Discriminative learning with latent articulatory variables

Eric Fosler-Lussier[1], Preethi Jyothi[1], Joseph Keshet[2],
Karen Livescu[3], Rohit Prabhavalkar[1], Hao Tang[3]

[1]The Ohio State University, USA [2]Bar-Ilan University, Israel [3]TTI-Chicago, USA

fosler@cse.ohio-state.edu, jyothi@ohio-state.edu, keshet@cs.biu.ac.il,
klivescu@ttic.edu, prabhava@cse.ohio-state.edu, haotang@ttic.edu

Abstract

We review our recent work, which includes several disparate approaches and tasks, all with the goal of using latent articulatory structure while learning discriminatively.[1][2]

1. Introduction

A major challenge in articulatory approaches to speech recognition is that it is difficult to obtain ground-truth articulatory data. This is certainly the case at test time, where it is infeasible to collect such data, but also the case at training time, where only relatively small databases including articulatory information exist. Moreover, the types of articulatory data that exist have some drawbacks. There are two main types of articulatory data available: continuous physical measurements via electromagnetic articulography, X-ray microbeam, MRI, and so on; and manual labels of discrete articulatory features. The former often exclude some measurements (e.g., velum, voicing) and are difficult to normalize across speakers; the latter is extremely time-consuming to obtain directly and quite noisy when obtained from phonetic transcriptions.

The approach we take here minimizes the use of any articulatory data, building articulatory structure into the model using knowledge from phonology and speech science. The articulatory variables are always hidden (latent) at both train and test time. In particular, we are interested in training such models discriminatively, which has been very successful for phone-based models but has not been widely applied for articulatory models. We review several threads of research united by this goal. We focus on conversational speech, where pronunciation variation still significantly hampers performance to this day [8].

Our starting point was a class of generative articulatory pronunciation models represented as dynamic Bayesian networks (DBNs) [9, 1], shown in Fig. 1. This model, inspired by ideas from articulatory phonology [10], describes alternative pronunciations as the result of either asynchrony between articulators or substitution of one articulatory value for another. On a lexical access task, where surface articulatory features serve as proxy for the acoustic signal (i.e., we assume that we have perfect articulatory classifiers), such models improve over typical models of phonetic substitutions, insertions, and deletions [1].

There are multiple ways to incorporate ideas from such a model into a complete system while taking advantage of discriminative training: the model can be converted to an HMM or FST and incorporated into a standard HMM- or FST-based recognizer (see Sec. 3 below and [1]); it can be applied to score/align new examples to obtain feature functions for a discriminative linear/log-linear model (see Sec. 2 and [4]); or parts of the model can be represented directly as feature functions in a linear/log-linear model (see Sec. 4 and [5, 6, 7]).

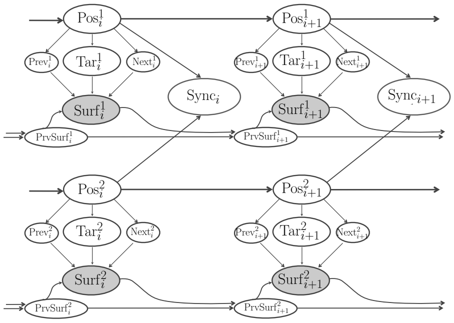

Figure 1: Articulatory DBN for pronunciation modeling [1], showing two articulatory streams over two frames i, $i+1$. Each stream includes a position variable (state index in the dictionary pronunciation); target and surface (observed) articulatory states; context of previous and next targets and previous surface state; and a "sync" variable modeling inter-articulator asynchrony.

2. Large-margin model for lexical access

Lexical access is the task of inferring, given a surface pronunciation $\overline{p}$ (a sequence of surface phonetic units), which single word w this represents from a vocabulary $\mathcal{V}$ (i.e., $|\mathcal{V}|$-way classification). We use data from the Switchboard Transcription Project (STP) [11], which is transcribed at a narrow phonetic level including nasalization, fricated stops, and so on. We define our word classifier as $w^* = \mathrm{argmax}_{w \in \mathcal{V}}\ \boldsymbol{\theta} \cdot \boldsymbol{\phi}(\overline{p}, w)$, where $\boldsymbol{\phi}(\overline{p}, w)$ is a vector of feature functions and $\boldsymbol{\theta}$ is a vector of corresponding weights, learned by a large-margin approach [4].

Some of the feature functions $\boldsymbol{\phi}$ are based on a context-independent variant of the DBN in Fig. 1. The DBN aligns each input surface pronunciation with each possible word, and we define feature functions based on counts of asynchronous articulatory states, counts of substitutions of one articulatory value for another, and the alignment score itself. Other (non-articulatory) feature functions are based on phonetic alignment between the surface form and the dictionary, word-specific counts of phonetic units, deviations in length between the dictionary and surface form, and a dictionary lookup function.

Tab. 1 shows the error rates obtained on data from STP. The large-margin model shows a sizable improvement over both previous work and a conditional log-likelihood (CLL) train-

[1] Many details and citations are left out for brevity; please see [1, 2, 3, 4, 5, 6, 7] for a more complete description and related literature.

[2] This research was supported by NSF grants IIS-0905633 and IIS-0905420. The opinions expressed in this work are those of the authors and do not necessarily reflect the views of the funding agency.

Model	ER
lexicon lookup (from [9])	59.3%
phonetic S/I/D model (from [1], based on [12])	32.1%
articulatory DBN [1]	29.1%
discriminative model, CLL training [4]	21.5%
discriminative model, large-margin training [4]	**15.2%**

Table 1: Lexical access error rates (ER). Lexicon lookup is a sanity check baseline using a lookup of the surface form in the dictionary. Phonetic S/I/D is a typical model of context-dependent phonetic substitutions, insertions, and delections.

ing approach that maximizes the word posterior $P(w|\overline{p}) \propto \exp(\boldsymbol{\theta} \cdot \boldsymbol{\phi}(\overline{p}, w))$ Most of the gain comes from discriminative training, and a small gain comes from the articulary features; however, the articulatory model used here is context-independent. Ongoing work is extending the feature functions and incorporating the model into a complete recognizer using a segmental model to allow for our whole-word features.

