Cross-lingual and Multi-stream Posterior Features for Low Resource LVCSR Systems

Samuel Thomas[1], *Sriram Ganapathy*[1,2], *Hynek Hermansky*[1,2]

[1]Department of Electrical and Computer Engineering,
[2]Human Language Technology Center of Excellence,
Johns Hopkins University, Baltimore, USA.
{samuel,ganapathy,hynek}@jhu.edu

Abstract

We investigate approaches for large vocabulary continuous speech recognition (LVCSR) system for new languages or new domains using limited amounts of transcribed training data. In these low resource conditions, the performance of conventional LVCSR systems degrade significantly. We propose to train low resource LVCSR system with additional sources of information like annotated data from other languages (German and Spanish) and various acoustic feature streams (short-term and modulation features). We train multilayer perceptrons (MLPs) on these sources of information and use Tandem features derived from the MLPs for low resource LVCSR. In our experiments, the proposed system trained using only one hour of English conversational telephone speech (CTS) provides a relative improvement of 11% over the baseline system.

Index Terms: Cross-lingual posterior features, Multi-stream features, Low resource ASR, Tandem features

1. Introduction

Conventional feature extraction techniques like Mel Frequency Cepstral Coefficients (MFCC) [1] and Perceptual Linear Prediction (PLP) [2] are motivated by properties of human auditory. In an alternate approach, these acoustic features can be used to train MLPs to estimate phoneme posteriors. Tandem features [3] are derived from these posteriors after decorrelation and dimensionality reduction. By training MLPs on sufficient amounts of data, these nets are able to classify sound classes, while discarding irrelevant details. With multiple frames of acoustic features being used to train the MLP, Tandem features also model information in larger temporal context [3]. Additionally, since multiple streams of acoustic features can be combined at the posterior level, more accurate and robust estimates of posteriors can be transformed into features for ASR [5].

In LVCSR systems, an important factor that impacts performance, is the amount of available transcribed training data. When LVCSR systems are built for new languages or domains with only few hours of transcribed data, the performance is rather low. To improve performances, unlabeled data from this new language or domain has been used to increase the size of the training set [4]. This is done by first recognizing the unlabeled data and incrementally adding reliable portions to the original training set. For these techniques to be effective, a low error rate recognizer is required to annotate the unlabeled data. However in several scenarios like ASR systems for new languages, low error rate recognizers are impractical using limited amounts of training data. Hence, additional improvements are not easily achieved.

In [6], a task independent approach has been used to first train MLPs with large amounts of data. Features derived from these nets are then shown to reduce the requirement of task specific data to train subsequent HMM stages. More recently this approach has been shown useful also in cross-domain and cross-lingual LVCSR tasks [7, 8]. In [8], Tandem features trained on English CTS data are shown to improve performances when used in other domains (meeting data) within the language and even in other languages (Mandarin and Arabic). Even though MLPs are trained on different phonesets in different languages, Tandem features are able to capture common phonetic distinctions among languages and improve performances of conventional acoustic features.

In this paper, we investigate an alternate approach to build LVCSR system using limited amounts of training data with additional sources of knowledge - multi-stream features and cross-lingual data. With a goal of enhancing the features derived from limited amount of in-language CTS training data, we try to achieve the following -

- We propose to use CTS data from other languages to train cross-lingual MLPs and use limited amount of in-language training data to adapt these MLPs. Unlike in [8] where several hundreds of hours of CTS data are used to train MLPs on different languages, we investigate if smaller amounts of cross-lingual data will suffice to provide useful representation of speech.

- We explore the usefulness of different feature representations of the speech signal as additional sources of information, for LVCSR system. In our previous work, we have shown that combining feature streams at the posterior level is useful for ASR tasks [5]. In this paper, we show the effectiveness of two types of feature streams - short-term spectral features and long-term modulation features for low resource LVCSR.

The reminder of the paper is organized as follows. In section 2, we briefly describe the features that we use for our experiments. We discuss how we train MLPs using these features and combine them to derive Tandem features for the low resource task in Sec. 3. A mutual information based scheme to adapt cross-lingual MLP systems is discussed in Sec. 4. Sec. 5 talks about the experiments and results using the proposed technique. In Sec. 6, we conclude with a discussion of the results.

