively. In this study, we examine both context independent and context-dependent source acoustic models:

1. The Malay monophone acoustic model with 102 states (i.e. 34 phones x 3 states/phone).
2. The Malay triphone acoustic model with 1592 tied-states.

These N_S scores are mapped to N_T states of the target language. N_T can be 120 states for the English monophone model or 243 tied-states for the English triphone model.

Table 2 shows the results for word recognition with 16 minutes of English training data. The first two rows are the WERs for the two monolingual baseline acoustic models. The next rows are the result of the cross-lingual tandem approach. The last four rows represent the results for the proposed cross-lingual acoustic models which use the HMM/GMM and hybrid HMM/MLP source models. From the table, we have four major observations.

First, all proposed cross-lingual phone mappings outperform the monolingual baseline models significantly. Our WERs are also considerably lower than the cross-lingual tandem approach although both approaches use source acoustic scores as the input feature. In this case, the phone mapping approach is more advantageous than modeling source acoustic scores by Gaussian distributions as in the tandem approach.

Second, by comparing the last four rows of Table 2, it is clear that using source triphone as the input of the cross-lingual phone mapping produces better results than using source monophone. This is due to the fact that the source triphones states provide more detailed representation of the target speech than the source monophone states. This result is contrary to the result on the cross-lingual tandem approach. It can be explained that MLP phone mapping can handle all inputs and does not lose information via the dimensionality reduction step as in the tandem approach. And hence, it can take the advantage of higher resolution feature generated by the source context-dependent model.

Third, by comparing the second and third columns of the table, it is observed that using target language triphone states as the label of the phone mapping consistently outperforms using target language monophone states. The best performance of the cross-lingual phone mapping is WER = 16.7% and 16.4% for the source HMM/GMM and source HMM/MLP, respectively. These results are obtained by using triphone representation in both the source and target languages.

Fourth, using the source hybrid HMM/MLP can produce a small improvement over the source conventional HMM/GMM. However, with the best configuration (i.e. triphone-to-triphone mapping), the performance of the two systems is almost the same. Also note that it took more than one week to train the source Malay hybrid triphone model while the training process of the conventional HMM/GMM was done in few hours.

In summary, the results in Table 2 shows that the cross-lingual phone mapping outperforms the conventional acoustic modeling techniques and the cross-lingual tandem features approach. In addition, the proposed context-dependent cross-lingual phone mapping produces significantly better results than the context independent cross-lingual phone mapping in [11].

3.5 Effect of Training Data Size

In this section, we will examine the effect of training data size on the performance of the cross-lingual phone mapping acoustic model. Three training data sizes are in our study,

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

Method	Amount of training data		
	7 minutes	16 minutes	55 minutes
Baseline monolingual acoustic model			
Monolingual HMM/GMM	30.9	23.1	14.8
Monolingual HMM/MLP	29.9	22.5	14.4
Baseline cross-lingual tandem acoustic model			
Source monophone	26.1	19.4	13.5
Source triphone	26.5	20.1	13.6
Proposed cross-lingual acoustic model (source HMM/GMM)			
Source monophone	21.7	18.3	13.4
Source triphone	20.3	16.7	11.7
Proposed cross-lingual acoustic model (source HMM/MLP)			
Source monophone	19.7	17.6	12.9
Source triphone	19.3	16.4	11.9
Combination of source HMM/GMM and source HMM/MLP (monophone)			
Feature combination	18.8	16.3	12.1
Probability combination	18.5	16.3	12.2
Combination of source HMM/GMM and source HMM/MLP (triphone)			
Feature combination	19.3	15.6	10.8
Probability combination	17.9	14.8	10.5

i.e. 7 minutes, 16 minutes, and 55 minutes of English target data.

In the previous sections, we have shown that the context-dependent triphone is a better choice than the monophone model for the target acoustic modeling. Therefore, in this section, the target language speech units are always triphone while the source language units could be either monophone or triphone. For each training data size, we follow the training steps in Sect. 2.3 to build the phone mapping. Note that we keep the number of triphone tied-states in the target language acoustic models to be always 243 for fair comparison.

