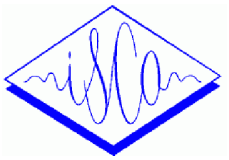

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Cross-Lingual and Ensemble MLPs Strategies for Low-Resource Speech Recognition

Yanmin Qian[1], *Jia Liu*[1]

[1]Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing, China

qianym07@mails.tsinghua.edu.cn, liuj@mail.tsinghua.edu.cn

10.21437/Interspeech.2012-11

Abstract

Recently there has been some interest in the question of how to build LVCSR systems for the low-resource languages. The scenario we focus on here is having only one hour of acoustic training data in the "target" language, but more plentiful data in other languages. This paper presents approaches using MLP based features: we construct a low-resource system with additional sources of information from the non-target languages to train the cross-lingual MLPs. A hierarchical architecture and multi-stream strategy are applied on the cross-lingual phone level, to improve the neural network more discriminatively. Additionally, an elaborate ensemble system with various acoustic feature streams and context expansion lengths is proposed. After system combination with these two strategies we get significant improvements of more than 8% absolute versus a conventional baseline in this low-resource scenario with only one hour of target training data.

Index Terms: low-resource language; cross-lingual posterior features; hierarchical architectures; ensemble system

1. Introduction

In recent years the performance of automatic speech recognition (ASR) systems has improved dramatically, but state-of-the-art systems require a large amount of language-specific transcribed speech data for training. However, demand exists for speech recognition systems in languages that have only limited available training data, and in these cases performance is still quite poor. Rapid development of ASR systems for resource-insufficient domains or languages is a research topic that has recently attracted interest [1][2].

Several strategies have been previously proposed. Developing a multilingual speech recognition system is a popular approach to deal with the low-resource problem [3][4]. In these systems a universal phone set is obtained based on the principle that the speech units with similar sounds across different languages are grouped together and represented by a single phonetic symbol. However the universal phone set is not as accurate as the language-specific one, and phone clustering induces more confusion among models, so the performance of these systems is not very promising.

Recently multilingual training with Subspace Gaussian Mixture Models (SGMM) [5][6] have also been presented for the low-resource application. The compact parameter structure and intrinsic ability to borrow out-of-domains or languages for model training makes SGMM suited for this scenario.

More recently various MLP based posterior features have been increasingly applied in this research area. An important point that impacts performance of these features is the amount of data used to train the neural network. For low-resource languages with only limited transcribed training data, the traditional training method is useless. A potential solution to this problem is to use transcribed data available from other languages to build models which can be shared with the low-resource language. Automatic speech attribute transcription (ASAT) based framework is one focused approach, whatever the task is phone level [7] or word level based transcription [8]. Articulatory Features (AF) [9] are more fundamental units shared across languages than phones, so that borrowing data from non-target languages becomes more reasonable. Another method mainly focuses on the NN parameters initialization in the training procedure, which can be obtained from another MLP training using the non-target languages, such as the work in [10]. The good initialization makes the target language MLP more robust. Besides other authors have previously looked at using MLPs in a cross-domain or cross-language setting [11][12].

In this paper we present a refined method based on the MLP architecture for this low-resource scenario. We construct a MLP system with additional sources of information from the non-target languages to train the cross-lingual MLPs, then a hierarchical architecture and multi-stream strategy are applied on the cross-lingual phone level to refine the NN outputs. Besides an elaborate ensemble system with various acoustic feature streams and context expansion lengths is developed. Finally we combine these two strategies MLPs to achieve the more improved performance.

The remainder of this paper is organized as follows: In Section 2, we describe the cross-lingual training strategy in a hierarchical architecture with the target and non-target language training data. Section 3 presents an elaborate ensemble based MLP framework to improve low resource MLPs only using the limited target languages data. Our experimental setup and experimental results are presented and compared in Section 4. Finally, we summarize and give conclusions in Section 5.

