AN EXEMPLAR-BASED APPROACH FOR CROSS-LINGUAL SPEECH RECOGNITION

Do Van Hai[1, 2, 3]

[1]*Faculty of Computer Science and Engineering, Water Resources University, Vietnam*

[2]*School of Computer Engineering, Nanyang Technological University, Singapore*

[3]*Temasek Laboratories@NTU, Nanyang Technological University, Singapore*

Email: haidv@wru.edu.vn

1. INTRODUCTION

Among thousands of spoken languages are used today, few ofthem are studied by the speech recognition community [3]. Oneof the major hurdles of automatic speech recognition (ASR) system deployment in new languages is that ASR system relies on a large amount of training data for acoustic modeling. Usually, to build a reasonableacoustic model for a largevocabulary continuous speech recognition (LVCSR) system, tens to hundreds of hours of trainingdata are required, which makes a full-fledged acoustic modelingprocess impractical especially for under-resourced languages.This motivates us to investigate methods to automatically transfer well-trained acoustic models to under-resourced languages.

Various methods have been proposed for cross-lingualspeech recognition such as universal phone set [4], tandemapproach [5], subspace GMMs (SGMMs) [6], KL-HMM[19], cross-lingual phone mapping [7] and our recentfinding, context-dependent phone mapping [8, 9, 14, 15]. Allthe techniques described above are referred to eager (offline)learning [17]. That means models are built offline before thetest data comes. In the case of limited training data, someclasses have only a few training examples, model parameters estimated for this class may be unreliable. Exemplar-based techniques such as k-nearest neighbors (k-NNs) which are calledlazy learning in the machine learning community build localmodels once a testing sample comes. These techniques do notsuffer from data sparsity issues since they require minimal orno training, and information is used from individual examplesin the training set when making assignments, thereby avoidingthe drawbacks of eager learning such as the GMM modeling [18]. Exemplar-based methods have been applied for monolingual ASR in several tasks. In [2], the authors indicated thatthe k-NN method is comparable and even better than the traditional HMM/GMM approach whenless than three hours oftraining data is used. The k-NNs method was also used in [16],a consistent improvement is achieved when applying smoothingacross frames.

In this paper, we apply the exemplar-based approach forcross-lingual speech recognition. Specifically, cross-lingualbottleneck feature is used as the input for the exemplar-basedmodel. Bottleneck features proposed by Grezl et.al [10] havebeen used widely in both monolingual and cross-lingual ASR[10, 14]. In cross-lingual ASR, a bottleneck neural networktrained on a language called the source language is used togenerate bottleneck feature for another language called the target language. In this study, we apply the exemplar-based approach at the state level that means the GMM part in the conventional HMM/GMM model is replaced by an exemplar-basedmodule to estimate the state likelihood scores. This approach isvery flexible since we can apply for existing HMM-based systems from small to large vocabulary. We can also utilize thetraditional Viterbi decoders in the testing phase. In addition,to achieve a further gain from the exemplar-based approach,we propose a technique called discriminative exemplar-basedmethod. The basic idea of this approach is to use a neural network to discriminatively calibrate the likelihood scores generated by the exemplar-based model.

2. EXEMPLAR-BASED APPROACH FOR CROSS-LINGUAL LVCSR

In this study, we apply the exemplar based approach at the HMM state level. It means that

when a testing sample o_t comes, likelihood probability $p(o_t|s_j)$ of HMM state s_j is estimated using the training exemplars.

In the other word, by using the exemplar-based approach we can remove the GMM part in the conventional HMM/GMM model. One difficulty to apply exemplar-based approaches such as k-nearest neighbors (k-NNs) at the state level is that these methods originally do not provide class probability $p(o_t|s_j)$ (soft decision). They are normally based on voting schemes (hard decision) which are not suitable to plug into the traditional HMM model at the state level. In this paper, we use the kernel densities approach [1] that allows for class probability estimation. In [2], this approach was applied successfully for speech recognition and achieved promising results.

Fig. 1 illustrates our cross-lingual LVCSR system usingthe exemplar-based approach. There are four steps to build thesystem.

1. Use the target language data x_t to build the conventional triphone HMM/GMM model and then apply forced alignment to generate the state frame label.
2. The bottleneck deep neural network (BN-DNN) trained by the source language training data is used to generate cross-lingual bottleneck feature o_t for the target language data.
3. Use the exemplar-based approach to estimate the likelihood probability $\hat{p}(o_t \mid s_i)$ for every target language test frame o_t.
4. Use the state likelihoods, together with the target language model and lexicon for Viterbi decoding.