Table 3 shows the performance of different acoustic models with three different amounts of training data. The first four rows are the monolingual and cross-lingual tandem acoustic models. We can see that the performance of all baseline models degrades quickly when less training data is used. The cross-lingual tandem approach outperforms the two monolingual models significantly for all data sizes. Also note that using source triphone MLP to generate posterior feature for the cross-lingual tandem approach does not help to improve performance over using the source monophone MLP. This result is also consistent with the recent research in [17] where the tandem approach was applied in a monolingual model. Using monophone states as the output representation of MLP is adequate to generate tandem feature. Increase the number of outputs (i.e. using tied-state triphone) generally does not help to improve even sometime it performs worse than using the monophone MLP.

The next four rows of Table 3 show the performance of our cross-lingual acoustic models which use the source HMM/GMM and hybrid HMM/MLP models. It is observed that the relative improvement of our cross-lingual phone mapping over the monolingual as well as the cross-lingual baselines increases as the amount of training data decreases. This shows that the proposed cross-lingual phone mapping is especially useful when a small amount of target training data is available. Although both our phone mapping and the cross-lingual tandem approaches use source acoustic scores as the input feature, modeling them by a mixture of Gaussian distributions in the tandem approach does not work well for the case of very limited training data. In this case, finding the "mapping" between the source phone set and the target phone set as in the proposed method is more effective.

We can also see that in case of source monophone models, using the source hybrid HMM/MLP can bring a small benefit over the source HMM/GMM. However, in the triphone cases, there are no much difference between the two source acoustic models.

Another observation is that using source triphone representation as the input of the mapping consistently outperforms using source monophone in all the training data sizes. However, this benefit reduces when less target training data is available. It can be explained that by using acoustic scores from the source triphone model, the number of inputs for mapping is much higher than those of the source monophone. When we have a extremely small amount of target training data, using source triphone may suffer from overfitting.

3.6 Combination of Different Types of Source Acoustic Models

As we discussed in Sect. 2.4, we can benefit from combining multiple input streams if they provide complementary information. In this section, the two phone mappings which use source HMM/GMM and source HMM/MLP respectively are combined at feature and probability levels. The result for the combined models is shown in the last four rows of Table 3. It is clearly shown that the combined systems outperform both the individual phone mappings significantly. Although the two source acoustic models are trained with the same data from the same language, the different model structures and different training criteria make information generated by the two acoustic models complementary. While the HMM/GMM is trained on a maximum likelihood criterion, the HMM/MLP models use a discriminative criterion to optimize the model parameters.

There is also a difference between the two combination methods. In case of both the two source acoustic models are monophone, the probability combination outperforms the feature combination method when a very small amount of training data is available i.e. 7 minutes. However, when we have more training data i.e. 55 minutes, the feature combination method performs better than the probability combination. In the case of both the two source acoustic models are triphone, combination at probability level gives a lower WER over combination at feature level especially when a very small amount of training data is available. It can be explained that although the feature combination approach seems a better choice for multi-stream input for MLPs, it may suffer from overfitting when a very small amount of training data is available. Especially, when source triphone

models are used, the number of inputs for the mapping is large if we concatenate multistreams at input level. In this case, the probability combination method can be a better option.

3.7 Using Multiple Source Languages

In this section, we investigate the performance of our phone mapping method with different source languages. We use a well-trained monophone hybrid MLP Hungarian phone recognizer downloaded from Brno University of Technology (BUT)† [19]. Given an English speech waveform, the Hungarian recognizer produces 186 monophone state posterior probabilities. These probabilities are mapped to 243 English tied states as in the Malay-to-English case. Note that BUT Hungarian recognizer uses 8-kHz sampling waveform as the input while our target English corpus (i.e. Aurora-4) is 16-kHz. We need to down-sample the target corpus to 8-kHz before applying to the Hungarian recognizer.