2. Cross-lingual hierarchical training strategy

Traditional MLP modeling with limited target data is usually useless. We investigate an alternate approach to build a new MLP structure using limited amounts of target training data with additional source of knowledge from cross-lingual training data. To enhance the features derived from the low-resource in-language training data, we first use data from the non-target languages to train cross-lingual MLPs and use the relative

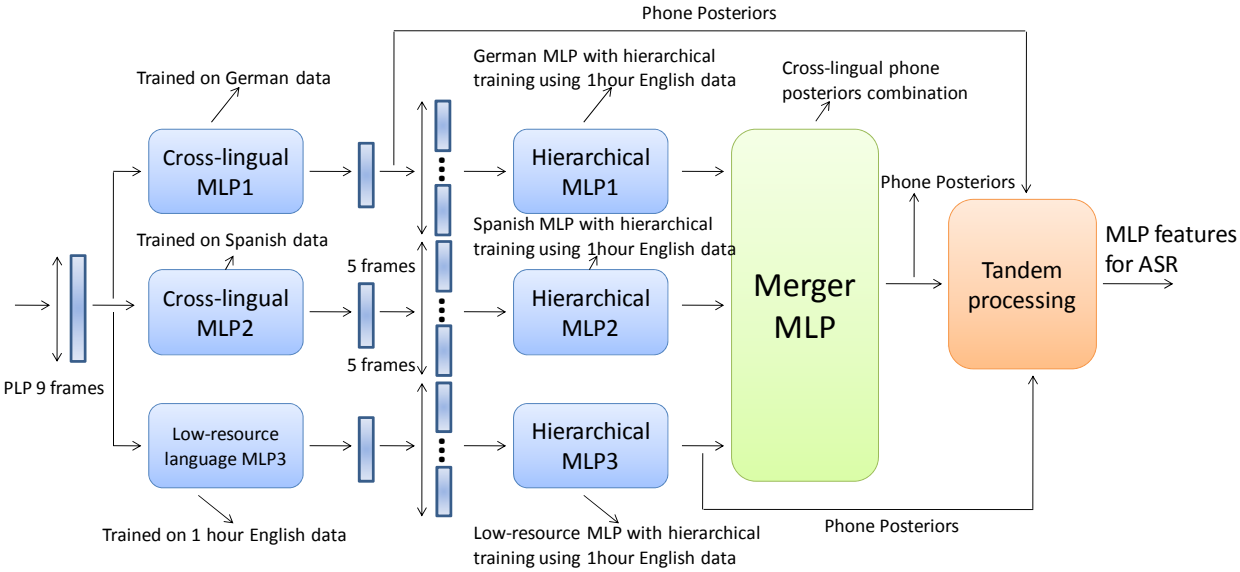

Figure 1: *Cross-lingual hierarchical training strategy for low resource speech recognition.*

language-dependent data or the target language data to finish the hierarchical training. A Multi-stream strategy is applied on the cross-lingual phone posteriors level to refine the final NN outputs.

A normal tandem process [13] is applied on the phone posteriors to generate the MLP feature. Logs of the phone posteriors are taken to approximately gaussianized values, and principal component analysis (PCA) is used to orthogonalize the features and retain the most important components which contribute most of the variance of the data.

2.1. Cross-lingual phone level hierarchical architecture training

To enhance the MLP features derived from the low-resource language data, we apply the paradigm illustrated in Fig. 1. We use PLP features with 9 frames of context to train the language-specific MLPs, including the non-target languages and target language, with their own phone sets (German, Spanish and English in our experiments). Even though the non-target languages have different phone sets from the target language, they share several common phonetic attributes of speech. The cross-lingual MLPs capture these attributes from each of the different posterior streams for that language.

To continue to refine the MLPs performance, we then apply the hierarchical training procedure [14] on the cross-lingual phone level. Another classifier is connected with the outputs of the previous language-specific MLPs, using a phone context of 110ms. When doing the hierarchical refinement on the cross-lingual MLPs, we can use the relative language-dependent data or only the target language data to train.

When utilizing only the target data to do the hierarchical architecture on the cross-lingual MLPs, we first need obtain the mapping relationship to determine that which phone in the non-target language set we can map each target language phone to. In our work, we used the State Time Alignment (STA) algorithm described in [15] to find the phone mapping between two languages. When the phone relationship is constructed, the hierarchical training is applied as normal. Experimentally, applying a hierarchical architecture with only the limited target language data is able to discriminate between phonetic classes of the low-resource language, better than that using the language-dependent data.

