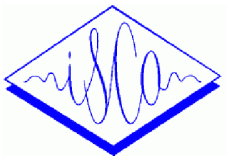

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Articulatory Feature based Multilingual MLPs 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

Abstract

Large vocabulary continuous speech recognition is particularly difficult for low-resource languages. In the scenario we focus on here is that there is a very limited amount of acoustic training data in the target language, but more plentiful data in other languages. In our approach, we investigate approaches based on Automatic Speech Attribute Transcription (ASAT) framework, and train universal classifiers using multi-languages to learn articulatory features. A hierarchical architecture is applied on both the articulatory feature and phone level, to make the neural network more discriminative. Finally we train the multilayer perceptrons using multi-streams from cross-languages and obtain MLPs for this low-resource application. In our experiments, we get significant improvements of about 12% relative versus a conventional baseline in this low-resource scenario.

Index Terms: low-resource language; multilayer perceptrons; articulatory features; hierarchical architectures

10.21437/Interspeech.2012-16

1. Introduction

State-of-the-art systems require a large amount of language-specific transcribed speech data for training. However, demand exists for speech recognition in languages that have only limited available training data, and in these cases sparse data may result in poor model estimation and performance. Rapid development of ASR systems for resource insufficient languages is a research topic that has recently attracted interest [1][2].

Several strategies have been previously proposed. Currently there are three key approaches to address low-resource applications: the universal phone based multilingual ASR method, the Subspace Gaussian Mixture model and the automatic speech attribute transcription (ASAT) strategy.

In the first method a universal phone set is obtained across different languages which can be used for the new language modeling [3][4]. 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. The Subspace Gaussian mixture model (SGMM) is a recently proposed acoustic model that is suited for low-resource applications [5]. It uses a more compact model parameter structure than a typical GMM based system, and has an intrinsic ability to borrow data out of domains or languages for model training. We have also done some work recently with SGMMs for the low-resource scenario, and the introduced data borrowing method [6] for borrowing closely non-target-language data to train acoustic states in the target language makes the model more refined.

Another way to deal with this scenario is automatic speech attribute transcription (ASAT) [7]. Articulatory Features (AF) are the more fundamental units shared across languages than the phones. This addresses the problem of low coverage of universal phone sets such as the IPA in limited data situations. The idea is to train classifiers to recognize articulatory features such as frication, voicing, nasality, etc.; even though a particular target-language phone may not have been seen in other languages, its attributes most likely will have been seen.

The approaches mentioned above are all model-level approaches. Our previous work [8], which using the ASAT framework, focuses on the feature level and obtained some promising results. In this paper we present a more elaborate multilingual MLP architecture based on our previous work [8] for this same scenario. We first construct a MLP system based on ASAT framework, and train all the articulatory features classifiers universally. A hierarchical architecture is applied on both the AF and phone level, to make the NN more discriminatively. A last merger NN is used to combine all the knowledge from the target and non-target languages. This refined AF based multilingual training strategy achieves even better performance than the multi-streams system combination system in our previous work [8].

Our experimental setup is similar to our previous work [8]: where we have limited amounts of training data in English, Spanish and German to imitate the low-resource situation. We will show significant improvements versus the traditional PLP-HMM-GMM method and the baseline MLP system.

The remainder of this paper is organized as follows: In Section 2, we describe the training strategy for the proposed AF based MLP architecture using multi-languages. A hierarchical architecture and cross-lingual posterior combination approaches are explained in detail. Our experimental setup and experimental results are presented and compared in Section 3. Finally, we summarize and give conclusions in Section 4.

2. Neural networks training strategy

2.1. Artifulatory feature based MLPs

Articulatory based MLPs were first developed in [9] and evaluated in an English monolingual system, demonstrating comparable improvements over traditional phone based MLPs. In our previous work [8], we developed an articulatory feature detector based MLPs framework. Fig. 1 reviews the outline of our AF based MLP system, consisting of two main blocks: (1) Articulatory feature classifiers, which consist of a bank of speech event detectors, and (2) a phone classifier, which take as input the outputs of the articulatory-feature detectors, and are trained to classify phones.

