ISCA Archive
http://www.isca-speech.org/archive

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

Data-driven Posterior Features for Low Resource Speech Recognition Applications

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

[1]Center for Language and Speech Processing,
[2]Human Language Technology Center of Excellence,
The Johns Hopkins University, Baltimore, USA.
{samuel, ganapathy, aren, hynek}@jhu.edu

Abstract

In low resource settings, with very few hours of training data, state-of-the-art speech recognition systems that require large amounts of task specific training data perform very poorly. We address this issue by building data-driven speech recognition front-ends on significant amounts of task independent data from different languages and genres collected in similar acoustic conditions as the data in the low resource scenario. We show that features derived from these trained front-ends perform significantly better and can alleviate the effect of reduced task specific training data in low resource settings. The proposed features provide a absolute improvement of about 12% (18% relative) in an low-resource LVCSR setting with only one hour of training data. We also demonstrate the usefulness of these features for zero-resource speech applications like spoken term discovery, which operate without any transcribed speech to train systems. The proposed features provide significant gains over conventional acoustic features on various information retrieval metrics for this task.

Index Terms: Low-resource speech recognition, spoken term discovery, posterior features.

10.21437/Interspeech.2012-242

1. Introduction

Acoustic models for state-of-the-art speech recognition systems are typically trained on several hundred hours of task specific training data. In low resource scenarios, there is often very less annotated acoustic data to properly train these models. A potential solution to this problem is to use transcribed data available from other languages to augment models for the low resource language. In [1, 2] multilingual data available from other high resource languages is incorporated using a common phone set with data from the low resource language. Recently multilingual training with Subspace Gaussian Mixture Models (SGMM) have also been proposed to train acoustic models [3].

An alternative approach to this problem moves the focus from using the shared data to build acoustic models, to training data-driven feature front-ends. The key element in this data-driven approach [4], is a multi-layer perceptron (MLP) which is trained on large amounts of task independent data. Posterior features corresponding to limited task specific data are derived using the trained MLP. Since the neural network has been trained on large amounts of data, raw acoustic feature vectors are transformed by suppressing unwanted variabilities like channel and speaker effects. The discriminative training also enhances the features to differentiate between speech classes. Since prior training is used for feature extraction, the data-driven features can compensate significantly for the lack of large amounts of task specific data in downstream speech applications.

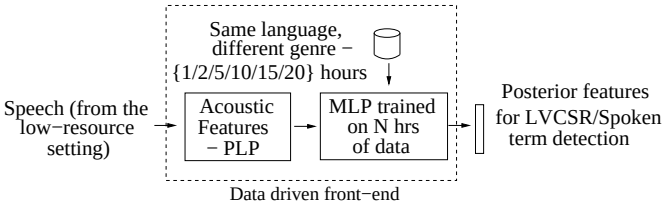

Figure 1: *Data driven front-end built using data from the same language but from a different genre*

We build on this approach for training data-driven front-ends for speech applications in low-resource settings. Two kinds of task independent data sources are used in building the front-end:

1. Up to 20 hours of data from the same language collected for a different task. Although this data has a different genre, it has similar acoustic channel conditions as the low resource data.
2. 200 hours of data from a different language but with similar acoustic channel conditions.

We train different configurations of MLPs on various amounts of these matched and mismatched data.

The rest of the paper is organized as follows. In the next section we describe how we build the data-driven front-ends. Sections 3 and 4 describe applications of the proposed features for two tasks - low-resource speech recognition and spoken term discovery. The paper concludes with a discussion in Section 5.

2. Data-driven Front-ends

As described earlier, the central elements of the proposed data-driven feature front-ends are trained MLPs. We build two kinds of front-ends on varying amounts of task independent training data from different languages.

1. A *monolingual* front-end trained on varying amounts of data from the same language as the low-resource task.

The research presented in this paper was partially funded by the DARPA RATS program under D10PC20015 and the JHU Human Language Technology Center of Excellence. 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 DARPA or the JHU HLTCOE.