2. Spectral envelope and Modulation frequency features

FDLP is an efficient technique for auto regressive (AR) modeling of temporal envelopes of a signal. In this technique, we first apply the discrete cosine transform (DCT) on long segments of speech to obtain a real valued spectral representation of the signal. The DCT of the signal is decomposed using critical-band-sized windows. Linear prediction is performed on each sub-band DCT signal to obtain a parametric model of its temporal envelope. We compute a spectrogram of speech by stacking the individual sub-band temporal trajectories derived using FDLP [5]. Short segments of this spectrogram are used to yield short-term spectral representations of the speech signal. We also extract modulation frequency components after compressing the sub-band temporal envelopes using static and adaptive compression techniques [5].

We derive short-term features (FDLP-S) from sub-band temporal envelopes, modeled using FDLP by integrating the envelopes in short term frames (of the order of 25 ms with a shift of 10 ms). These short term sub-band energies are converted into 13 cepstral features along with their first and second derivatives [5]. Each frame of these spectral envelope features is used with a context of 9 frames for training an MLP network. To extract modulation frequency features (FDLP-M), we first compress the sub-band temporal envelopes statically using the logarithmic function and dynamically with an adaptation circuit consisting of five consecutive nonlinear adaptation loops. The compressed temporal envelopes are then transformed using the DCT in long term windows (200 ms long, with a shift of 10 ms). We use 14 modulation frequency components from each cosine transform, yielding modulation spectrum in the 0-35 Hz range with a resolution of 5 Hz [5]. The static and dynamic modulation frequency features of each sub-band are stacked together and used to train an MLP network. For telephone channel speech, we use 17 bark spaced bands for extracting these features.

3. Deriving Cross-lingual and Multi-stream posterior features

We use spectral and modulation features described in the previous section to train MLP systems. In our earlier work [5], we train an MLP on each of the feature streams and combine the features at the phoneme posterior level. These phoneme posteriors are then gaussianized by transforming them using the log function and decorrelated using the Karhunen-Loeve Transform (KLT) [3]. This reduces the dimensionality of the feature vectors by retaining only the feature components which contribute most to the variance of the data. Experiments using these Tandem features provided significant improvements over the baseline systems for several ASR tasks [5]. To enhance the features derived from limited amounts of training data using these features, we use the paradigm illustrated in Figure 1.

- We train cross-lingual MLP systems on data from two other languages - German and Spanish using a phoneset that covers phonemes from both the languages. We derive spectral envelope and modulation frequency features from 15 hours of German and 16 hours of Spanish data. Even though these languages have different phones from English, they share several common phonetic attributes of speech. The cross-lingual MLPs capture these attributes from each of the different features streams for that language.
- We train a set of low resource MLP systems for each of the feature streams by adapting the cross-lingual system using 1 hour of English data. By adapting the nets it is observed that the systems are able to discriminate better between phonetic classes of the low resource language. The primary challenge in adapting an MLP system using additional data from different language is to effectively map the phonetic units of the new language to the phoneset on which the system has already been trained. We use a mutual information technique (Section 4) to find a mapping between the existing and new language phonesets. This adaptation allows the systems to capture information about phonetic classes from the acoustic signal enhanced along with common phonetic attributes from the other languages.
- Posterior features from the two acoustic streams (FDLP-M and FDLP-S) are combined at the posterior level. This allows us to obtain more accurate and robust estimates of posteriors. Posterior features corresponding to 1 hour of data are gaussianized, decorrelated and dimensionality reduced to 30 dimensional Tandem features. These features are used to train the subsequent HMM-GMM system.

4. Adapting Cross-lingual systems

The phoneset of the cross-lingual system is different from that of the low resource language. In order to describe the low resource training data in terms of cross-lingual phoneset, we use a mutual information (MI) approach. When trained on sufficient amounts of data, it has been shown that MLPs estimate Bayesian a posteriori probabilities of phoneme classes [12]. The accumulated posterior outputs can also be considered as soft counts corresponding to the presence or absence of different phoneme classes.

In our approach, the first step is to forward pass the low resource training data (in-language) through the cross lingual MLP to obtain phoneme posteriors. Using these posterior probabilities (described in terms of cross-lingual phoneset) and its true label from the low resource phoneset, we now estimate the following counts -

$c(x)$ - total instances when a particular label x of low resource phoneset is present in the input.

$c(y)$ - accumulated posterior value for cross-lingual phoneme y.

$c(x, y)$ - accumulated posterior value for cross-lingual phoneme y when x is the true label.

Using these counts, we now find

$$MI(x, y) = \frac{c(x, y)}{c(x)c(y)}$$

For each label x, the more frequently a particular label y occurs, higher the value of $MI(x, y)$. This measure can hence be used to map a label in cross-lingual phoneset with a particular label in low-resource phoneset.