Our detectors are built similarly to those in [7], using 3 feed-forward NNs with 500 hidden nodes, and the merger NN is also feed-forward NN with one hidden layer and 1500 hidden nodes. The AF targets for the detectors training are obtained from deterministic phone-to-AF mapping of forced phonetic alignments from a baseline PLP-HMM-GMM system. The acoustic representations is STC feature (STCF) [10] for training, expect the baseline MLP system described later.

We applied tandem processing [11] on the phone merger outputs to generate the MLPs. For each frame, phone posteriors are taken the log function 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.

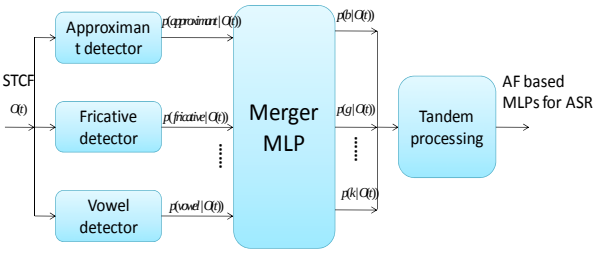

Figure 1: *Articulatory feature based MLPs framework.*

2.2. Multilingual training strategy

Previous work [12] has demonstrated that sounds described by the set of articulatory features share common acoustic properties across languages, e.g. among both the target language and the non-target languages. Articulatory features can be considered as more fundamental units than phonemes, since they are independent of the underlying language. All phones can be modeled by a set of articulatory attributes and most of the attributes, such as voicing, frication, can be identified in any particular language.

Using the multilingual advantage of articulatory features, we train the multilingual networks to obtain the universal AF detectors. We pool the target and non-target languages' data, and train the AF detectors on this data. As a result, these detectors are trained with much more data than is available in the target language, which contribute much more robust and discriminative articulatory feature classifiers.

2.3. Enhanced MLP training architecture

Besides utilizing multilingual superiority to obtain the universal AF classifiers, which can use significantly more out of languages data, in our new framework, we refined the training structure using several other techniques.

2.3.1. Articulatory feature level hierarchical architecture enhancement

The hierarchical Tandem system is first proposed in [13], applied on the phone MLP level. In our approach, we first use the hierarchical architecture on the universal articulatory feature level to enhance the AF classifiers. As the left part of Fig. 2 shows, a second MLP classifier is trained on the log posteriors features estimated by the first AF detectors with a temporal context of 200ms (this temporal context is empirically best in our experimental setup, and the frame AF accuracy achieve the best at an around 200ms context). The outputs of every second MLP are used as inputs to the language-specific merger NN in the same way as described above.

When enhancing the universal AF detectors with the hierarchical architecture, we fed the corresponding detectors outputs into merger NNs to train several language-specific phone classifiers, including the target and non-target languages, with their own phone sets (English, German and Spanish in our setup).

2.3.2. Cross-lingual phone level hierarchical architecture enhancement

To continue to improve the MLPs performance, we then apply the hierarchical training procedure on the phone level. As illustrated in the middle part of Fig. 2, another classifier is connected with the outputs of the language-specific phone merger, using a context of 110ms. When doing the hierarchical refinement on the non-target language MLPs, we can use the relative language-dependent data or only the target language data to train.

When we utilize only the target data to do the hierarchical architecture on the non-target 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 [14] to find the phone mapping between two languages. When the phone relationship is constructed, the hierarchical training is applied as normal.

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.

2.3.3. Cross-lingual phone posteriors combination

To sufficiently 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 strategy. As shown in the right part of Fig. 2, we pass all the language-specific merger outputs, including target and non-target languages, into another NN merger, and train this to classify the target language phones, which posteriors are used in the Tandem process.

