Cross-lingual Techniques and Use of Articulatory Features in Acoustic Modeling of Low-Resource Languages

Basil Abraham

Indian Institute of Technology-Madras, India

ee11d032@ee.iitm.ac.in

With recent advancements in deep neural networks (DNN), there has been a significant improvement in the performance of speech recognizers. However such robust systems are mainly limited to popular languages like English, French etc. To build a robust speech recognizer in any language, large amount of transcribed speech data is required. In many of the under resourced languages like Indian and African languages, data sparsity is a critical problem in building good speech recognizers. Throughout this work, we use the term *high-resource* language to refer to a language having abundant resources in terms of transcribed training data and *low-resource* language to the one with limited training data.

In this work we have tried to leverage the resources form high-resource languages to build better acoustic model for low-resource languages. We proposed three approaches to overcome the data sparsity in acoustic modeling. In the first approach, speech data was pooled from high-resource language to train acoustic model for the low-resource language. Secondly, the cross-lingual approaches of borrowing the model parameters from a well-trained model were used. In the third case, pseudo articulatory features were obtained from articulatory classifiers trained in high-resource languages.

1. Pooling Data

The problem of data insufficiency is addressed by pooling data from closely related languages. The pooled data along with the data from low-resource language is used for building acoustic models like continuous density hidden Markov model (CDHMM), subspace Gaussian mixture Model (SGMM) [1], phone cluster adaptive training (Phone-CAT) [2], deep neural network (DNN) [3] and convolutional neural network (CNN) [4]. The major challenge in pooling data was the difference in phone set between the languages. In our work, we proposed an automated technique to map the phones in one language to another to facilitate data pooling [5]. The proposed mapping technique is based on the center-phone capturing property of interpolation vectors emerging from the recently proposed Phone-CAT method.

Phone-CAT is an acoustic modeling technique that belongs to the broad category of canonical state models (CSM) that includes SGMM. In Phone-CAT, the interpolation vector belonging to a particular context-dependent state has maximum weight for the center-phone of monophone clusters. In the proposed technique the context-dependent states of the low-resource language is represented as the interpolation of the high-resource language monophone clusters as shown in Figure 1. The conventional mapping technique uses decoding of low-resource language data with high-resource model and compare it against the true phone sequence [6]. As compared to the conventional technique, the interpolation vectors of Phone-CAT uses data belong to a context-dependent state covering different utterances to generate the mapping. To achieve further improvements, the data pooled models are then adapted towards the low-resource language.

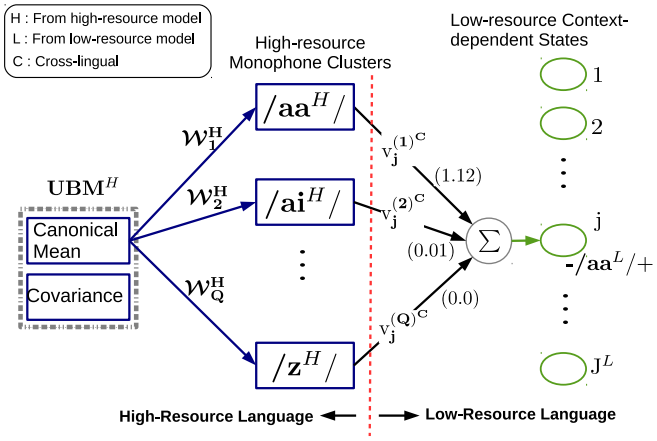

Figure 1: *Block schematic of phone mapping technique using Phone-CAT*

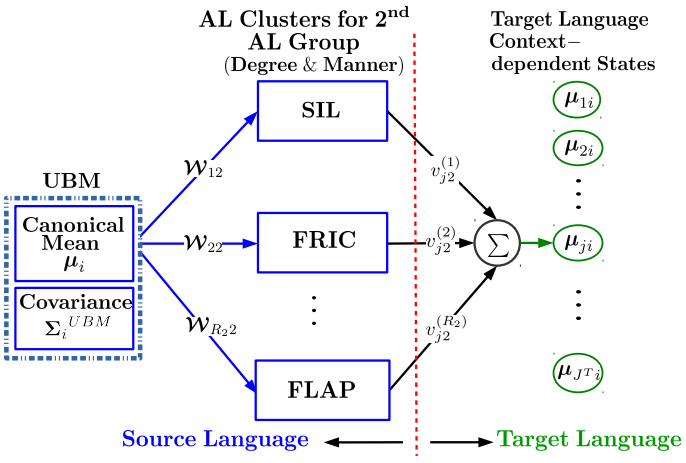

Figure 2: *Block schematic of phone-to-AL mapping technique using Phone-CAT(AFV-CAT) in Degree & Manner AL group*

