CONTEXT DEPENDANT PHONE MAPPING FOR CROSS-LINGUAL ACOUSTIC MODELING

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

[1]School of Computer Engineering, Nanyang Technological University, Singapore
[2]School of Electrical Engineering and Telecommunications, University of New South Wales, Australia
[3]Temasek Laboratories@NTU, Nanyang Technological University, Singapore
{dova0001, xiaoxiong, aseschng}@ntu.edu.sg, hli@i2r.a-star.edu.sg

ABSTRACT

This paper presents a novel method for acoustic modeling with limited training data. The idea is to leverage on a well-trained acoustic model of a source language. In this paper, a conventional HMM/GMM triphone acoustic model of the source language is used to derive likelihood scores for each feature vector of the target language. These scores are then mapped to triphones of the target language using neural networks. We conduct a case study where Malay is the source language while English (Aurora-4 task) is the target language. Experimental results on the Aurora-4 (clean test set) show that by using only 7, 16, and 55 minutes of English training data, we achieve 21.58%, 17.97%, and 12.93% word error rate, respectively. These results outperform the conventional HMM/GMM and hybrid systems significantly.

Index Terms— speech recognition, under-resourced language, cross-lingual LVCSR, context dependant, phone mapping.

1. INTRODUCTION

Automatic speech recognition (ASR) technology has made significant progress over the past decades. Unfortunately, speech researchers have focused only on a small number out of thousands of spoken languages in the world [1]. One of the limitations of current ASR systems is that they rely on a large amount of training data for acoustic modeling. Usually, to build a reasonable acoustic model for a large-vocabulary continuous speech recognition (LVCSR) system, tens to hundreds of hours of training data are required, which makes a full fledged acoustic modeling process impractical especially for under-resourced languages. This motivates us to investigate methods to automatically transfer well-trained acoustic models to under-resourced languages.

Several methods have been proposed for cross-lingual speech recognition [2–6]. Among these techniques, a group techniques called cross-lingual phone mapping is of our interest. In cross-lingual phone mapping, to recognize a target language (usually under-resourced), the target language speech data is first recognized into the phone sequences or phone posteriogram of a source language (using the well-trained source acoustic model), and then mapped to the phone sequences or posteriogram of the target language. 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 source phone sequence/posteriogram for speech recognition purpose, given that 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, due to the fact that it requires fewer data to to train a phone-to-phone mapping system than to train a feature-to-phone mapping system from scratch. In cross-lingual phone mapping, the source acoustic model acts as a feature extractor that generates high-level and meaningful features for the mapping. This then allows the use of a simple mapping trained with very little data to map the source phones to the target phones.

Several cross-lingual phone mapping methods have been studied in the past. In [4], Schultz and Waibel used a hard-decision phone mapping to build the seed model of the target language from a well-trained source language. Each target language phone is mapped to a fixed source language phone. In [5], 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 there is rarely to have a one-to-one mapping between phones of two languages, it is desirable to have “soft-mapping” of phones rather than “hard-mapping”. In [6], 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 [5]. The mapping of source phone posteriors to target phone posteriors is implemented by using a product-of-expert method which is realized by a 3-layer multilayer perceptron (MLP).

In this paper, we aim at building an LVCSR system for a language with very few training data by leveraging on existing well-trained acoustic models of another language. We use a similar phone mapping framework in [6] with two major modifications. First, to retain good resolution of the acoustic space, we map from source triphone states to target triphone states rather than map from source monophone states to target monophone states. We call this context-dependent cross-lingual phone mapping. Second, we use a conventional HMM/GMM (Hidden Markov Model / Gaussian Mixture Model) triphone model as the source acoustic model rather than using hybrid HMM/MLP (Hidden Markov Model / Multilayer Perceptron) source model. This makes our approach easier to be applied as HMM/GMM acoustic models are easier to build.

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

978-1-4673-2507-3/12/$31.00 © 2012 IEEE

2. CROSS-LINGUAL PHONE MAPPING

2.1. Prior Work

In cross-lingual phone mapping, the first step is to convert the speech data of the target language to either phone sequences [5] or phone posteriors [6] of the source language. Take monophone state posteriors [6] 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} state of the source language acoustic model. The source posterior vector is denoted as $\mathbf{u}_t = [p(s_1|\mathbf{o}_t), ..., p(s_{N_F}|\mathbf{o}_t)]^T$, where N_F is the number of states in the source language. The representation $\mathbf{u}_t$ is then mapped to, e.g. the phone state posteriors of the target language states:

$$p(q_j|\mathbf{o}_t) = f(\mathbf{u}_t), j = 1, ..., N_T \quad (1)$$

where $f(\cdot)$ can be any mapping function, q_j is the j^{th} state of the target language acoustic model, and N_T is the number of target phone states. In [6], the mapping function is a product of expert system and is realized by using a 3-layer MLP. 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 [6], 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 target and source languages. In this way, the resolution of the system is increased. The other improvement is to use conventional triphone HMM/GMM as the source acoustic model. We will explain these two improvements in the following section.

