Strategies for Using MLP based Features with Limited Target-Language Training Data

Yanmin Qian [#1], Ji Xu [#2], Daniel Povey [*3], Jia Liu [#4]

[#] *Tsinghua National Laboratory for Information Science and Technology*
Department of Electronic Engineering, Tsinghua University
Beijing, China
[1] qianym07@mails.tsinghua.edu.cn
[2] xuji010@gmail.com
[4] liuj@mail.tsinghua.edu.cn

[*] *Microsoft Research, Redmond, WA, USA*
[3] dpovey@microsoft.com

***Abstract*—Recently there has been some interest in the question of how to build LVCSR systems when there is only a limited amount of acoustic training data in the target language, but possibly more plentiful data in other languages. In this paper we investigate approaches using MLP based features. We experiment with two approaches: One is based on Automatic Speech Attribute Transcription (ASAT), in which we train classifiers to learn articulatory features. The other approach uses only the target-language data and relies on combination of multiple MLPs trained on different subsets. After system combination we get large improvements of more than 10% relative versus a conventional baseline. These feature-level approaches may also be combined with other, model-level methods for the multilingual or low-resource scenario.**

***Index Terms*—Low resource ASR, Multi-Layer Perceptrons, Articulatory features, Tandem features**

I. 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].

One approach [3][4] builds a multilingual system using a universal phone set, typically based on the International Phonetic Alphabet (IPA) or using data-driven approaches. Essentially this amounts to sharing phones across languages. The advantage of this system is that once one has a lexicon in the target language and knows the IPA symbols for the phones, it is possible to build a system with no target-language training data at all. Unfortunately, in practice there is never an exact correspondence between phones in different languages, so results with this method never approach the results from conventionally trained monolingual systems, and it is not clear that this method makes the best use of any data that is available in the target language.

The Subspace Gaussian mixture model (SGMM) is a recently proposed acoustic model that is especially suited for low-resource applications [5]. The majority of the trainable parameters of an SGMM are typically globally shared and not specific to any individual acoustic state; the only parameters specific to acoustic states are some relatively low-dimensional (e.g. 40-dimensional) vectors that represent fewer parameters than a typical GMM-based system. Therefore, when training SGMMs we can borrow other languages' data for model training without sharing the acoustic states, and obtain more robust estimates of the globally shared parameters [6]. We have recently worked with SGMMs for the low-resource scenario, and we introduced a method [7] for borrowing closely non-target-language data to train acoustic states in the target language. Our method worked with SGMMs but the idea is applicable to conventional GMM-based systems (the take-home message of that paper is that it is better to share context-dependent states than phones)

Another way to deal with this scenario is so-called automatic speech attribute transcription (ASAT) [8][9]. 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. Most of the research in this area has up till now been focused on phone-level rather than word-level transcription.

The approaches mentioned above are model-level approaches, in that they modify the acoustic model rather than the features. In this paper, we focus on the feature level. The ideas we present in this paper are not exclusive with other approaches for the low-resource scenario that we described above, and in future we hope to investigate combinations of methods. Specifically, we are looking at features based on Multi-Layer Perceptrons (MLPs). The general framework we are looking at is the Tandem framework [10] in which an MLP is trained to classify phones, and the features consist of log phone posteriors processed with PCA. Many authors concatenate these with the original PLP features, but we do not do this here. For good performance, MLPs have to be trained with a reasonably large amount of data, but this does

not necessarily have to be the same data used to train the acoustic model [11]. Other authors have previously looked at using MLPs in a cross-domain or cross-language setting [12][13].

In this paper we investigate improved MLP training methods for building ASR systems using extremely limited target-language training data (one hour). We investigated two quite different methods and also combined them for further improved results.

1) Multilingual Articulatory based MLP: Similar to the front end of the ASAT system [8], we train a number of MLPs to detect various binary-valued articulatory features. These classifier outputs can either be directly processed using PCA as Tandem features, or can be fed into another MLP trained to classify phones, and the output of this MLP used in Tandem processing; we experiment with both methods. Note that unlike [8], we are using the articulatory detectors in the feature extraction phase rather than as probabilities in the model.

2) Ensemble MLPs: The basic idea of the ensemble MLPs is to train separate phone-classifier MLPs on disjoint subsets of the target-language data and combine these with another MLP trained on all the data. Then we use the output of this MLP in Tandem feature extraction, which essentially does PCA on log phone posteriors.

3) Multi-Stream Combination: We explore combinations of the methods described above. Our favoured combination method is to train a phone classifier MLP on the outputs of MLPs in the stages described above, and use the outputs of this MLP as features in a Tandem approach.

