MULTILINGUAL MLP FEATURES FOR LOW-RESOURCE LVCSR SYSTEMS

Samuel Thomas, Sriram Ganapathy and Hynek Hermansky

Center for Language and Speech Processing,
Department of Electrical and Computer Engineering,
The Johns Hopkins University, USA.
{samuel,ganapathy,hynek}@jhu.edu

ABSTRACT

We introduce a new approach to training multilayer perceptrons (MLPs) for large vocabulary continuous speech recognition (LVCSR) in new languages which have only few hours of annotated in-domain training data (for example, 1 hour of data). In our approach, large amounts of annotated out-of-domain data from multiple languages are used to train multilingual MLP systems without dealing with the different phoneme sets for these languages. Features extracted from these MLP systems are used to train LVCSR systems in the low-resource language similar to the Tandem approach. In our experiments, the proposed features provide a relative improvement of about 30% in an low-resource LVCSR setting with only one hour of training data.

Index Terms— Multilingual training, multilayer perceptrons, MLP features for low-resource LVCSR.

1. INTRODUCTION

MLP based posterior features are increasingly being used to improve the performance of LVCSR systems [1, 2]. An important factor that impacts performance of these features is the amount of data used to train the MLP systems. For new languages with only few hours of transcribed data, the performance of these data driven features is low. 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. However training such systems requires all the multilingual data to be transcribed using a common phoneset across the different languages. This common phoneset can be derived either in a data driven fashion or using phonetic sets such as the International Phonetic Alphabet (IPA) [3]. More recently cross-lingual training with Subspace Gaussian Mixture Models (SGMM) [4] have also been proposed for this task.

The research presented in this paper was partially funded by IARPA BEST program under contract Z857701 and DARPA RATS program under D10PC20015. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the IARPA or DARPA.

In our previous work [5], we explored a data driven approach for finding a common phoneset across different languages. Using this approach we adapt a multilingual MLP trained on 30 hours of Spanish and German using one hour of English (considered as the low-resource language). Tandem features [1] derived from such a system were then used for an LVCSR task using one hour of English. In this paper we propose a different MLP architecture and training method for the same task. The primary advantage of this new architecture is that it does not require the multilingual data to be mapped using a common phoneset across various languages.

In the proposed architecture, we train a 4 layer multilayer perceptron. The MLP has a linear input layer with a size corresponding to the dimension of the input feature vector, followed by two non-linear layers and a final linear layer with a size corresponding to the phoneset of the language the MLP is being trained. While training on multiple languages with different phonesets, the first 3 layers are shared. The last layer that is specific to the phoneme set of each language is then modified. Modifying only this layer allows us to train across different languages.

Section 2 describes the training procedure for the proposed MLP architecture using multiple languages. Section 3 talks about how we derive features from these multilingual MLPs. We investigate the usefulness of this approach in Section 4 with experiments on English, Spanish and German CTS data. The paper concludes with a discussion in Section 5.

2. TRAINING THE NETWORKS

In this section we describe the training approach for the proposed MLP system on two languages - P and Q. P is the out-of-domain language with larger amounts of training data compared to the low-resource in-domain language Q. Both languages have different phoneme sets of size p and q. The network is trained using an acoustic representation with dimension d in the following steps -

A. Train the MLP on language P - We start by training a 4 layer MLP of size d,$h1$,$h2$,p on the high resource language

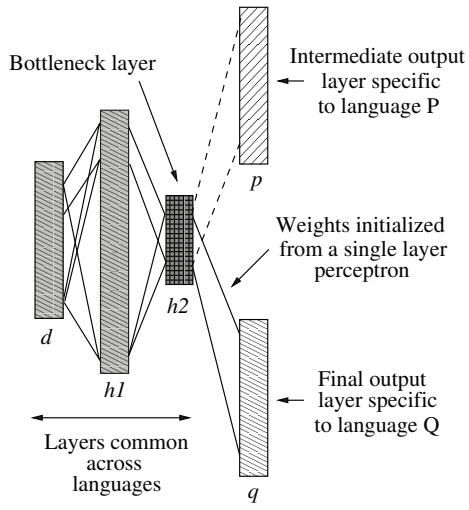

Fig. 1. Block schematic of the proposed MLP system.

with randomly initialized weights. While the input and output nodes are linear, the hidden nodes are non-linear. Similar to bottleneck MLPs [6] or HATS [7], while the dimension of $h1$ is high, $h2$ is low dimensional and is known as the 'bottleneck' layer. We are motivated to introduce the bottleneck layer to allow the network to learn a common low dimensional representation among the languages.