2. Borrowing Model Parameters

In the second approach, acoustic model parameters are borrowed from models SGMM, Phone-CAT, DNN and CNN built with high-resource language and are then further refined using low-resource data. In both SGMM and Phone-CAT, global

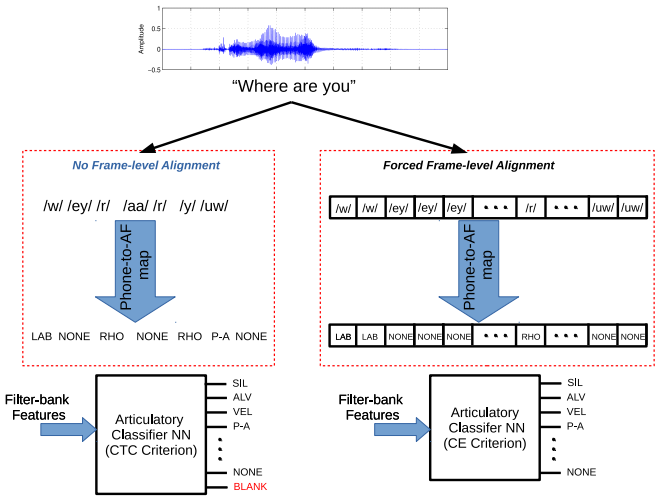

Figure 3: *Comparison of articulatory classifier training using CTC criterion and cross-entropy criterion*

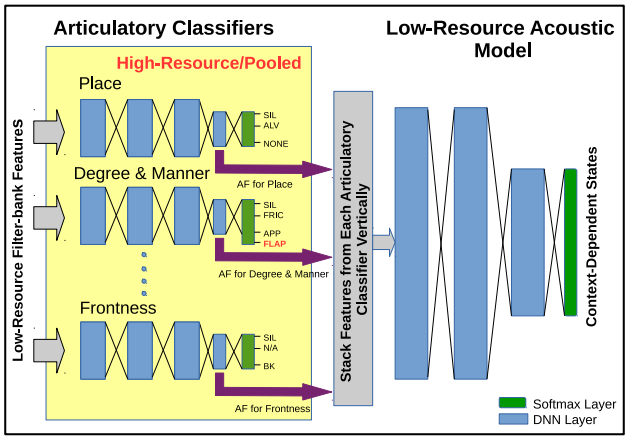

Figure 4: *Proposed joint estimation framework for training acoustic model using articulatory features*

parameters are borrowed from the the corresponding high-resource model and the language specific parameters are trained with the low-resource language [7, 5]. In the case of DNN and CNN, the hidden layers of the high-resource language model are borrowed and the output layer with low-resource language tied states is then trained with low-resource language data [8, 9].

3. Articulatory Features for ASR

Recent studies have shown that in the case of under-resourced languages, use of articulatory features (AF) emerging from an articulatory model results in improved automatic speech recognition (ASR) compared to conventional mel frequency cepstral coefficient (MFCC) features. Articulatory features are more robust to noise and pronunciation variability compared to conventional acoustic features [10, 11, 12]. To extract articulatory features, one method is to take conventional acoustic features like MFCC and build an articulatory classifier that would output articulatory features (known as pseudo-AF). However, these classifiers require a mapping from phone to different articulatory labels (AL) (e.g., place of articulation and manner of articulation), which is not readily available for many of the under-resourced languages. In our work, we have proposed an automated technique to generate phone-to-articulatory label (phone-to-AL) mapping for a new target language based on the knowledge of phone-to-AL mapping of a well-resourced language like English [13]. The proposed mapping technique also exploit the center-phone capturing property of interpolation vectors in Phone-CAT. In this context the Phone-CAT acoustic model was modified to include the articulatory information of the phones. We call this as AFV-CAT and is shown in Figure 2. In AFV-CAT, the interpolation vector belonging to a particular context-dependent state has maximum weight for the articulatory label corresponding to the center-phone. These relationships from the various context-dependent states are used to generate a phone-to-AL mapping.