Our experimental setup is similar to our previous work on SGMMs [7]: 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 our articulatory feature based MLPs and explain how to use non-target language data to obtain more robust MLPs. In Section 3, we propose an ensemble based MLP framework to improve low resource MLPs. Our experimental setup and experimental results are presented and compared in Section 4. Finally, we summarize and give conclusions in Section 5.

II. Multilingual Articulatory Based MLP

Articulatory based MLPs were first developed in [14] and evaluated in an English monolingual system, demonstrating comparable improvements over traditional phone based MLPs. In contrast, here we develop multilingual trained MLPs, both on the articulatory-feature and phone level, which generate more discriminative features.

A. Multilingual Articulatory Feature

In our work, we assume 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. Previous work [15] has demonstrated that articulatory features can be considered as more fundamental units than phonemes, since they are independent of the underlying language.

In this work, we have selected English as our target language, and Spanish and German as the non-target languages. Table I shows a part of the set of articulatory attributes used in our experiments, along with the attributes to phone mapping for these three languages. We consider all articulatory features as binary although they may take continuous values. Shown as Table I, we see that each phone possesses several articulatory features and every articulatory feature exists across a set of phones, shared commonly among different languages.

TABLE I

THE CORRESPONDING RELATION BETWEEN ARTICULATORY FEATURE AND PHONE WITHIN ENGLISH, SPANISH AND GERMAN

Articulatory Feature	**English**	**Spanish**	**German**
Alveolar	D C J s z n N T t S Z	l s n r R z	t d ts s z n l
Approximant	r y w l	B w D j G	l j
Fricative	f v R H h s z S Z	F s S x z	F v s z S Z x h r
Front	@ W E e I I X	i e	i I Y y e E W w @
Vowel	@ W a c Y E O e o I I X U u x A	a e i o u	a e i o u A @ E I O U W w Y y &
Nasal	n G m M N	p b m n N	G m n
Voiced	b D g v R J z Z n l r G m M N w	b B l m w d D J n y r R N g G z	NULL
Unrounded	@ W a c Y E e i I X A	NULL	I i e E @
……	……	……	……

B. Articulatory Feature based MLP System

Fig. 1 shows the outline of our multilingual articulatory feature based MLP system, consisting of two main blocks: (1) Articulatory MLPs, which consist of a bank of speech event detectors, (2) Phone MLPs, which take as input the outputs of the articulatory-feature detectors, and are trained to classify phones.

1) Articulatory MLPs

The goal of each detector is to analyse the speech signal and produce the posterior probability of some articulatory attribute. Our detectors are built similarly to those in [9], using 3 feed-forward NNs with 500 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.

Energy trajectories in mel-frequency bands, organized in split-temporal context as in [16], are used as parametric representations of the speech signal. In this work all MLPs used these STC features (STCF) for training, except the baseline MLP system described later.

We applied tandem processing [10] on the AF detector outputs to generate the Articulatory MLPs. For each frame, the posteriors from the AF detectors are joined together, taking the logarithm to approximately gaussianized the values, and principal component analysis (PCA) is used to orthogonalize the features and retain the most important components which account for the 95% of the total variance.

2) *Phone MLPs*

Besides using AF posteriors to obtain Articulatory MLPs, we trained a phone classifier (a "merger NN") on the outputs of the AF classifiers. The merger MLP is a feed-forward NN with one hidden layer and 1500 hidden nodes. Then the tandem process is applied: i.e. we do PCA on the log phone posteriors and these are the features we use.

3) *Multilingual MLPs Training*

When training multilingual MLPs, 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. The target-language data is then processed using these multilingual AF detectors, and the outputs are used as inputs to the merger NN, which is trained on the target-language data as a phone classifier. We experiment with two different feature extraction approaches: we either use the outputs of the original AF detectors directly in Tandem feature extraction, or we use the output of the merger NN in the same way. Fig. 1 illustrates both approaches.

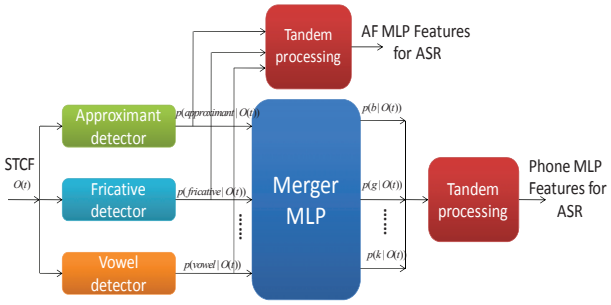

Fig. 1. Multilingual articulatory based MLP system

III. Ensemble MLPs

Our ensemble approach for MLP feature extraction is related to the approach described in [17]. This is not a multilingual 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 experiments). We then train N different phone classifiers; each phone classifier is trained on N-1 out of the N data subsets. We use 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 phone log-posteriors at the output of this MLP are then processed with PCA in the typical Tandem fashion.