Table 4 shows the results of the two phone mapping systems with two different source languages. The first row is WER of the phone mapping with the source language Malay as in the previous sections. In this case source acoustic scores are generated by the Malay triphone HMM/GMM model. The result for Hungarian-to-English mapping is shown in the second row. It can be seen that when a very small amount of target training data is available (i.e. 7 and 16 minutes) using the Hungarian source acoustic model provides better performance than using the Malay model. One possible reason is that Hungarian and English are more similar than the pair Malay-English. As a result, the Hungarian-to-English phone mapping can be implemented easier and hence provides better performance even with an extremely small amount of target training data. However, when more training data is available, using the source Malay model gets better performance. This can be explained as the BUT Hungarian MLP is a monophone recognizer while the Malay HMM/GMM model is a triphone recognizer which provides higher resolution feature input for phone mapping. This will be useful when we have more target data to train the mapping.

The last two rows of Table 4 show the result when Malay-to-English and Hungarian-to-English mappings are combined at feature and probability levels. Interestingly, the combined models provide a big improvement over both the individual mappings. It demonstrates that the two models of the two source languages provide complementary information. In other words, combination of different source languages can give a better acoustic coverage for phone mapping. It is also noted that with 55 minutes of English training data, the best combined phone mapping can give 9.1% WER which is close to 7.9% WER of the monolingual model trained with the whole 15 hours of English training data.

There is a also a slight difference between the two combination approaches. While probability combination outperforms feature combination for the case of 7 minutes of training data, with bigger amounts of training data i.e. 16 and 55 minutes, feature combination is a better choice. Note that the dimension of posterior vectors generated by the Hungarian phone recognizer is only 186, which is much lower than the number of tied states in the Malay triphone model. Hence, it is less likely to have overtraining problem when 16 minutes or more English training data is available. This may explain why feature combination is better than probability combination in Table 4 and vice versa in Table 3 (the last two rows).

Table 4 The WER (%) of the proposed phone mapping for two source languages with different amounts of target training data (the target acoustic model is triphone).

Method	Amount of training data		
	7 minutes	16 minutes	55 minutes
Malay-to-English	20.3	16.7	11.7
Hungarian-to-English	18.2	16.1	12.3
Combined models			
Feature combination	15.0	13.1	9.1
Probability combination	14.7	13.3	9.4

†http://speech.fit.vutbr.cz/software/phoneme-recognizer-based-long-temporal-context

3.8 Discussion on Mapping Structure

In the previous sections, we showed that for the case of under-resourced language, our phone mapping method using MLPs is much more effective than the tandem approach which tries to model source acoustic scores using Gaussian distributions. It demonstrates that in this case choosing an appropriate model to relate the source acoustic scores to target phone posteriors is very important. In all of our previous experiments, 3-layer MLPs were used as the phone mapping from source language to the target language. In this section, we will explain why this MLP topology is chosen for phone mapping.

As we stated in the previous sections, source acoustic scores i.e. posteriors, likelihoods can be considered as higher level features as compared to conventional features such as MFCCs for speech recognition. This raises the question of whether we can use a simple mapping to perform this task, for example linear combination. To answer this question, we conduct phone mapping experiments using 2-layer neural networks (NN) (i.e. with no hidden layer) using linear activation function as the output layer. In the decoding phase, the output of NN is normalized using softmax function. In the case of monolingual hybrid HMM/MLP, the MLP is also considered as a mapping from the input cepstral feature i.e. MFCCs to HMM states.

Table 5 shows the phone mapping results with different inputs and different NN architectures for 16 minutes of target training data. Three different types of input are: (i) MFCCs with 351 inputs (i.e. 9 frames of 39-dimensional MFCC vectors), (ii) source monophone likelihoods with 102 inputs, (iii) source triphone likelihoods with 1592 inputs. 243 outputs are used to model 243 tied states in the target context-dependent acoustic model. The second column rep-

Table 5 The WER (%) of different mapping architectures with 16 minutes of target training data (the target acoustic model is triphone). Number in (.) in the first and the last column indicates the number of inputs and number of hidden units, respectively. Number in (.) of the third column represents relative improvement over the corresponding 2-layer NN.