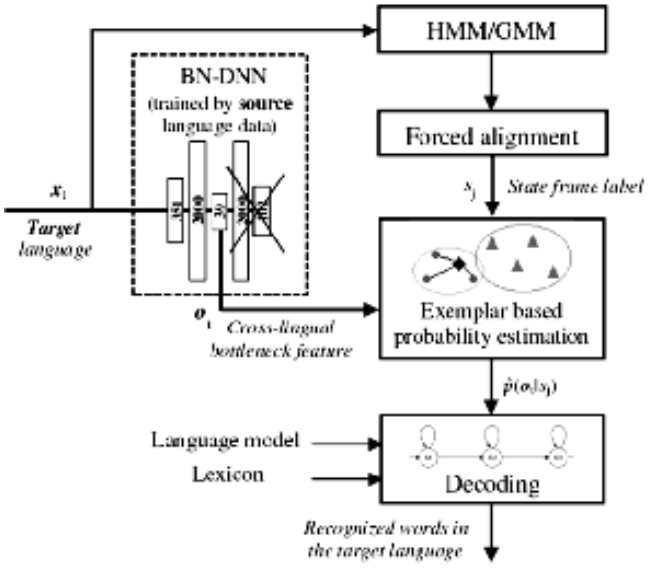

Figure 1: *Exemplar-based approach for cross-lingual ASR.*

3. DISCRIMINATIVE SCORE CALIBRATION

In the previous section, we presented a method using exemplar-based approach for cross-lingual LVCSR. The acoustic model in that method can be considered as a generative model since the state likelihood $p(o_t|s_j)$ is estimated for each state s_j independently. It is well known that using discriminative models e.g., multilayer perceptron (MLP), deep neural network (DNN) [13] or discriminative training criteria [20] can improve performance of speech recognition significantly. To achieve a further gain from the exemplar-based approach, we propose a technique called discriminative exemplar-based method. The basic idea of this approach is to use a neural network to discriminatively calibrate the likelihood scores generated by the exemplar-based model.

4. EXPERIMENTS

4.1. Experimental setup

Tasks and databases: To examine the effect of the proposed method, we use Malay - an Asian language as the source language and English (Aurora-4 task [11]) as the presumed under-resourced language. The Aurora-4 task has been chosen as the target under-resourced language as the effect of sufficient training data for it is well known, and we can hence clearly demonstrate the effect of reduced resources and our proposed work.

The source BN-DNN is trained with more than 100 hours of Malay read speech data [12]. While the target acoustic models are trained from limited amounts of English training data which is randomly selected from the clean training data of the Aurora-4 task. For testing, we used the small clean test set of Aurora-4, which consists of 166 sentences, or 20 minutes of speech.

4.2. Speech recognition with limited amounts of targettraining data

In this study, we investigate performance of monolingual and cross-lingual models when only limited amounts of target language training data are available. As indicated in our previous studies [14, 15, 30], using context-dependent triphone acoustic models in the target language brings a significant improvement over the monophone models even in the case of very limited training data. In this paper, all of the target acoustic models are context-dependent triphone with 243, 243, 501 and 1005 tiedstates for the cases of 7, 16, 55 and 220 minutes of training data, respectively.

The first row of Table 1 shows performance of the monolingual HMM/GMM model in word error rate (WER) with four different amounts of English training data. We also build an HMM/GMM acoustic model using all 15 hours of training data in Aurora 4 and get a WER of 7.9%. It can be seen that the performance of the HMM/GMM model drops significantly when less training data is available. We also build a monolingual exemplar-based system using MFCC feature. As shown in row 2 of Table 1, the exemplar-based method does not work well with the low-level MFCC feature where in all four cases, the exemplar-based model is outperformed by the HMM/GMM model.

Now, we examine cross-lingual models which use bottleneck feature generated by the Malay BN-DNN. The WERs of cross-lingual tandem and exemplar-based models are listed in row 3 and 4 of Table 1, respectively. We can see that using cross-lingual bottleneck feature achieves a significant improvement over MFCCs in both the two models except the case of the HMM/GMM model using 4 hours of training data. In this case, using the exemplar-based model can help to reduce WER significantly over the tandem approach. One reason can be the variability of cross-lingual bottleneck feature is smaller than the raw MFCC feature and hence a simple model like exemplar-based approach can benefit in case of limited training data.

Table 1: The WER (%) of the monolingual and cross-lingual models with 7, 16, 55 and 220 minutes of English training data.

#	Acoustic model	Training data (minutes)			
		7	16	55	220
Monolingual (MFCC feature)					
1	HMM/GMM	30.9	23.1	14.8	9.1
2	Exemplar-based	33.7	26.2	15.5	11.5
Cross-lingual (cross-lingual bottleneck feature)					
3	Tandem (HMM/GMM)	24.6	18.5	11.3	10.1
4	Exemplar-based	18.8	15.8	11.1	8.2
5	+ score calibration	**15.8**	**13.3**	**9.9**	**7.9**
6	Hybrid [14]	17.5	15.3	10.3	9.1

4.3. Discriminative score calibration

A 2-layer neural network is used to discriminatively calibrate the scores generated by the exemplar-based model. Experiment 5 in Table 1 indicates that by using the neural network as a discriminative score calibrator can improve performance of the exemplar-based model (row 4) significantly. Note that with 4 hours of training data, our cross-lingual system can produce a WER as low as the monolingual HMM/GMM system trained with whole 15 hours. These results are also 36% relatively better than the cross-lingual tandem model for the cases of 7 minutes of training data and this gain reduces to 22% for the case of 220 minutes of training data even both the two models use them same cross-lingual bottleneck feature.

The last row of Table 1 shows our previous result where cross-lingual bottleneck feature is used directly as the input of a 3-layer neural network. The WER obtained by that approach is consistently higher than WER given by the discriminative exemplar-based method presented in this paper. It proves that by using an exemplar-based module to estimate state probabilities before applying to a neural network is a better option than feeding the cross-lingual feature to the neural network directly even a more complicated neural network is used.

5. CONCLUSION

In this paper, we present a method using exemplar-based approach for cross-lingual LVCSR. Basically, an exemplar-based module is used to estimate the state likelihood scores. These scores are then calibrated in a discriminative manner using a 2-layer neural network. Experiments have been conducted to show that using our proposed framework can provide a significant improvement over the cross-lingual tandem approach as well as the cross-lingual hybrid model even they use the same input feature. In the future, different types of distances and different types of cross-lingual features are investigated.

REFERENCES

[1]. B.W. Silverman, "Density Estimation for Statistics and Data Analysis", Chapman and Hall, New York, 1986.

[2]. T. Deselaers, G. Heigold, and H. Ney, "Speech recognition with state-based nearest neighbour classifiers," in Proc. Interspeech, 2007, pp. 2093-2096.

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

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