B. Initialize the network to train on language Q - To continue training on the low-resource language which has a different phoneme set size, we create a new 4 layer MLP of size d,$h1$,$h2$,q. The first 3 layer weights of this new network are initialized using weights from the MLP trained on the high resource language. Instead of using random weights between the last two layers, we initialize these weights from a separately trained single layer perceptron.

To train the single layer perceptron, non-linear representations of the low-resource training data are derived by forward passing the data through the first 3 layers of the MLP. The data is then used to train a single layer network of size $h2$,q.

C. Train the MLP on language Q - Once the 4 layer MLP of size d,$h1$,$h2$,q has been initialized, we re-train the MLP on the low-resource language. By sharing weights across languages the MLP is now able to train better on limited amounts of in-domain data. Figure 1 is a schematic of the proposed MLP system.

3. FEATURE EXTRACTION

The proposed 4 layer MLP are trained to estimate phoneme posterior probabilities using the standard back propagation algorithm with cross entropy error criteria. We derive two kinds of features for LVCSR task from these networks -

A. Tandem features - These features are derived from the posteriors estimated by the MLP at the fourth layer. When networks are trained on multiple feature representations, better posterior estimates can be derived by combining the outputs from different system using posterior probability combination rules. Phoneme posteriors are then converted to features by gaussianizing the posteriors using the log function and decorrelating them by using the Karhunen-Loeve transform (KLT) [1]. A dimensionality reduction is also performed by retaining only the feature components which contribute most to the variance of the data.

B. Bottleneck features - Unlike Tandem features, bottleneck features are derived as linear outputs of the neurons from the bottleneck layer [6]. These outputs are used directly as features for LVCSR features without applying any transforms. When bottleneck features are derived from multiple feature representations, these features are appended together and a dimensionality reduction is performed using KLT to retain only relevant components.

Both of these MLP features are derived using two acoustic feature representations - short-term spectral PLP [8] features and long-term modulation features using frequency domain linear prediction (FDLP-M) [9].

4. EXPERIMENTS AND RESULTS

We use the English, German and Spanish parts of the Callhome corpora collected by LDC for our experiments [10, 11, 12]. The conversational nature of speech 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 English database consists of 120 spontaneous telephone conversations between native English speakers. 80 conversations corresponding to about 15 hours of speech, form the complete training data [10]. We use 1 hour of randomly chosen speech covering all the speakers from the complete train set for our experiments as an example of data from a low-resource language. The English MLPs and subsequent HMM-GMM systems use this one hour of data. Two sets of 20 conversations, roughly containing 1.8 hours of speech each, form the test and development sets. Similar to the English database, the German and Spanish databases consist of 100 and 120 spontaneous telephone conversation respectively between native speakers. 15 hours of German and 16 hours of Spanish are used as examples of out-of-domain high resource languages for training the MLPs. Each of these languages use different phoneme sets - 47 phonemes for English, 46 for German and 28 for Spanish.

We train a single pass HTK based recognizer with 600 tied states and 4 mixtures per state on the 1 hour of data. We use fewer states and mixtures per state since the amount of training data is low. The recognizer uses a 62K trigram