Fig. 2 illustrates the architecture of our ensemble approach. All the networks are a three layer networks with 1500 hidden nodes, and the outputs are the phone targets. We also use the STC features [16] as the inputs to the first NNs.

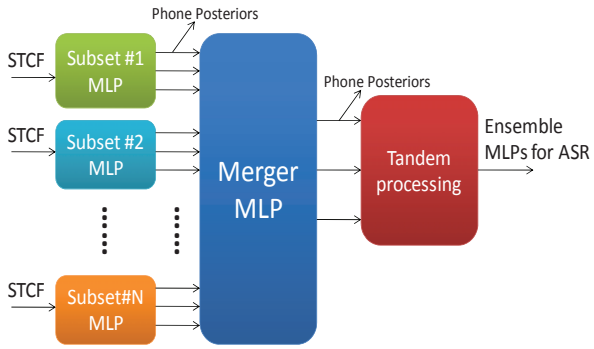

Fig. 2. Ensemble MLPs system

IV. Experiments And Results

A. *Experimental Data and Baseline System*

Our experiments are on the Callhome English, German and Spanish databases [18]. 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, corresponding to 28 AFs for English, 29 for German and 27 for Spanish, which have 17 common AFs across the three 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 [19], 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. 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 [20] 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 [21] and the Gigaword corpus [22]. We use HDecode as the recognizer, and score the results with the NIST scoring scripts.

The first two lines of Table II summarize 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 II

PERFORMANCE COMPARISON OF DIFFERENT SYSTEMS USING ONLY 1 HOUR OF TARGET LANGUAGE DATA

System description	WER
Conventional PLP-HMM-GMM	72.57%
Baseline Tandem feature derived from PLP feature with 9frame context	71.23%
Tandem feature derived from monolingual AF based Articulatory MLPs	72.32%
Tandem feature derived from monolingual AF based Phone MLPs	71.87%

B. *Evaluation of Multilingual Articulatory based MLPs and Ensemble MLPs*

We build multilingual articulatory based MLPs systems as described in Section II, and also develop the monolingual trained systems using only the 1 hour of target English data for comparison. The last two lines of Table II present the results of Articulatory MLPs and Phone MLPs utilizing the approach in Section II, but only using 1 hour of target training data. We can see that these two systems perform similarly to the normal MLP system, and the AF MLPs have comparable performance to the Phone MLPs.

Lines 1 and 2 of Table III show the experimental results of using multilingual articulatory based MLP, using both Articulatory MLPs and Phone MLPs. There are clear improvements over all monolingual systems shown in Table II, and the Phone MLP system is slightly better than the Articulatory MLP system.

The last line of Table III presents the results of adding the ensemble method to the MLP system. This proposed framework uses different subsets of training data, and generates diverse models, with a merger MLP combining the results to produce more accurate posteriors. This ensemble approach gives about 4% absolute WER improvement compared with the baseline PLP-HMM-GMM system.

TABLE III

PERFORMANCE COMPARISON OF DIFFERENT SYSTEMS USING TANDEM FEATURES ENHANCED WITH THE APPROACHES PROPOSED IN THIS PAPER

System description	WER
Tandem feature derived from multilingual AF based Articulatory MLPs	69.17%
Tandem feature derived from multilingual AF based Phone MLPs	68.37%
Tandem feature derived from Ensemble MLPs	68.58%

C. *System Combination*

The next step is to combine the different streams at the posterior level in order to get more improved performance. Based on the three systems shown in Table III, we concatenate the posterior streams and use another merger MLP to generate the final phone posteriors. After this, Tandem feature extraction and HMM model training are performed as before. This final system is constructed as shown in Fig. 3. This system not only comprises complementarity of different unit-based MLPs, AF vs. Phone, but also combines several different criteria training strategies, including the AF based strategy and the ensemble strategy.

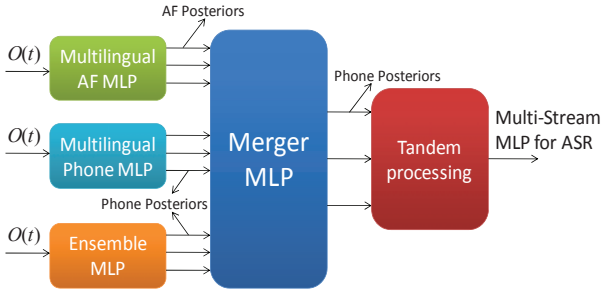

Fig. 3. Multi-Stream combination MLP system

Table IV shows that the multi-stream combination system results in a 7.85% absolute WER improvement (relative gain 10.8%) compared with the baseline PLP system.