One of the major problems faced in building articulatory classifiers is the requirement of speech data aligned in terms of articulatory feature values at frame level. Manually aligning data at frame level is a tedious task and alignments obtained from the phone alignments using phone-to-AL feature mapping are prone to errors. In this work, a technique using connectionist temporal classification (CTC) criterion [14] to train an articulatory classifier using bidirectional long short-term memory (BLSTM) recurrent neural network (RNN) is proposed [15]. The CTC criterion eliminates the need for forced frame level alignments as shown in Figure 3. Articulatory classifiers were also built using different neural network architectures like deep neural networks (DNN), convolutional neural network (CNN) and BLSTM with frame level alignments and were compared to the proposed approach of using CTC. Among the different architectures, articulatory features extracted using articulatory classifiers built with BLSTM gave better recognition performance. Further, the proposed approach of BLSTM with CTC gave the best overall performance.

The performance of the articulatory features depends on the efficacy of this classifier. But, training such a robust classifier for a low-resource language is constrained due to the limited amount of training data. We can overcome this by training the articulatory classifier using a high resource language. This classifier can then be used to generate articulatory features for the low-resource language. However, this technique fails when high and low-resource languages have mismatches in their environmental conditions. In this work, we address both the aforementioned problems by jointly estimating the articulatory features and low-resource acoustic model [16] as in Figures 4.

In this study, we have used three under-resourced Indian languages namely Assamese, Hindi and Tamil. Experiments were performed in the cross-lingual scenario by assuming a small training data set ($\approx$ 2 hours) from each of the Indian languages and high-resource data-set ($\approx$ 22 hours) from other languages. All proposed techniques gave improved performance compared to the monolingual acoustic models built in that low-resource language.

4. Acknowledgements

I thank my adviser Prof. Umesh S for all his help and guidance. This work was supported in part by the consortium project titled "Speech-based access to commodity price in six Indian languages", funded by the TDIL program of DeITY of Govt. of India. The authors would like to thank consortium members involved in collecting Assamese, Hindi and Tamil corpus.

5. References

[1] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. C. Rose, P. Schwarz, and S. Thomas, "The Subspace Gaussian Mixture Model - A Structured Model for Speech Recognition," *Computer Speech & Language*, vol. 25, no. 2, pp. 404–439, 2011.

[2] V. Manohar, B. S. Chinnari, and S. Umesh, "Acoustic Modeling using Transform-Based Phone-Cluster Adaptive Training," in *Proc. ASRU*, December 2013, pp. 49–54.

[3] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[4] T. N. Sainath, A.-r. Mohamed, B. Kingsbury, and B. Ramabhadran, "Deep convolutional neural networks for lvcsr," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8614–8618.

[5] Basil Abraham, Neethu Mariam Joy, Navneeth K., and S. Umesh, "A Data-Driven Phoneme Mapping Technique Using Interpolation Vectors of Phone-Cluster Adaptive Training," in *Proc. SLT*, December 2014, pp. 36–41.

[6] T. Schultz and A. Waibel, "Experiments on cross-language acoustic modeling." in *INTERSPEECH*, 2001, pp. 2721–2724.

[7] L. Lu, A. Ghoshal, and S. Renals, "Cross-lingual Subspace Gaussian Mixture Models for Low-resource Speech Recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 1, pp. 17–27, Jan 2014.

[8] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7304–7308.

[9] Basil Abraham S. Umesh and Neethu Mariam Joy, "Overcoming data sparsity in acoustic modeling of low-resource language by borrowing data and model parameters from high-resource languages," in *INTERSPEECH*, 2016.

[10] K. Kirchhoff, G. A. Fink, and G. Sagerer, "Combining acoustic and articulatory feature information for robust speech recognition," *Speech Communication*, vol. 37, no. 3, pp. 303–319, 2002.

[11] J. Frankel, M. Magimai-doss, S. King, K. Livescu, and Ö. Çetin, "Articulatory feature classifiers trained on 2000 hours of telephone speech," in *In Proc. Interspeech*, 2007.

[12] O. Cetin, A. Kantor, S. King, C. Bartels, M. Magimai-Doss, J. Frankel, and K. Livescu, "An articulatory feature-based tandem approach and factored observation modeling," in *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 4. IEEE, 2007, pp. IV–645.

[13] Basil Abraham and S. Umesh, "An automated technique to generate phone-to-articulatory label mapping," *Speech Communication*, vol. 86, pp. 107 – 120, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0167639316300206

[14] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.

[15] Basil Abraham S. Umesh and Neethu Mariam Joy, "Articulatory feature extraction using ctc to build articulatory classifiers without forced frame alignments for speech recognition," in *INTERSPEECH*, 2016.

[16] Basil Abraham S. Umesh and Neethu Mariam Joyh, "Joint estimation of articulatory features and acoustic models for low-resource languages," in *to be presented in INTERSPEECH*, 2017.