Input	Mapping architecture		
	2-layer NN	3-layer NN (500 HUs)	3-layer NN (*)
MFCC (351)	30.1	22.5 (25.1%)	26.0 (144 HUs)
Source monophone (102)	21.6	18.3 (15.3%)	20.3 (72 HUs)
Source triphone (1592)	17.6	16.7 (5.0%)	16.9 (211 HUs)

resents WER using 2-layer NN. The third column is the result using 3-layer NN with 500 neurons in the hidden layer. The number in (.) represents relative improvement over 2-layer NN mapping. By comparing these two columns of Table 5, it is observed that in all three cases, 3-layer NNs outperform corresponding 2-layer NNs significantly, especially for the case of MFCC input. The results show that for low level MFCC features, 3-layer MLP performs much better than 2-layer NN due to that the system needs to be powerful enough to accurately map MFCC to states of the target language. For high level features such as source triphone scores, although the difference between the 2-layer and 3-layer networks is smaller, a more flexible 3-layer network is still preferred. Our result is also consistent with the result reported in [11] where the posterior weighted product-of-expert approach realized by a 3-layer NN outperformed the product-of-posterior model realized by a 2-layer NN in a cross-lingual phone recognition task.

Now we investigate whether the improvement of 3-layer NNs over 2-layer NNs comes from larger number of parameters or its 3-layer architecture. The result shown in the last column of Table 5 is obtained using 3-layer NNs those have the same number of parameters as the corresponding 2-layer NNs. Number in (.) represents number of hidden units in the 3-layer NN. It is observed that although the performance deteriorates when smaller hidden layers are used, 3-layer NNs perform better than 2-layer NNs even with the same number of parameters. One advantage of using 3-layer NNs is that while it is impossible to change number of parameters in 2-layer NNs, it is very easy to select a suitable model complexity for different phone mapping problems by changing number of hidden units in 3-layer NNs. Note that in all of our mapping experiments, we simply choose 3-layer NNs with 500 hidden units. We believe that further improvement can be obtained if this parameter is chosen carefully for each experiment.

To investigate in details the advantage of 3-layer NNs over 2-layer NNs for phone mapping, we extend the previous experiment by using 3 different amounts of target training data: 7, 16, and 55 minutes. It is observed in Table 6 that the 3-layer NN mapping outperforms the corresponding 2-layer NN mapping in all cases and the relative improvement increases consistently when we have more training data. It demonstrates that when more training data is available, a more powerful mapping is required. However, even in the case of extremely small amount of training data (i.e. 7 minutes), using 3-layer NN mapping brings more benefit.

Table 6 The WER (%) of 2 different phone mapping acoustic models with 3 different amounts of target training data (the target acoustic model is triphone). Percentages are indicated in (.) is the relative improvement of the 3-layer NN over the 2-layer NN acoustic model.

Input	Amount of training data		
	7 minutes	16 minutes	55 minutes
Mapping using 2-layer neural network			
MFCC (351)	36.9	30.1	23.5
Source monophone (102)	23.9	21.6	16.6
Source triphone (1592)	20.8	17.6	14.3
Mapping using 3-layer neural network (500 HUs)			
MFCC (351)	29.9 (19.1%)	22.5 (25.1%)	14.4 (38.9%)
Source monophone (102)	21.7 (9.2%)	18.3 (15.3%)	13.4 (19.7%)
Source triphone (1592)	20.3 (2.1%)	16.7 (5.0%)	11.7 (18.0%)

Recently, deep neural networks have been applied successfully for speech recognition [20]–[23]. They show significant improvements over 3-layer NNs. However in our preliminary experiments, deep neural networks do not help to improve the phone mapping performance. As we proved in this section mapping from acoustic scores of the source language to the target language is simpler than from raw features such as MFCCs. One possible reason is that in phone mapping, 3-layer NNs is powerful enough while deep neural network can suffer from over-fitting in the case of under-resourced language where a small amount of training samples may not be able to train a deep network with many hidden layers. We will investigate this issue in details in the future work.