2.2. 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 higher 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 triphones is that, many acoustic modeling techniques, such as discriminative training and 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 those 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 frame $\mathbf{o}_t$ is encoded into a vector of log likelihoods as follows:

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

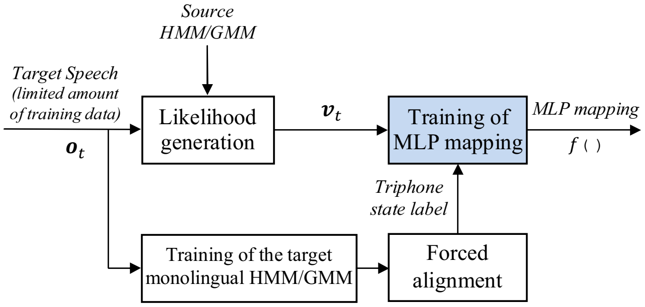

Fig. 1. *A diagram of the training process for cross-lingual phone-mapping.*

where N_F in this case is the number of tied-states in the source acoustic model. The s_i in (2) is the i^{th} tied-state in the source acoustic model. Similar to the monophone state mapping, in (3) the source triphone likelihood vector $\mathbf{v}_t$ is mapped to the target triphone tied-states:

$$p(q_j|\mathbf{o}_t) = f(\mathbf{v}_t), j = 1, ..., N_T \quad (3)$$

where N_T is the number of tied-states in the target language acoustic model. The mapping function is implemented by a 3-layer MLP [6]. It is also possible to convert the log likelihood vector $\mathbf{v}_t$ into tied-states posterior vector before applying phone mapping. However, our preliminary study found that the conversion to posteriors does not improve performance. Hence, we will use the log likelihood vector as the input of the phone mapping. This is different from [6] where posteriors generated from a hybrid phone recognizer were used.

The training of the cross-lingual phone mapping is illustrated in Fig. 1 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 Extract the triphone tied-states of the well-trained HMM/GMM source acoustic model. Evaluate the feature vectors $\mathbf{o}_t$ of the target language training data on the source tied-states to generate the log likelihood vector $\mathbf{v}_t$ as in (2).

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

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

Step 1 Generate the log likelihood 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 $p(q_j|\mathbf{o}_t)$.

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

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

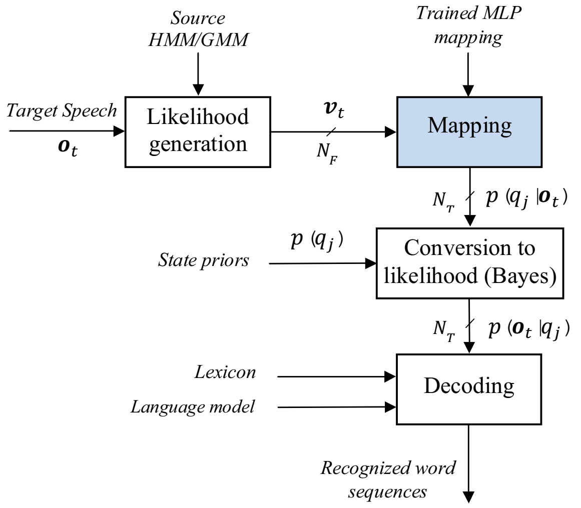

Fig. 2. *A diagram of the decoding process using cross-lingual phone mapping.*

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 as the target language. The English training data is randomly selected from the clean training of the Aurora-4 task [7]. Malay-to-English phone mapping functions are trained from 7, 16, or 55 minutes of English training data. 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: The Malay acoustic model uses the conventional HMM/GMM triphone structure and is trained from 100 hours of training data [8]. There are 1592 tied-states in the model and each state contains 32 Gaussian mixture components. We also trained a monophone acoustic model for comparison purpose which contains 102 states and 32 Gaussian mixture components per state.

Target language corpus: In this study, English is used as the target language, we use the large vocabulary Aurora-4 corpus [7]. The small clean test set consists of 166 sentences.

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 25ms and the frame shift is 10ms. To reduce recording mismatch between the source and target corpora, utterance-based mean and variance normalization (MVN) are applied to both training features of Malay and training and testing features of English. The English hybrid baseline HMM/MLP systems use feature vectors concatenated from 9 frames of MFCC features.

Language model and dictionary: The Wall Street Journal English bigram language model is used in word recognition experiments. The test set contains a vocabulary of 5k words.

MLP network training: To train the MLP neural networks (for both phone mapping and the hybrid baseline system) to generate the state posterior probabilities, the training set is separated into two parts randomly. The first part is used as the training data and contains around 90% of the training set. 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 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 state, the probability of jumping to the next state to 0.5. The probability of remaining in the state is hence also 0.5.

3.3. Baseline Monolingual Acoustic Models

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 dependant 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 mixtures) outperforms the best monophone model (8 mixtures), although the two acoustic models contains comparable number of Gaussian mixture components. The results show that triphone model is more robust than monophone model even with an extremely small amount of training data.