In our experiments we first train a cross-lingual MLP using German and Spanish data on a phoneset of 52 phoneset (combined set of phonemes which cover German and Spanish data). One hour of English data is forward passed using the cross lingual MLP to obtain phoneme posteriors in terms of 52 cross-lingual phones. The true labels for English data contains 47 English phonemes. Using the mapping technique described above

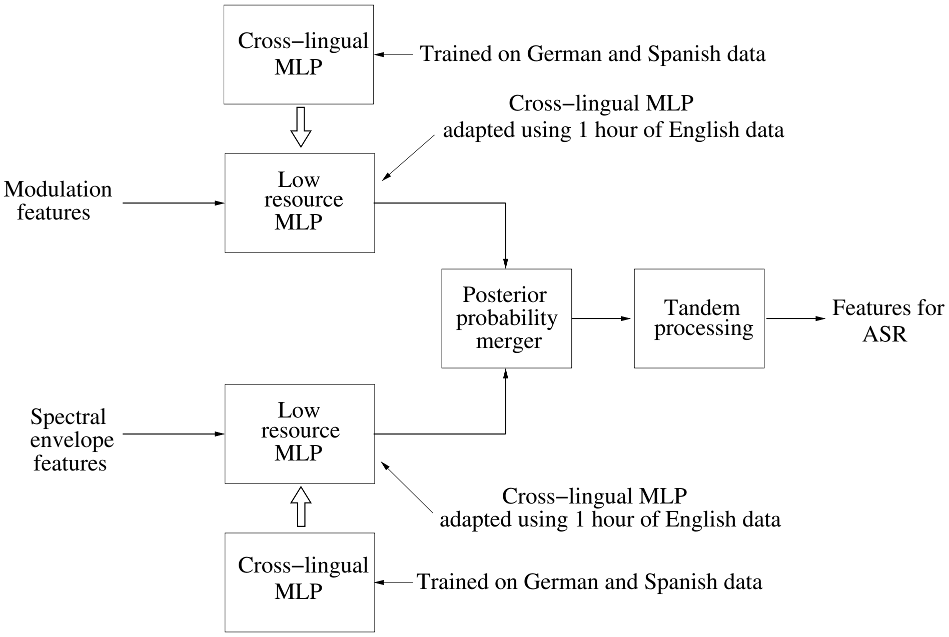
Cross−lingual MLP
Trained on German and Spanish data
Cross−lingual MLP adapted using 1 hour of English data
Modulation features
Low resource MLP
Posterior probability merger
Tandem processing
Features for ASR
Spectral envelope features
Low resource MLP
Cross−lingual MLP adapted using 1 hour of English data
Cross−lingual MLP
Trained on German and Spanish data

Figure 1: Deriving cross-lingual and multi-stream posterior features for low resource LVCSR systems

we then determine to which phone in the German-Spanish set we can map phones each English phoneme to. Each English phoneme is mapped to the phone which gives the highest MI score in the German-Spanish set. Once the English data has been mapped, the cross-lingual MLPs are adapted using 1 hour of English data. We adapt the MLP by retraining it using the new data after initializing it with its original weights.

5. Experiments and Results

The Callhome English, German and Spanish databases [9, 10, 11] used in our experiments are part of the Callhome corpora collected by LDC. 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 [9]. We use 1 hour of randomly chosen speech covering all the speakers from the complete train set for our experiments. The 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. We use 15 hours of German and 16 hours of Spanish for training the MLPs. To train the MLP systems, we use 47 phone targets for English, 46 for German and 28 for Spanish. To train systems on both German and Spanish we use a combined phoneset with 52 targets. These are derived from the Callhome lexicons for these 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. A three layered MLP with soft max nonlinearity at the output nodes is used to estimate the phoneme posterior probabilities. The network is 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. 5000 hidden nodes are used for the MLP systems while training the cross-lingual MLP systems. The base-line Tandem systems (Table 1) are trained on 1 hour of data using 800 nodes in the hidden layer of MLP.

Table 1: Word Recognition Accuracies (%) using different Tandem features derived using only 1 hour of English data

39 dimensional PLP features used directly to train HMM-GMM system	28.8
Tandem features derived from PLP features with 9 frame context	28.7
Tandem features derived from FDLP-S features with 9 frame context	29.3
Tandem features derived from 476 dimensional FDLP-M features	27.2

We use 30 dimensional Tandem features to train the subsequent single pass HTK based recognizer with 600 tied states and 4 mixtures per state. We use fewer states and mixtures per state since the amount of training data is low. The recognizer uses a 62K trigram 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.