TABLE IV

BEST PERFORMANCE USING MULTI-STREAM FEATURES COMBINATION

System description	WER
Conventional PLP-HMM-GMM	72.57%
Multi-Stream combination Tandem feature	64.72%

Fig. 4 shows a performance comparison of all the methods investigated in this paper. Compared to traditional systems, the proposed approaches show substantial improvements. The final system using multi-stream MLPs has the best overall

performance, with significant improvements over individual stream MLP systems. The improvements are not quite as large as our previously described improvements using the SGMM framework (we got down to about 60% error), but since the methods we describe here are all feature-level, we can in principle combine them with the model-level techniques we previously described.

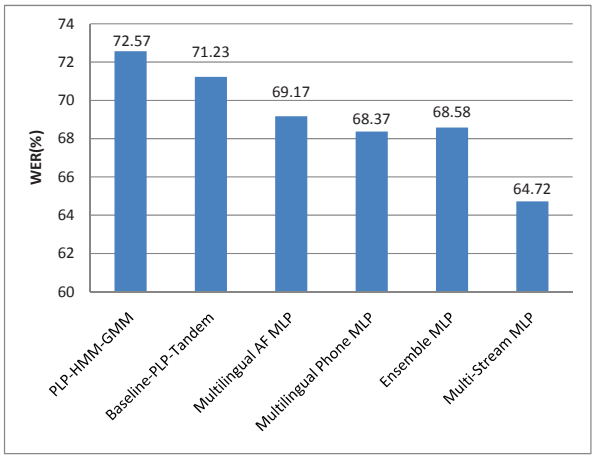

Fig. 4. The performance comparison of the methods investigated in this paper

V. Conclusion

In this paper, we presented some ideas and experimental results for using MLPs in the low-resource scenario where out-of-language training data may be available. We combined elements of the traditional Tandem feature-processing method, with an ensemble method for MLP training, and a multi-level MLP training method in which we train phone classifiers on the outputs of earlier classifiers. Overall, we got about 10% relative improvement over a conventional PLP-HMM-GMM system. In the future we hope to combine these ideas with previously published model-level approaches such as SGMMs [6], and investigate other MLP based approaches similar to [23].

Acknowledgment

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.

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] L. Burget, P. Schwartz, et al., "Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models," in Proc. Of ICASSP, pp:4334-4337, 2010.
[7] Y. Qian, D. Povey, J. Liu, “State-Level Data Borrowing for Low-Resource Speech Recognition based on Subspace GMMs,” in Proc. Of INTERSPECH, 2011.
[8] S. M. Siniscalchi, T. Svendsen, and C. H. Lee, "Toward bottom-up continuous phone recognition," in Proc. Of ASRU, 2007.
[9] S. M. Siniscalchi, T. Svendsen, and C. H. Lee, "Toward A Detector-Based Universal Phone Recognizer," in Proc. Of ICASSP, pp:4261-4264, 2008.
[10] 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.
[11] S. Sivadas and H. Hermansky, “On use of task independent training data in tandem feature extraction,” in Proc. Of ICASSP, pp:541-544, 2004.
[12] Q. Zhu, B. Chen, N. Morgan and A. Stolcke, “On using MLP features in LVCSR,” in Proc. Of INTERSPEECH, pp:921-924, 2004.
[13] 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.
[14] O. Cetin et al., “An articulatory feature-based tandem approach and factored observation modeling,” in Proc. Of ICASSP, pp: 645-648, 2007
[15] S. Stuker, T. Schultz, F. Metze, and A. Waibel, “Multilingual articulatory features,” in Proc. Of ICASSP, pp: 144-147, 2003.
[16] P. Schwarz, P. Matejaka, and J. Cernocky, “Hierarchical structures of neural networks for phoneme recognition,” in Proc. Of ICASSP, pp: 325-328, 2006.
[17] X. Chen and Y. Zhao, “Data sampling ensemble acoustic modeling,” in Proc. Of ICASSP, pp: 3805-3808, 2009.
[18] A. Canavan, D. Graff, and G. Zipperlen, “CALLHOME American English/German/Spanish Speech,” Linguistic Data Consortium, 1997.
[19] ICSI QuickNet Software Package, http://www.icsi.Berkeley.deu/speech/qn.htm.
[20] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," in Proc. Of ICSLP,pp:901-904, 2002.
[21] J.J. Godfrey el at., “Switchboard: Telephone speech corpus for research and development,” in Proc. Of ICASSP, 1992.
[22] D. Graff, “English Gigaword,” Linguistic Data Consortium, 2003.
[23] S. Thomas, S. Ganapathy and H. Hermansky, “Cross-lingual and Multilingual-stream Posterior Features for Low Resource LVCSR Systems,” in Porc. Of INTERSPEECH, pp: 877-880, 2010.