3.9 Target Model Complexity Optimization

In the previous experiments, the number of tied states in the target acoustic model was kept fixed at 243 for comparison purpose. It is an appropriate choice when a small amount of training data is available such as 7 or 16 minutes. However, in the case we have more training data i.e. 55 minutes, higher number of tied states may improve the performance. In this section, we re-run all experiments for the case of 55 minutes of training data with different numbers of tied states.

Table 7 shows the WERs of different models with three different numbers of tied states in the target acoustic model i.e. 243, 501 and 1003. The first two rows are results given by the two monolingual baseline models. The WER of the cross-lingual tandem system is reported in the next row, while the next three rows present the WERs of our phone mapping approach. The last four rows show the performance of combined models with different combination schemes. From the table, it can be observed that although the performance of different systems varies differently for different number of states generally, with 55 minutes of training data, using 501 tied states provides lowest WERs for almost all models compared with the cases of 243 and 1003 states. Combined models outperform individual models significantly for all three cases. The feature combination method gives a lower WER in the case when the triphone

Table 7 The WER (%) of different acoustic models with 55 minutes of target training data for 3 different number of tied states in the target acoustic model.

Method	Number of tied states		
	243	501	1003
Baseline monolingual acoustic model			
HMM/GMM	14.8	14.0	13.9
HMM/MLP	14.4	14.1	13.9
Baseline cross-lingual tandem acoustic model			
Tandem	13.5	12.9	13.4
Propose cross-lingual acoustic model			
Triphone Malay HMM/GMM	11.7	11.0	11.8
Triphone Malay HMM/MLP	11.9	11.3	12.4
Monophone Hungarian HMM/MLP	12.3	11.4	11.3
Malay HMM/GMM and Malay HMM/MLP combination			
Feature combination	10.8	10.4	10.9
Probability combination	10.5	9.8	11.0
Malay HMM/GMM and Hungarian HMM/MLP combination			
Feature combination	9.1	9.0	9.2
Probability combination	9.4	9.5	9.3

Malay HMM/GMM and monophone Hungarian models are combined. While in the case of combination triphone Malay HMM/GMM and HMM/MLP models, the probability combination method outperforms the feature combination approach except in the 1003 states case.

4. Conclusion

In this paper, we proposed a context-dependent cross-lingual phone mapping for fast training of acoustic model for under-resourced languages. Our experimental results verified the effectiveness of the proposed phone mapping technique for building LVCSR models. There are two advantages in our method, i.e. the use of triphone states for improved acoustic resolution of both the source and target models and the ability of using various types of source acoustic models which results in an additional improvement when combining them even they are trained from the same data. Our paper also indicated that combination of different source languages can significantly improve the performance of phone mapping as we may have a better acoustic coverage for the target language. In this work, we conclude that using a "phone mapping" is a better choice than modeling source acoustic scores by a mixture of Gaussian distributions in the case of under-resourced speech recognition.

In the paper, we showed that although the mapping from acoustic scores of the source language to the target language is simpler than from low level features such as MFCCs, using 3-layer neural networks as the mapping can achieve better results over weak mappings such as linear combination even for the case of very limited amount of target training data. Our preliminary results indicated that using deeper structure mapping does not help to improve the performance. However, it is an interesting issue for future research as deep neural networks can handle unlabeled training data effectively in the unsupervised pre-training process [23]. We can benefit from using unlabeled training data from many sources.

Cross-lingual phone mapping is a relatively new topic and many aspects of the technique are not well known yet. For example, how do we measure the similarity of the acoustic spaces of two languages and how does this similarity affect cross-lingual phone mapping performance. We will examine these questions in future work.