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

% Number of Gaussian mixture components	Monophone Model	Triphone Model
2	37.50	27.73
4	31.05	24.38
8	26.02	25.38
16	26.14	-

3.3.2. Monolingual hybrid HMM/MLP acoustic models

We have also conducted two English monolingual hybrid HMM/MLP models [9] using the same 16 minutes of training data to compare against the two models reported in Subsection 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. In addition, there is no required detailed assumption about input distribution. They have been applied successfully for phone recognition [10] and recently for word recognition [11].

In this experiment, MLP is used to predict the posterior probabilities of the monophone states and triphone tied-states respectively. 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 25.89% and 23.72%, respectively. These results are slightly better than the corresponding HMM/GMM models.

3.4. Cross-lingual 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. 2, 39-dimensional MFCC feature vectors, $\boldsymbol{o}_t$ are passed through the source HMM/GMM acoustic model to obtain N_F likelihood scores from N_F source state models. In this study, we examine two different source acoustic models:

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

These N_F likelihood 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. However, since the range of likelihood scores is very large, before the mapping step they are taken logarithm and normalized to zero mean and unit variance over the training set.

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 baseline monolingual acoustic models. The next two rows represent the results for proposed cross-lingual acoustic models. From the table, we have two major observations. First, by comparing the last two rows of the table, 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. Second, by comparing the second and third columns of the table, we observe that using target language triphone states as the label of the phone mapping consistently outperforms using target language monophone states. The best performance of cross-lingual phone mapping is WER=17.97% and is obtained by using triphone representation in both the source and target languages. This result is significantly better than all the baseline results. It is also 3.87% lower than the WER of 21.84%, obtained when monophone states are used in both the source and target languages.

Table 2. The WER (%) of the 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		
HMM/GMM	26.02	24.38
Hybrid HMM/MLP	25.89	23.72
Proposed cross-lingual acoustic model		
Source monophone($N_F = 102$)	21.84	19.59
Source triphone($N_F = 1592$)	20.59	17.97

3.5. Effect of Training Data Size

In this section, we will further examine the effect of amount of training data on the performance of the cross-lingual phone mapping acoustic model. The training data sizes are in our study, i.e. 7 min, 16 min, and 55 min of target training data.

In the previous two sections, we have shown that the context dependant triphone states are a better choice than monophone states for cross-lingual phone mapping. Therefore, in this section, the target language speech unit is always triphone while the source language unit could be either monophone or triphone. For each training data size, we follow the training steps in Section 2.2 to build the phone mapping. Note that we keep the number of triphone tied-states in the target language to be always 243 for fair comparison.

Table 3. The WER (%) of different acoustic models with different amount of target training data (the target acoustic model is triphone). Relative improvement of the proposed cross-lingual models over the best baseline model is indicated in (.) at the last row.

Method	Amount of training data		
	7 mins	16 mins	55 mins
Baseline monolingual acoustic model			
HMM/GMM	32.08	24.38	15.95
Hybrid HMM/MLP	30.90	23.72	15.45
Proposed cross-lingual acoustic model			
Source monophone	23.06	19.59	14.59
Source triphone	21.58 (30.2%)	17.97 (24.2%)	12.93 (16.3%)

Table 3 shows the performance of different acoustic models with three different amounts of training data. The first two rows are the two baseline monolingual models. We can see that the performance of the both two baseline models degrades quickly when less training data is used. The hybrid HMM/MLP model outperforms the conventional HMM/GMM slightly for all data sizes.

The last two rows of Table 3 show the performance of the cross-lingual acoustic models. The results show that using source triphone representation as the input of the mapping consistently outperforms using source monophone in all the training data sizes. It is also observed that the relative improvement of the cross-lingual phone mapping over the best baseline (i.e. hybrid HMM/MLP) increases as the training data decrease. The relative improvements are 30.2%, 24.2%, and 16.3% when the target training data is 7 min, 16 min, and 55 min, respectively. This shows that the proposed cross-lingual phone mapping is especially useful when only very small amount of target training data is available.

4. CONCLUSION

In this paper, we proposed a context-dependent cross-lingual phone mapping. This technique is used for fast training of acoustic model for under-resourced languages. There are two advantages in our method, i.e. the use of triphone states for improved acoustic resolution and the use of HMM/GMM source acoustic model which is easy to be obtained. Our experimental results on English verified the effectiveness of the proposed phone mapping technique for building LVCSR model.

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. If multiple source acoustic models are used to cover the target language acoustic space, will the performance be improved. We will try to answer these questions in the future.

5. REFERENCES

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

[2] L. F. Lamel and J. L. Gauvain, "Cross-Lingual Experiments with Phone Recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1993.

[3] S. Thomas, S. Ganapathy, and H. Hermansky, "Cross-lingual and Multi-stream Posterior Features for Low Resource LVCSR Systems," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2010, pp. 877–880.

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

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

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

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

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

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

[10] P. Matejka P. Schwarz and J. Cernocky, "Towards lower error rates in phoneme recognition," in *TSD, Brno, Czech Republic*, 2004.

[11] A. Abad, T. Pellegrini, I. Trancoso, and J. Neto, "Context dependent modelling approaches for hybrid speech recognizers," in *Conference of the International Speech Communication Association (INTERSPEECH)*, 2010, pp. 2950–2953.