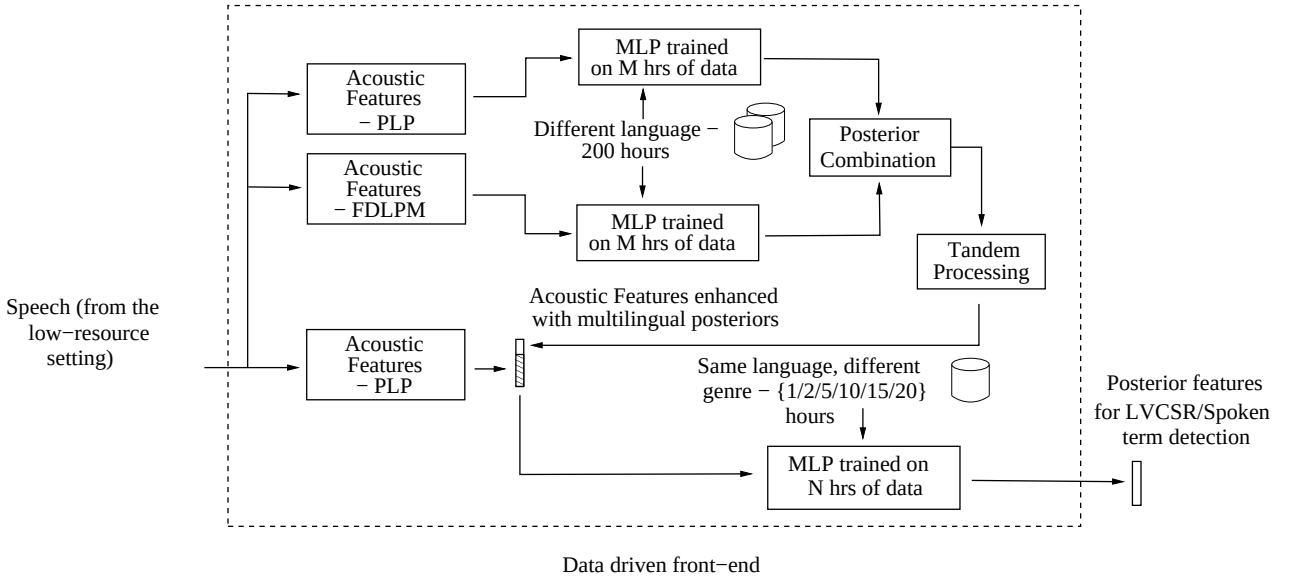

Figure 2: *A cross-lingual front-end built with data from the same language and with large amounts of additional data from a different language but with same acoustic conditions*

As shown in Fig. 1, we train different configurations of this front-end on 1 to 20 hrs of data (*N* hours). The primary advantage of this kind of a front-end is that even though the genre is different, the MLP learns useful information that characterizes the acoustics of the language. This improves as the amount of training data increases. For our current experiments we also choose task independent data from similar acoustic conditions as the low resource setting. Features generated using this front-end are hence enhanced with knowledge about the language and have unwanted variabilities from the channel and speaker removed. We use conventional short-term acoustic features to train these nets.

2. A *cross-lingual* front-end that uses large amounts of data from a different language. In most low-resource settings, it is less likely to have sufficient transcribed data in the same language to train a monolingual front-end. However considerable resources in other languages might be available. Fig. 2 outlines the components of the cross-lingual front-end that we train to include additional data from a different language. This front-end has two parts. The first part is similar to the monolingual front-end described above and consists of an MLP trained on various amounts of data from same language but different genre (*N* hours). The second part includes a set of MLPs trained on large amounts of data from a different language (*M* hours). Outputs from these MLPs are used to enhance the input acoustic features for the former part.

Although languages have common attributes between themselves, since they are transcribed using different phone sets, they first need to be combined before they can be used. In [5, 6] we use two different approaches to deal with this - a count based data driven approach to find a common phone set and an MLP training scheme with intermediate language specific layers. Both these approaches finally involve adaptation of multilingual MLPs to the low-resource language. In this work, we do not adapt any MLPs, instead we keep the front-end fixed by using the multilingual MLP to derive posterior features.