When these enhanced languages-specific phone classifiers are trained, we forward pass the limited target data through the cross-lingual MLPs to obtain different phone posteriors sets. These posteriors could be directly used in Tandem feature extraction as normal MLPs.

2.2. Multi-stream on the cross-lingual phone posteriors

To effectively explore the available knowledge from different languages, which may contribute for the low-resource target language, we combine the cross-lingual phone posteriors in our training architecture. As shown in Fig. 1, we pass all the language-specific NN outputs, including target and non-target languages, into another NN merger, and train these to classify the target language phones, which posteriors are used in the Tandem process.

3. Elaborate ensemble MLPs

An ensemble approach has been previously evaluated on some conversational speech recognition tasks [16]. Through data sampling, multiple training data sets are produced, and each sampled training data set is used to train one set of models. The ensemble set generates diverse models, which is particularly important and effective when data sparsity is an obstacle in low resource language development. In this paper we integrate a more elaborate ensemble framework into the MLPs training procedure, based our previous attempt [8]. This method is not a cross-lingual approach, as we only use the target-language training data.

We divide the target-language training data randomly into N equal-sized subsets (we used N=5 in our setup). We then train N different phone classifiers; each phone classifier is trained on N-1 out of the N data subsets. Diversity is obtained by using different acoustic representations as the inputs of different NNs,

such as PLP, MFCC and TRAP features. This helps the model sets extract distinct information from different speech acoustic features sufficiently, and make the ensemble models much more complementary and powerful. In addition we also use different context expansion lengths for each NN stream individually, such as acoustic features with 7/9/11 frames of context, to capture different speech context information efficiently. Using these elaborate ensemble models, we take the outputs of the N phone classifiers to train a merger MLP, which is trained on all the training data as a phone classifier. The outputs of this merger MLP are then processed with the typical Tandem fashion.

Fig. 2 illustrates the architecture of our elaborate ensemble approach. All the networks are a three layer networks with 1500 hidden nodes, and the outputs are the low-resource phone targets.

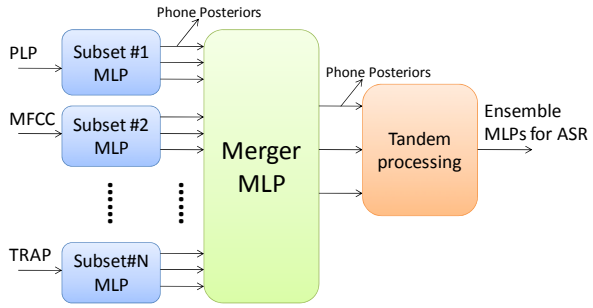

Figure 2: *Ensemble MLPs framework.*

4. Experiments and Results

4.1. Experimental data and Baseline system

Our experiments are on the Callhome English, German and Spanish databases. The conversational nature of speech in the Callhome database along with high out-of-vocabulary rates, use of foreign words and telephone channel distortions make the task of speech recognition on this database challenging.

The database contains 80 spontaneous telephone conversations in each of English, German and Spanish, with about 15 hours of speech per language to be used as training data. To imitate the low-resource application, we select English as the target language and use 1 hour of randomly chosen speech from the English corpus as the target-language training data. Besides this, we use the entire 15 hours of German and 16 hours of Spanish training data. The 20 conversations of the English evaluation set, roughly containing 1.8 hours of speech, form our test set.

To train the MLPs, we use a 42-phone set for English, 46 for German and 28 for Spanish. We use force-aligned phone labels for the 1 hour of English training data, 15 hours of German data and 16 hours of Spanish data. All NNs are three-layer built using the ICSI QuickNet neural network software package [17], with the classical back-propagation algorithm and cross entropy error criterion. The learning rate and stopping criterion are controlled by the frame-based classification error on the cross validation data. 5000 hidden nodes are used for all the MLP systems using cross-lingual data, while we use 1500 nodes for the other NN training those only use the limited target training data. The baseline tandem system, using PLP features with 9 frames of context as the MLP inputs and phone posteriors as the MLP outputs, are trained on the 1 hour of English data using 1500 nodes in the hidden layer of the MLP.