3. Experiments and Results

3.1. 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

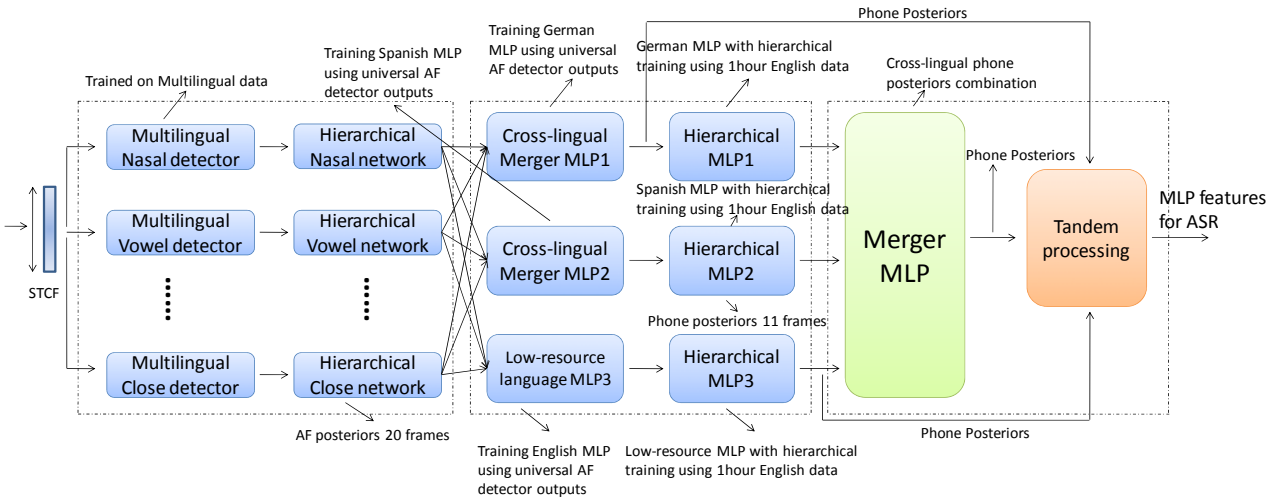

Figure 2: *Proposed Neural Network training strategy.*

evaluation set, 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, corresponding to 28 AFs for English, 29 for German and 27 for Spanish, which have 17 common AFs across all the three languages and total 10 common AFs existing between each two languages. 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, 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. 1500 hidden nodes are used for all the phone MLPs except the AF detectors which use 500 nodes configuration. 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 a 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 several baselines, include PLP-HMM-GMM baseline, MLP-HMM-GMM baseline and AF based monolingual MLP baseline. It is clear that the ASR systems built with low resource dada sets perform poorly, the MLP based technique achieves better performance than traditional features, and AF based MLPs perform similarly as the normal MLP system. Our proposed approaches aim to improve the system in such a low resource environment.

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%
Baseline Tandem feature derived from monolingual AF based Phone MLPs	71.87%

3.2. Evaluation of Our Proposed Multilingual MLP training architecture

We construct the AF detectors using different strategies, and evaluate the performance of AF classifiers. Fig. 3 shows several articulatory features' frame accuracy on the test set, including monolingual trained, multilingual trained, and multilingual plus hierarchical architecture AF detectors. Results indicate that multilingual AF classifiers utilize all the data across languages, using much more training data than the monolingual one, giving robust improved performance. The multilingual training mode with the hierarchical architecture enhances the AF detectors further, and produces more discriminative and better outputs.

We take the best universal AF detectors using a hierarchical strategy, as our system's final detectors, and the outputs of these classifiers are used in the following experiments.