When MLPs trained on a particular language are used to derive phoneme posteriors from a different language, the language mismatch results in less sharp posteriors than from an MLP trained on the same language. However an association can still be seen between similar speech sounds from the different languages. We use this information to enhance acoustic features of the task specific language. Phoneme posteriors from two complimentary acoustic streams are combined to improve the quality of the posteriors before they are converted to features using the Tandem technique [7]. The multilingual posterior features are finally appended to short-term acoustic features to train a second level of MLPs on varying amounts of data from the same language as the low-resource task. This procedure is identical to the monolingual front-end training described above.

We compare the use of both these front-ends for two different low-resource applications. The first example is an application where there is transcribed data, however the amount of data (1 hour) is miniscule compared with the typical amounts of data used in LVCSR tasks. The second example assumes no transcripts and tries to find repeated words or phrases in speech.

3. Low-resource LVCSR Experiments

A. Low-resource setting - We design a low-resource setting using only 1 hour of training data from Callhome English database [8]. The complete English database consists of 15 hours of speech from 120 spontaneous telephone conversations between native English speakers. 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 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 test set contains roughly 3.6 hours of speech.

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 language model with an OOV rate of 0.4%, built using the SRILM tools. The

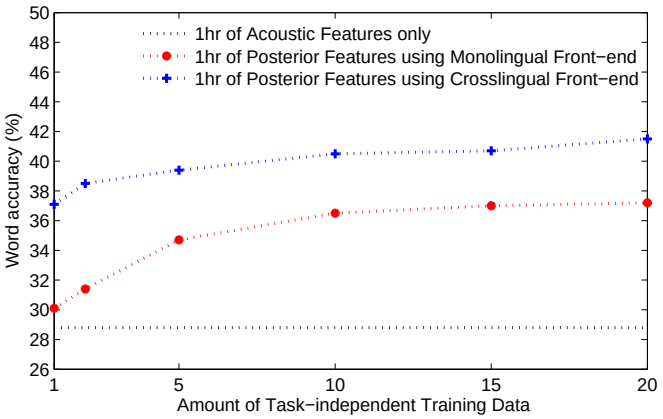

Figure 3: *LVCSR word recognition accuracies (%) with 1 hour of task specific training data using the proposed front-ends*

language model is interpolated from individual models created using the English Callhome corpus, the Switchboard corpus [9], the Gigaword corpus [10] 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.

B. Data-driven front-ends - We train two data-driven front-ends for the low-resource LVCSR task as described in Sec. 2. We train the monolingual front-end on a separate task independent training set of 20 hours from the Switchboard corpus. This training set is independent to the Callhome task but has the similar telephone channel conditions. The phone labels for this set are obtained by force aligning word transcripts to previously trained HMM/GMM models using a set of 45 phones [11]. 39 dimensional PLP features [12] (13 cepstral + Δ + $\Delta\Delta$ features) are used along with a context of 9 frames. We train seperate MLPs on subsets of 1, 2, 5, 10, 15 and 20 hours to understand how the amount of task independent data affects performance on these features.

In addition to the Switchboard corpus, we train Spanish MLPs on 200 hours of telephone speech from the LDC Spanish Switchboard and Callhome corpora for the cross-lingual front-end. Phone labels for this database are obtained by force aligning word transcripts using BBN's Byblos recognition system using 27 phones. We use two acoustic features - short-term 39 dimensional PLP features with 9 frames of context and 476 dimensional long-term modulation features (FDLPM)[13]. When networks are trained on multiple feature representations, better posterior estimates can be derived by combining the outputs from different systems using posterior probability combination rules. We use the Dempster-Shafer rule of combination for our experiments [14]. Posteriors from multiple streams are combined to reduce the effects of language mismatch and improve posteriors. 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) [7]. A dimensionality reduction is also performed by retaining only the top 20 feature components which contribute most to the variance of the data.