All the above mentioned Tandem features are reduced to 30 dimensions to train the subsequent single pass HTK based recognizer, with 550 tied states and 4 Gaussians per state. For comparison we also train the HMM-GMM system with the normal 39-dimensional PLP parameters, plus per-speaker mean and variance normalization, using only the 1 hour English data. We used the SRILM tools to build a trigram language model with a word-list of 62K words obtained by interpolating individual models trained from English Callhome corpus, the Switchboard corpus and the Gigaword corpus. We use HDecode as the recognizer, and score the results with the NIST scoring scripts.

Table 1 summarizes the PLP-HMM-GMM baseline and MLP-HMM-GMM baseline results for our experiments. It is clear that the ASR systems built with low resource perform poorly, and MLP based technique achieves better performance than traditional features. Our proposed approaches aim to improve the low resource system.

Table 1. *Performance comparison of different systems using only 1 hour English data.*

System description	WER
Conventional PLP-HMM-GMM	72.57%
Baseline Tandem feature derived from PLP feature with 9frame context	71.23%

4.2. Evaluation of the cross-lingual hierarchical MLPs and ensemble MLPs

The first two lines of Table 2 show the performance of basic cross-lingual MLP systems, and we can see that though the different phone set from the target language, the cross-lingual systems obtain comparable or even better performance to the normal MLP system.

We apply a hierarchical strategy on the phone posteriors level to elaborate the MLPs. When implementing on the cross-lingual MLPs, we can use the language-dependent data (as System 3 and 4 show) or only the target language data (as the System 5 and 6 show) to train the hierarchical NN. We can see that phone-level hierarchical architectures refine the MLPs further, and that training the hierarchical MLP only with the target data is much better than with the language-dependent data, which make the cross-lingual MLPs more discriminative on the low-resource language.

We implement the multi-stream architecture on the cross-lingual posterior level in order to get more improved performance, which utilizes the outputs of System 5, 6 and 7. System 8 shows that the multi-stream cross-lingual phone posteriors combination explores more useful knowledge from various languages, and achieve more significant improvement.

The last line of Table 2 presents the results of our ensemble MLP. This proposed framework uses different subsets of training data, and each subset uses distinct acoustic feature and context lengths. Our elaborate framework generates more diverse models, and produces more accurate posteriors than the previous attempt [8]. This ensemble approach gives about 6% absolute WER improvement compared with the baseline PLP-HMM-GMM system.

Table 2. *Performance comparison of systems using cross-lingual hierarchical MLPs and ensemble MLPs on the English test set.*

System description	WER
System 1: trained on German data	70.78%
System 2: trained on Spanish data	71.60%
System 3: trained on German data, hierarchical with German data	68.72%
System 4: trained on Spanish data, hierarchical with Spanish data	70.03%
System 5: trained on German data, hierarchical with English data	67.19%
System 6: trained on Spanish data, hierarchical with English data	68.56%
System 7: trained on English data, hierarchical with English data	69.88%
System 8: three languages' phone posteriors combination (combining system 5 + 6 + 7)	66.17%
System 9: elaborate ensemble MLPs	66.62%

4.3. Strategy for phone posteriors combination

Furthermore we combine the above mentioned two strategies MLPs at the posterior level in order to get more improved performance. Based on System 8 and 9 in Table 2, we concatenate the posterior streams and use another merger MLP to generate the final phone posteriors. This system comprises complementarity of different training criteria, including cross-lingual strategy and the ensemble strategy. This best system results in an 8% absolute WER improvement (relative gain 11.2%) compared with the baseline PLP system.