Table 1 summarizes the baseline results for our experiments using different features with only 1 hour of English data. In our second set of experiments we derive Tandem features for

Table 2: Word Recognition Accuracies (%) using Tandem features enhanced using cross-lingual posterior features

Tandem features derived from Cross-lingual systems	FDLP-S	FDLP-M
System 1 - Trained on German data	30.6	27.9
System 2 - Trained on German and Spanish data	30.9	29.4
System 3 - Trained on German (System 1) and adapted with 1 hr of English	32.3	29.9
System 4 - Trained on German and Spanish (System 2) further adapted with 1 hr of English	33.1	30.2

Table 3: Word Recognition Accuracies (%) using multi-stream cross-lingual posterior features

Baseline PLP features	28.8
Multi-stream Cross-lingual Tandem features	36.5

the 1 hour of English data from the cross-lingual systems. It is clear that systems built using low amounts of training data perform very poorly. Our subsequent experiments aim to improve these performances using multi-stream and cross-lingual data. Table 2 shows the experiments using Tandem features derived from the spectral envelope and modulation features using the cross-lingual systems. These experiment show the improvements as more cross-lingual data is used. Adapting the systems with the limited amount of in-domain language improves the performance of each system further. As described earlier posterior streams derived from two different feature representations are now combined to derive better representations.

Table 3 shows the results of combining posterior streams from the final cross-lingual systems (System 4 of Table 2) of both streams the using the Dempster Shafer (DS) theory of evidence [15]. The results show significant improvements after combining posterior streams over the results from individual streams compared to the baseline PLP system. We also note the recent work reported in [16] which uses a similar experimental setup with cross-lingual training of Subspace Gaussian Mixture Models (SGMM). While the SGMM approach improves acoustic models, the proposed approach focuses on improving feature representations for low resource applications.

6. Conclusions

In this paper we investigated how the performance of recognizers built using limited amounts of training data can be improved using Tandem features. We examined the use of cross-lingual data by training MLPs on two different languages. Adapting these cross-lingual systems with limited amounts of data provides an absolute improvement of 4% in our system. These improvements are obtained using relatively less amounts of cross-lingual data. We also explored the use of multiple feature streams in improving performances. Combining Tandem features from various posteriors streams provides a further improvement of 3.5% in the LVCSR system. Future work will include a more studies on how in-domain data can be used along with this framework and how the performances will improve with more amounts of cross-lingual data.

7. References

[1] S. Davis and P. Mermelstein, “Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences”, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. 28, no. 4, pp. 357-366, 1980.

[2] H. Hermansky, “Perceptual linear predictive (PLP) analysis of speech”, The Journal of the Acoustical Society of America, vol. 87, pp. 1738-1752, 1990.

[3] H. Hermansky, D.P.W. Ellis and S. Sharma, “Tandem connectionist feature extraction for conventional HMM systems”, in IEEE ICASSP, 2000, pp. 1635-1638.

[4] G. Zavaliagkos et.al., “Using Untranscribed Training Data to Improve Performance”, in ISCA ICSLP, 1998.

[5] S. Thomas, S. Ganapathy and H. Hermansky, “Tandem Representations of Spectral Envelope and Modulation Frequency Features for ASR”, in ISCA INTERSPEECH, 2009.

[6] S. Sivadas and H. Hermansky, “On use of task independent training data in tandem feature extraction”, in IEEE ICASSP, 2004, pp. 541-544.

[7] Q. Zhu, B. Chen, N. Morgan and A. Stolcke, “On using MLP features in LVCSR”, in ISCA INTERSPEECH, 2004, pp. 921-924.

[8] A. Stolcke et. al., “Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons”, in IEEE ICASSP, 2006, pp. 321-324.

[9] A. Canavan, D. Graff, and G. Zipperlen, “CALLHOME American English Speech,” Linguistic Data Consortium, 1997.

[10] A. Canavan, D. Graff, and G. Zipperlen, “CALLHOME German Speech,” Linguistic Data Consortium, 1997.

[11] A. Canavan and G. Zipperlen, “CALLHOME Spanish Speech,” Linguistic Data Consortium, 1997.

[12] M.D. Richard and R.P. Lippmann, “Neural network classifiers estimate Bayesian a posteriori probabilities”, Neural Computation, vol. 3, pp. 461-483, 1991.

[13] J.J. Godfrey el. al., “Switchboard: Telephone speech corpus for research and development,” in IEEE ICASSP, 1992, pp. 517-520.

[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 IEEE ICASSP, 2007, pp. 1129-1132.

[16] L. Burget et. al., “Multilingual Acoustic Modeling for Speech Recognition based on Subspace Gaussian Mixture Models”, in IEEE ICASSP, 2010.