Table 1: Word Recognition Accuracies (%) using different amounts of Callhome data to train the LVCSR system with conventional acoustic features

	1hr	2hr	5hr	10hr	15hr
PLP features	28.8	33.60	39.70	43.80	46.50

The English MLPs in the cross-lingual setting are trained on enhanced acoustic features. These features are created by appending posterior features derived from the Spanish MLPs to the PLP features used in monolingual training. We similarly also train seperate MLPs on subsets of 1, 2, 5, 10, 15 and 20 hours of task independent data.

C. Experiments with various front-ends - In our first experiment we use 39 dimensional PLP features directly for the 1 hour Callhome LVCSR task. As we hypothesize, the acoustic models are insufficiently trained, resulting in a low word accuracy of 28.8%. These features are then replaced by 25 dimensional posterior features using the monolingual and cross-lingual front-ends, each trained on varying amounts of task independent data from the Switchboard corpus. Fig. 3 shows how the performance changes for both the monolingual and cross-lingual systems. Using the data-driven front-ends, the word accuracy improves from 28.8% to 30.1% and 37.1% with just 1 hour of task independent training data using the monolingual and cross-lingual front-ends respectively. These improvements continue to 37.2% and 41.5% with the same 1 hour of Callhome LVCSR training data as the amount of task-independent data is increased for both the front-ends. We draw the following conclusions from these experiments -

1. With very few hours of task specific training data, posteriors features can provide significant gains over conventional acoustic features. Table 1 shows the work accuracies when different amounts of Callhome data are used to train the LVCSR system. By using the cross-lingual front-end, features from only 1 hour of data perform close to 5-10 hours of the Callhome data with conventional features. This demonstrates the usefulness of our approach where we use task independent data in low-resource settings to generate better features.

2. When data from a different language is used, additional gains of 4-7% absolute are achieved over just using task independent data from the same language. It is interesting to observe that the performance with the cross-lingual front-end starts improving from the best performance achieved with the monolingual front-end.

4. Spoken Term Discovery Experiments

The low resource setting described above assumes the availability of transcibed data to train LVCSR systems. However in zero resouce settings, tasks such as spoken term detection attempt to automatically discover repeated words and phrases in speech without any transcriptions [15]. In recent approaches [15, 16, 17] to address this task, a dynamic time warping (DTW) search of the speech corpus is performed against itself to discover repeated patterns. With no transcripts to guide the process, results of the search largely depend on the quality of the underlying speech representation being used. In [18], multiple information retrival metrics have been proposed to evaluate the quality of different speech representations on this task. These metrics operate by using a large collection of presegmented word examples to first compute the DTW distance between all example pairs and then quantifying how well the DTW dis-

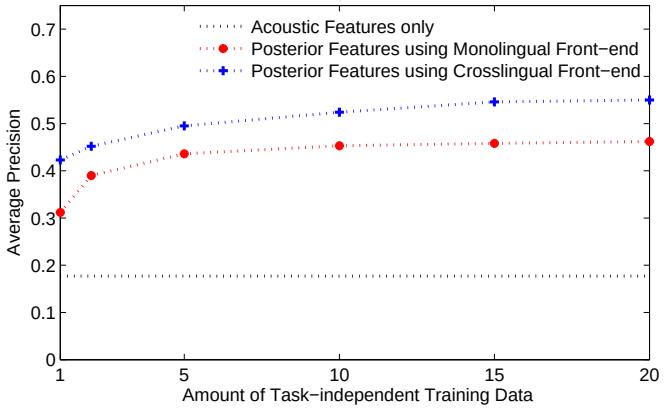

Figure 4: *Average precision for different configuraion of the proposed front-ends*

tances can differentiate between same or different the example pairs. Better scores with these metrics are indicative of good speaker independence and high word discriminabily of feature representations. Since these are also desirable properties of features for other downstream recognition applications, these metrics are also predictive of how different features will perform in those applications. In this work, we evaluate posterior features from both the multilingual and cross-lingual for spoken term discovery with information retrival metrics used in [18].