Table 3. *Best performance combining two strategies phone posteriors on the English test set.*

System description	WER
Conventional PLP-HMM-GMM	72.57%
Strategy for phone posteriors combination	64.45%

5. Conclusions

In this paper, we have presented some ideas and experimental results for using MLPs in the low-resource speech application where out-of-language training data may be available. We examined the use of cross-lingual data by training MLPs on the non-target languages. Hierarchical training only with the limited target data on the cross-lingual MLPs produces more discriminative phone posteriors, and a multi-stream architecture on the cross-lingual phone level incorporates knowledge from different languages, giving an absolute 6.5% improvement. A more elaborate ensemble MLPs with various acoustic feature streams and context lengths is implemented and achieves promising performance. Combining Tandem features from there two different MLP strategies provides a further improvement of 2% in this low-resource scenario. In the future we hope to combine these ideas with previously published model-level approaches such as SGMMs [5].

6. Acknowledgements

This work was supported by the National High Technology Research and Development Program of China (Project 2008AA040201), the Project 2009BAH41B01 supported by National Science and Technology Pillar Program of China, the Project 90920302 of NSFC (National Natural Science Foundation of China), and the Project 60931160443 of NSFC and RGC.

7. References

[1] P. Fung and T. Schultz, "Multilingual spoken language processing," *IEEE Signal Processing Magazine*, vol. 25, pp. 89–97, May. 2008.

[2] X. Cui, J. Xue, et al., "Acoustic Modeling with Bootstrap and Restructuring for Low-Resourced Languages," in Proc. Of INTERSPEECH, pp:2974-2977, 2010.

[3] B. D. Walker, B. C. Lackey, J. S. Muller, and P. J. Schone, "Language-Reconfigurable Universal Phone Recognition," in Proc. Of EUROSPEECH, 2003.

[4] H. Lin, L. Deng, et al., "A Study on Multilingual Acoustic Modeling for Large Vocabulary ASR," in Proc. Of ICASSP, pp:4333-4336, 2009.

[5] D. Povey, L. Burget, et al., "The Subspace Gaussian Mixture Model-A Structured Model for Speech Recognition," Computer Speech and Language, vol. 25, Issue 2, pp:404-439, 2011.

[6] Y. Qian, D. Povey, J. Liu, "State-Level Data Borrowing for Low-Resource Speech Recognition based on Subspace GMMs," in Proc. Of INTERSPECH, 2011.

[7] S. M. Siniscalchi, T. Svendsen, and C. H. Lee, "Toward A Detector-Based Universal Phone Recognizer," in Proc. Of ICASSP, pp:4261-4264, 2008.

[8] Y. Qian, J. Xu, D. Povey, J. Liu, "Strategies for Using MLP based Features with Limited Target-Language Training Data," in Proc. Of ASRU, 2011.

[9] S. Stuker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in Proc. Of ICASSP, pp: 144-147, 2003.

[10] S. Thomas, S. Ganapathy and H. Hermansky, "Cross-lingual and Multi-stream Posterior Features for Low Resource LVCSR Systems," in Porc. Of INTERSPEECH, pp: 877-880, 2010.

[11] Q. Zhu, B. Chen, N. Morgan and A. Stolcke, "On using MLP features in LVCSR," in Proc. Of INTERSPEECH, pp:921-924, 2004.

[12] A. Stolcke et.al., "Cross-domain and cross-language portability of acoustic feature estimated by multilayer perceptrons," in Proc. Of ICASSP, pp: 321-324, 2006.

[13] H. Hermansky, D.P.W. Ellis, and S. Sharma, "Tandem connectionist feature extraction for conventional HMM systems," in Proc. Of ICASSP, pp: 1635-1638, 2000.

[14] J. Pinto, M. Magimai.-Doss, and H. Bourlard, "MLP Based Hierarchical System for Task Adaptation in ASR," in Proc. Of ASRU, 2009.

[15] Yanmin Qian and Jia Liu, "Phone Modeling and Combining Discriminative Training for Mandarin-English Bilingual Speech Recognition", in Proc. Of ICASSP, pp:4918-4921, 2010.

[16] X. Chen and Y. Zhao, "Data sampling ensemble acoustic modeling," in Proc. Of ICASSP, pp: 3805-3808, 2009.

[17] ICSI QuickNet Software Package, http://www.icsi.Berkeley.deu/ speech/qn.htm.