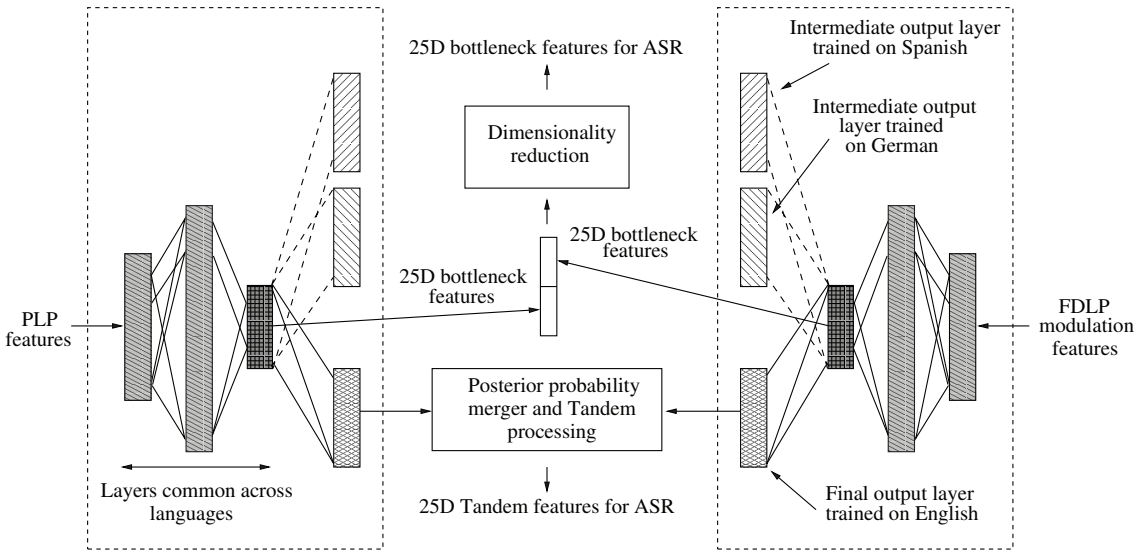

Fig. 2. Tandem and bottleneck features for low-resource LVCSR systems. We use 2 acoustic feature representations along with 2 languages - Spanish and German to train a multilingual system for 1 hour of English.

language model with an OOV rate of 0.4%, built using the SRILM tools. The language model is interpolated from individual models created using the English Callhome corpus, the Switchboard corpus [13], the Gigaword corpus [14] and some web data. The web data is obtained by crawling the web for sentences containing high frequency bigrams and trigrams occurring in the training text of the Callhome corpus. The 90K PRONLEX dictionary with 47 phones is used as the pronunciation dictionary for the system. The test data is decoded using the HTK decoder - HDecode, and scored with the NIST scoring scripts.

4.1. Training with 2 languages

In our first set of experiments we train a 4 layer MLP system on two languages - Spanish and English as outlined in Sec. 2. We start by training two separate networks on the out-of-domain language using 16 hours of Spanish. Both these systems have a first hidden layer of 1000 nodes, a bottleneck layer of 25 nodes and a final output layer of 28 nodes corresponding to the size of the Spanish phoneme set. 39 dimensional PLP features (13 cepstral + Δ + $\Delta\Delta$ features) are used along with a context of 9 frames to train the first network with architecture - 351 x 1000 x 25 x 28. A second system is trained on 476 dimensional modulation features derived using FDLP [9]. These features correspond to 28 static and dynamic modulation frequency components extracted from 17 bark spaced bands [9]. This system has an architecture of 476 x 1000 x 25 x 28. Both the systems are trained using the standard back propagation algorithm with cross entropy error criteria. The learning rate and stopping criterion are controlled by the error in the frame-based phoneme classification on the cross validation data.

After the out-of-domain networks have been trained, the in-domain networks to be trained on 1 hour of English are initialized in two stages as discussed in Sec. 2. In the first stage, all weights except the weights between the bottleneck layer and the output layer are initialized directly from the Spanish network. The second set of weights are initialized from a single layer network trained on non-linear representations of the 1 hour of English data derived by forward passing the English data through the Spanish network till the bottleneck layer. This network has an architecture of 25 x 47 corresponding to the dimensionality of the non-linear representations from the bottleneck layer of the Spanish network and the size of the English phoneme set. These networks are trained on both PLP and FDLPM features.

Once the networks has been initialized, PLP and FDLPM features derived from 1 hour of English are used to train the new in-domain low-resource networks. The networks trained on PLP and FDLPM features now have an architecture of 351 x 1000 x 25 x 47 and 476 x 1000 x 25 x 47 respectively. 47 dimensional phoneme posteriors from both the networks are combined using the Dempster Shafer (DS) theory of evidence [15] before deriving the 25 dimensional Tandem set (Section 3A). The 2 sets of 25 dimensional bottleneck features from each of the networks are appended together before applying a dimensionality reduction to form a final 25 dimensional bottleneck feature vector (Section 3B). Both the Tandem and bottleneck features are used to train the subsequent low-resource HMM-GMM system on 1 hour of training data.