The evaluation metric uses 11K words from the Switchboard corpus resulting in 60.7M word pairs of which 96K are same word pairs [18]. Similarity between word pairs (w_i, w_j) are measured using minimum DTW alignment cost - $DTW(w_i, w_j)$ between w_i and w_j. For a particular threshold τ, w_i and w_j are considered similar if $DTW(w_i, w_j) \leq \tau$. Computing DTW distances also requires a distance metric to be defined between feature vector frames that make up words. For this evaluation cosine distance is used for comparing frames of raw acoustic features corresponding to words. A more meaningful symmetric KL-divergence is used for accessing similarities on phoneme posteriors vectors generated by the proposed front-ends for words

The entire set of word pairs is now used in the context of an information retrieval task where the goal is to retrieve same word pairs from different word impostors for each front-end configuration. Sweeping τ allows us to create a standard precision-recall curve for each setting. The precision-recall curves can then be characterized by several criteria. We use the average precision metric defined as the area under the precision-recall curve for our experiments. This metric also summarizes the average system performance across all operating points.

Fig. 4 shows the average precision scores for the two front-ends with varying amounts of training data. The plot shows that posterior features perform significantly better than the raw acoustic features (39D PLP features with zero mean/unit variance) which have a very low score of only 0.177. As in the LVCSR case, posterior features from the cross-lingual front-end perform better. Both front-ends improve as the amount of task independent data increases. Since this evaluation metric is based on DTW distances over a moderately large set of words, improved performances on this metric imply more efficient spoken term discovery.

It is also interesting to observe that there is a very strong correlation of 0.985 and 0.948 between the average precision metric and the LVCSR word accuracies. This shows the usefulness of the metric in evaluating features and predicting their performance in downstream recognition tasks.

5. Conclusions

We have considered the use of data driven front-ends to address the poor performance of acoustic models for speech recognition applications in low resource setting. We have proposed two front-ends trained on large amounts of task independent data both from the same and different languages. Our experiments show that significant improvements can be obtained using these front-ends. The results demonstrate that in low-resource settings where task dependent training data is scarce, task independent multi-lingual data can still be used to compensate for performance drops.

6. References

[1] 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.

[2] D. Imseng, H. Bourlard, P.N. Garner, “Using KL-divergence and Multilingual Information to Improve ASR for Under-resourced Languages”, Proc. of IEEE ICASSP, 2012.

[3] Y. Qian, D. Povey and J. Lu, “State-Level Data Borrowing for Low-Resource Speech Recognition based on Subspace GMMs”, Proc. of ISCA Interspeech, 2011.

[4] S. Sivadas and H. Hermansky, “On Use of Task Independent Training Data in Tandem Feature Extraction”, Proc. of IEEE ICASSP, 2004.

[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] S. Thomas, S. Ganapathy and H. Hermansky, “Multilingual MLP Features For Low-resource LVCSR Systems”, Proc. of IEEE ICASSP, 2012.

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

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

[9] J.J. Godfrey et. al., “Switchboard: Telephone speech corpus for research and development”, Proc. of IEEE ICASSP, 1992.

[10] D. Graff, “English Gigaword”, Linguistic Data Consortium, 2003.

[11] T. Hain et. al., “The AMI system for the transcription of speech in meetings”, Proc. of ISCA Interspeech, 2007.

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

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

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

[15] A. Jansen, K. Church, and H, Hermansky, “Towards Spoken Term Discovery at Scale with Zero Resources”, Proc. of ISCA Interspeech, 2010.

[16] A. Muscariello, G. Gravier, and F. Bimbot, “Audio Keyword Extraction by Unsupervised Word Discovery”, Proc. of ISCA Interspeech, 2009.

[17] Y. Zhang and J.R. Glass, “Towards Multi-speaker Unsupervised Speech Pattern Discovery”, Proc. of IEEE ICASSP, 2010.

[18] M.A. Carlin, S. Thomas, A. Jansen, and H. Hermansky, “Rapid Evaluation of Speech Representations for Spoken Term Discovery”, Proc. of ISCA Interspeech, 2011.