References

[1] T. Schultz and K. Kirchhoff, Multilingual Speech Processing, 1st ed., Elsevier, Academic Press, 2006.
[2] T. Schultz and A. Waibel, "Experiments on cross-language acoustic modeling," Proc. International Conference on Spoken Language Processing (ICSLP), pp.2721–2724, 2001.
[3] T. Vu, F. Kraus, and T. Schultz, "Cross-language bootstrapping based on completely unsupervised training using multilingual A-stabil," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.5000–5003, 2011.
[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," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.321–324, 2006.
[5] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and multistream posterior features for low resource LVCSR systems," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.877–880, 2010.
[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," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4334–4337, 2010.
[8] L. Liang, A. Ghoshal, and S. Renals, "Maximum a posteriori adaptation of subspace Gaussian mixture models for cross-lingual speech recognition," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4877–4880, 2012.
[9] V-B. Le and L. Besacier, "Automatic speech recognition for under-resourced languages: Application to vietnamese language," IEEE Trans. Audio Speech Language Process., vol.17, no.8, pp.1471–1482, Nov. 2009.
[10] K.C. Sim and H. Li, "Context sensitive probabilistic phone mapping model for cross-lingual speech recognition," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.2715–2718, 2008.
[11] K.C. Sim, "Discriminative product-of-expert acoustic mapping for crosslingual phone recognition," Proc. IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), pp.546–551, 2009.
[12] N. Parihar and J. Picone, "Aurora working group: DSR front end LVCSR evaluation AU/384/02," Inst. for Signal and Infomation Process., Mississippi State Univ., Mississippi, Tech. Rep., 2002.
[13] X. Xiao, E.S. Chng, T.P. Tan, and H. Li, "Development of a malay LVCSR system," Proc. Oriental COCOSDA, 2010, pp.25–30.
[14] H. Bourlard and N. Morgan, "Continuous speech recognition by connectionist statistical methods," IEEE Trans. Neural Netw., vol.4, pp.893–909, 1993.
[15] P. Matejka, P. Schwarz, and J. Cernocky, "Towards lower error rates in phoneme recognition," TSD, Brno, Czech Republic, 2004.
[16] A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "Context dependent modelling approaches for hybrid speech recognizers," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), pp.2950–2953, 2010.
[17] Z. Tuske, M. Sundermeyer, R. Schluter, and H. Ney, "Context-dependent MLPs for LVCSR: TANDEM, hybrid or both?," Proc. Annual Conference of the International Speech Communication As-

sociation (INTERSPEECH), 2012.
[18] K. Kirchhoff, "Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments," Proc. Fifth International Conference on Spoken Language Processing, 1998.
[19] P. Schwarz, P. Matejka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.325–328, 2006.
[20] A. Mohamed, G.E. Dahl, and G. Hinton, "Deep belief networks for phone recognition," Proc. NIPS 2009 Workshop on Deep Learning for Speech Recognition and Related Applications, 2009.
[21] A. Mohamed, G.E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," IEEE Trans. Audio Speech Language Process., vol.20, no.1, pp.14–22, 2012.
[22] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," IEEE Trans. Audio Speech Language Process., vol.20, no.1, pp.30–42, 2012.
[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," Proc. Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 2011.
[24] X. Li, "Combination and generation of parallel feature streams for improved speech recognition," PhD thesis, Carnegie Mellon University, 2005.
[25] X. Cui, J. Xue, B. Xiang, and B. Zhou, "A study of bootstrapping with multiple acoustic features for improved automatic speech recognition," Proc. Annual Conference of the International Speech Communication Association (INTERSPEECH), 2009.
[26] C. Ma, H. Kuo, H. Soltau, X. Cui, U. Chaudhari, L. Mangu, and C.-H. Lee, "A comparative study on system combination schemes for lvcsr," Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp.4394–4397, 2010.