3. Discriminative training of pronunciation models via conversion to FSTs

Sec. 2 described articulatory model-based feature functions in a large-margin whole-word classifier. Such models can also be incorporated into traditional frame-based ASR pipelines. In [2], we developed an approach for compiling DBNs into cascades of Weighted Finite State Transducers (WFSTs). Both DBNs and WFSTs have a Markovian structure that encodes time; however, the state space of the DBNs within a time step is much larger than that of typical WFSTs in ASR. Variables in the DBN can be grouped based on the DBN's conditional independence assumptions to vastly speed up decoding: we subdivide the DBN into multiple levels, attempting to minimize the number of edges between levels. Each level becomes a WFST (across time); the composition of all of these WFSTs represents the whole DBN.

There are two advantages to this approach: first, we can apply discriminative training techniques for WFSTs in order to improve performance. In [2], we experimented with an averaged perceptron approach for updating WFST weights,which improved lexical access over a baseline DBN (from 43.0% to 35.9% error).[3] The second advantage is that this can be implemented in a standard WFST-based ASR system taking acoustic input. We developed a discriminative approach for training individual WFST factors within cascades to allow training pronunciation models with a small amount of data. We first applied this technique to a phone-based (rather than articulatory) WFST system, which significantly improved isolated-word recognition; promising preliminary results on continuous speech were observed as well [3]. Interestingly, discriminatively training the lexical WFST was more effective than discriminatively training the entire WFST cascade with a small amount of data. We have also seen some promising results using articulatory models for the isolated-word recognition task.

4. Discriminative articulatory models for spoken term detection

We have also explored the idea of jointly modeling the articulatory configuration space using factored conditional random fields (CRFs) structured similarly to the generative DBN of Fig. 1. While factored CRFs can be computationally expensive, restricting the amount of asychrony limits the CRF state space to the point where efficient polynomial-time inference can be used. Articulatory alignments produced by the factored CRF were more accurate than those produced by the original DBN from which the factored CRF was derived [5].

System	500	1000	2500	5000
HMM-triphone	0.828	0.855	0.899	0.920
Disc. phone	0.874^{*}	0.901^{*}	0.917	0.933^{*}
Disc. artic.	$0.888^{*,\dagger}$	0.898^{*}	0.915	0.937^{*}
Disc. ph.+artic.	$0.891^{*,\dagger}$	0.905^{*}	0.920^{*}	$0.938^{*,\dagger}$

Table 2: Spoken term detection AUC results [6, 7]. *, † = significant improvement over triphone and Disc. phone, respectively.

The ability to produce joint alignment scores allows us to implement articulatory spoken term detection systems using very little training data. Extending [13], we train a large-margin classifier to separate positive keyword examples from negative examples using features based on the factored CRF; training optimizes the area under the receiver operating characteristic curve (AUC) [6, 7]. In very low-resource conditions (training on 500-5000 Switchboard utterances), phone-based discriminative systems outperform HMM baselines, and articulatory models improve performance even further (both alone and combined with phone-based systems), as shown in Tab. 2.

5. Summary

Our work thus far has demonstrated that modeling articulatory variables as latent and using discriminative training can be effective in pronunciation modeling for lexical access, with initial results showing promise for continuous recognition and spoken term detection. Along the way, we have developed new techniques for representing such models as FSTs and new ways of applying discriminative training for FST cascades and spoken term detection. Much work remains to extend the articulatory models used in discriminative systems (which thus far have been only simplified models) and to scale up to larger tasks.

6. References

[1] P. Jyothi *et al.*, "Lexical access experiments with context-dependent articulatory feature-based models," in *ICASSP*, 2011.

[2] P. Jyothi *et al.*, "Discriminatively learning factorized finite state pronunciation models from dynamic Bayesian networks," in *Interspeech*, 2012.

[3] P. Jyothi *et al.*, "Discriminative training for WFST factors with application to pronunciation modeling," in *Interspeech*, 2013.

[4] H. Tang *et al.*, "Discriminative pronunciation modeling: A large-margin, feature-rich approach," in *ACL*, 2012.

[5] R. Prabhavalkar *et al.*, "A factored conditional random field model for articulatory feature forced transcription," in *ASRU*, 2011.

[6] R. Prabhavalkar *et al.*, "Discriminative spoken term detection with limited data," in *MLSLP*, 2012.

[7] R. Prabhavalkar *et al.*, "Discriminative articulatory models for spoken term detection in low-resource settings," in *ICASSP*, 2013.

[8] K. Livescu *et al.*, "Sub-word modeling for automatic speech recognition," *IEEE Sig. Proc. Mag.*, vol. 29, pp. 44–57, 2012.

[9] K. Livescu, "Feature-based Pronunciation Modeling for Automatic Speech Recognition," *PhD dissertation*, 2005.

[10] C.P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[11] S. Greenberg *et al.*, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," in *ICSLP*, 1996.

[12] M. Riley *et al.*, "Stochastic pronunciation modelling from hand-labelled phonetic corpora," *Speech Communication*, vol. 29, no. 2-4, pp. 209–224, 1999.

[13] J. Keshet *et al.*, "Discriminative keyword spotting," *Speech Communication*, vol. 51, no. 4, pp. 317–329, 2009.

[3] NB: these experiments did not take into account context or long-range features that were critical in [4].