Table 1 shows the results of using the proposed MLP based features. We train the 1 hour HMM-GMM system on 39 dimensional PLP features (13 cepstral + Δ + $\Delta\Delta$ features) as our baseline system.

Table 1. Word Recognition Accuracies (%) using two languages - Spanish and English

Baseline PLP features	28.8
Tandem features	34.9
Bottleneck features	35.4

4.2. Training with 3 languages

We extend our training on 2 languages to train a multilingual MLP system on 3 languages - Spanish, German and English. The training procedure starts as outlined earlier with 15 hours of Spanish. The networks are then initialized to train with the German data in two stages - with weights from the Spanish system till the bottleneck layer and with weights from single layer network trained to the German data. After the net has been trained on the German data, we do a re-training using the 1 hour of English data. Figure 2 is a schematic of the training and feature extraction procedure. Table 2 shows the results of using the proposed MLP based features.

Table 2. Word Recognition Accuracies (%) using three languages - Spanish, German and English

Tandem features	35.8
Bottleneck features	37.2

The above results show the advantage of the proposed approach to training MLPs on multilingual data. Unlike in earlier approaches we are able to train on multiple languages without using a common phoneset among the languages. On a low-resource task, features extracted from these multilingual MLP give up to 30% relative improvement over conventional features. While other techniques, for example the SGMM approach [4] improve acoustic models, the proposed approach focuses on improving feature representations for low-resource applications.

5. CONCLUSIONS

In this paper we introduce a new technique for training multilingual MLPs. We propose the use of an language dependent layer to conventional three layer MLPs which are used to derive phoneme posteriors. This approach allows for sharing resources across languages without needing to construct common phoneme sets. Future work will include more studies on using several hundreds of hours of multilingual data in a low-resource setting.

6. ACKNOWLEDGMENTS

Authors would like to thank Brian Kingsbury and Karen Livescu for the helpful discussions.

7. REFERENCES

[1] H. Hermansky, D.P.W. Ellis and S. Sharma, "Tandem Connectionist Feature Extraction for Conventional HMM Systems", Proc. of IEEE ICASSP, 2000.

[2] J. Park, F. Diehl, M.J.F. Gales, M. Tomalin and P.C. Woodland, "Training and Adapting MLP features for Arabic Speech Recognition", Proc. of IEEE ICASSP, 2009.

[3] H. Lin, L. Deng, D. Yu, Y. Gong, A. Acero, and C. Lee, "A study on Multilingual Acoustic Modeling for Large Vocabulary ASR", Proc. of IEEE ICASSP, 2009.

[4] L. Burget et. al., "Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models", Proc. of IEEE ICASSP, 2010.

[5] S. Thomas, S. Ganapathy and H. Hermansky, "Cross-lingual and Multi-stream Posterior Features for Low-resource LVCSR Systems", Proc. of ISCA Interspeech, 2010.

[6] F. Grezl, M. Karafiat, S. Kontar, and J. Cernocky, "Probabilistic and Bottle-neck Features for LVCSR of Meetings", in Proc. of IEEE ICASSP, 2007.

[7] B. Y. Chen, Q. Zhu, and N. Morgan, "Learning long-term temporal features in LVCSR using neural networks", Proc. of ICSLP, 2004.

[8] H. Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech", J. Acoust. Soc. Am., 1990.

[9] S. Ganapathy, S. Thomas and H. Hermansky, "Modulation Frequency Features For Phoneme Recognition In Noisy Speech", J. Acoust. Soc. Am. - Express Letters, 2008.

[10] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME American English Speech", Linguistic Data Consortium, 1997.

[11] A. Canavan, D. Graff, and G. Zipperlen, "CALLHOME German Speech", Linguistic Data Consortium, 1997.

[12] A. Canavan and G. Zipperlen, "CALLHOME Spanish Speech", Linguistic Data Consortium, 1997.

[13] J.J. Godfrey el. al., "Switchboard: Telephone speech corpus for research and development", in Proc. of IEEE ICASSP, 1992.

[14] D. Graff. "English Gigaword", Linguistic Data Consortium, 2003.

[15] F. Valente and H. Hermansky, "Combination of Acoustic Classifiers based on Dempster-Shafer Theory of Evidence", in Proc. of IEEE ICASSP, 2007.