Van Hai Do received the B.Eng. and M.Sc. degrees in Electronics and Telecommunications from the Hanoi University of Science and Technology, Vietnam, in 2002 and 2006, respectively. Since August, 2009 he is pursuing his Ph.D. at the School of Computer Engineering, Nanyang Technological University, Singapore. His research focuses on hybrid acoustic models, cross-lingual speech recognition for under-resourced languages. From September, 2012 to March, 2013 he was at the International Computer Science Institute (ICSI), Berkeley, USA as an attachment. At ICSI, he contributed to the BABEL project which aims to develop keyword search capability for under-resourced languages rapidly.

Xiong Xiao received his B.Eng. degree and PhD degree in 2004 and 2010, respectively, both from the School of Computer Engineering, Nanyang Technological University (NTU), Singapore. Since 2009, he has been a research staff in the speech team of Temasek Lab@NTU, where he is now a senior research scientist. His research interests include robust speech recognition, pattern recognition, and signal processing.

Eng Siong Chng is currently an Associate Professor in the School of Computer Engineering (SCE), Nanyang Technological University (NTU), Singapore. Concurrently, he is the deputy director of Emerging Research Lab (ER Lab@SCE) in the same school. Prior to joining NTU in 2003, he has worked in several leading research centers/companies, namely: Knowles Electronics (USA), Lernout and Hauspie (Belgium), Institute of Infocomm Research (I2R, Singapore), and RIKEN (Japan). He received both PhD and BENG (Hons) from Edinburgh University, Scotland, in 1996 and 1991 respectively. His area of focus is in speech research and signal processing. To date, he has received numerous external research grants as principal investigator with a total funding amount of S$3.4 million for the Speech and Language Technology Program (SLTP) at SCE. His publications include 2 edited books and over 100 journal/conference papers. He has graduated 4 PhD students. He has served as the publication chair for 3 international conferences (APSIPA-2010, APSIPA-2011, ISCSLP-2006), and has served as program committees, session chairs, and organizers in many technical sessions at various international conferences. He has been an associate editor for IEICE (special issue 2012), a reviewer for Speech Communications, Eupsico, IEEE Trans Man, System and Cybernectics Part B, Journal of Signal Processing System, ACM Multimedia Systems, IEEE Trans Neural Network, IEEE Trans CAS-II, and Signal Processing. Dr Chng is the recipient of the Tan Chin Tuan fellowship (2007) to visit Tsinghua University, the JSPS travel grant award (2008) to visit Tokyo Institute of Technology, and the Merlion Singapore-France research collaboration award in 2009.

Haizhou Li received the B.Sc., M.Sc., and Ph.D. degrees in electrical & electronic engineering from the South China University of Technology (SCUT), Guangzhou, in 1984, 1987, and 1990, respectively. Dr Li was a Research Assistant from 1988 to 1990 at the University of Hong Kong, Hong Kong, China. In 1990, he joined SCUT as an Associate Professor. From 1994 to 1995, he was a Visiting Professor at CRIN, Nancy, France. In 1995, he became the Manager of the ASR group at the Apple-ISS Research Centre in Singapore where he led the research of Apple's Chinese Dictation Kit for Macintosh. In 1999, he was appointed Research Director of Lernout & Hauspie Asia Pacific. From 2001 to 2003, he was the Vice President of InfoTalk Corp. Ltd. Since 2003, he has been with the Institute for Infocomm Research (I2R), Singapore, where he is now the Principal Scientist and Head of Human Language Technology Department. His current research interests include automatic speech recognition, speaker and language recognition and natural language processing. Dr Li was named one of the two Nokia Visiting Professors 2009 by the Nokia Foundation in recognition of his contributions to Speaker and Language Recognition research. He was a recipient of the National Infocomm Award 2002 in Singapore. He is now an Associate Editor for Springer International Journal of Social Robotics, IEEE Transactions on Audio, Speech and Language Processing, Computer Speech and Language, and ACM Transactions on Speech and Language Processing. He is also an elected Board Member of International Speech Communication Association (2009–2013).