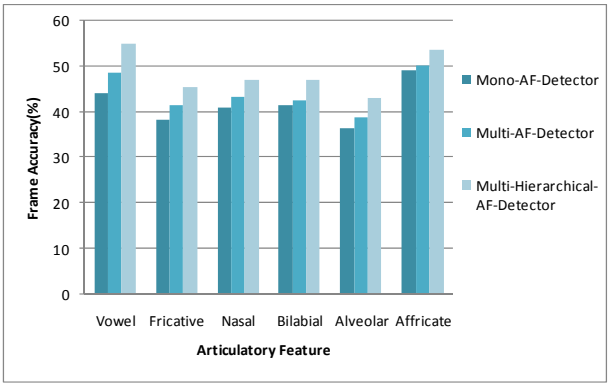

Figure 3: *Articulatory feature frame accuracy with different training strategies.*

Using the best AF classifiers' outputs described above, with the AF detectors trained in the multilingual training mode with a hierarchical architecture, we fed the corresponding detectors outputs to train language-specific phone classifiers, including

English, German and Spanish in our experiments. Table 2 shows the results of these three MLP systems, with forwarded limited English data. There are clear improvements over all monolingual systems shown in Table 1, and the cross-lingual MLP systems have comparable performance to the target phone MLP system.

Table 2. *Cross-lingual phone MLPs using multilingual trained articulatory feature frontend on English test set.*

System description	WER
Multilingual AF based English MLPs	68.37%
Multilingual AF based German MLPs	69.42%
Multilingual AF based Spanish MLPs	69.64%

We apply the hierarchical strategy on the phone level to elaborate the MLPs. When implementing on the cross-lingual MLPs, we could use the language-dependent data (as System 1 and 2 show) or only the target language data (as the System 3 and 4 show) to train the hierarchical NN. As Table 3 shows, we can see that phone-level hierarchical architectures refine the MLPs further, and training the hierarchical MLP only with the target data is much better than with the language-dependent data.

We combine the different streams at the cross-lingual phone posterior level to incorporate knowledge from different languages, including outputs of System 3, 4 and 5. The last line of Table 3 shows that the multi-stream cross-lingual phone posteriors combination adds useful information from various languages, and achieves significant improvement. This best system results in an 8.49% absolute WER improvement (relative gain 11.7%) compared with the baseline PLP system.

Table 3. *Performance comparison of different cross-lingual phone posterior systems on the English test set.*

System description	WER
System 1: Multilingual AF based German MLPs, + hierarchical with German data	68.65%
System 2: Multilingual AF based Spanish MLPs + hierarchical with Spanish data	69.05%
System 3: Multilingual AF based German MLPs + hierarchical with English data	67.36%
System 4: Multilingual AF based Spanish MLPs + hierarchical with English data	68.59%
System 5: Multilingual AF based English MLPs + hierarchical with English data	67.75%
System 6: Cross-lingual phone posteriors combination, using three languages' best hierarchical system	64.08%

4. Conclusions

In this paper, we introduce a new technique for training articulatory based multilingual MLPs for the low-resource scenario where out-of-language training data may be available. We investigate the AF based framework, and train the AF classifiers universally using multilingual data. A hierarchical architecture is first examined on the AF level to make the AF classifiers more discriminative. We also apply hierarchical training on the cross-lingual phone level, and doing that only with the limited target data on the cross-lingual MLPs produces better phone posteriors. In addition exploring common information from the articulatory feature level, we combine the available knowledge on the cross-lingual phone level from different languages to get a significant improvement of absolute 8.5% versus a conventional baseline in this low-resource scenario. In the future we hope to use several hundreds of hours of multilingual data and combine these ideas with previously published model-level approaches such as SGMMs [5] in this low-resource setting.

5. 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.

6. 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] O. Cetin et al., "An articulatory feature-based tandem approach and factored observation modeling," in Proc. Of ICASSP, pp: 645-648, 2007.

[10] P. Schwarz, P. Matejaka, and J. Cernocky, "Hierarchical structures of neural networks for phoneme recognition," in Proc. Of ICASSP, pp: 325-328, 2006.

[11] 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.

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

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

[14] 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.