

JOINT VERSUS INDEPENDENT PHONOLOGICAL FEATURE MODELS WITHIN CRF PHONE RECOGNITION

Ilana Bromberg*, Jeremy Morris†, and Eric Fosler-Lussier*†

*Department of Linguistics

†Department of Computer Science and Engineering

The Ohio State University, Columbus, OH

bromberg@ling.ohio-state.edu, {morrijer, fosler}@cse.ohio-state.edu

Abstract

We compare the effect of joint modeling of phonological features to independent feature detectors in a Conditional Random Fields framework. Joint modeling of features is achieved by deriving phonological feature posteriors from the posterior probabilities of the phonemes. We find that joint modeling provides superior performance to the independent models on the TIMIT phone recognition task. We explore the effects of varying relationships between phonological features, and suggest that in an ASR system, phonological features should be handled as correlated, rather than independent.

1 Introduction

Phonological features have received attention as a linguistically-based representation for sub-word information in automatic speech recognition. These sub-phonetic features allow for a more refined representation of speech by allowing for temporal desynchronization between articulators, and help account for some phonological changes common in spontaneous speech, such as devoicing (Kirchhoff, 1999; Livescu, 2005). A number of methods have been developed for detecting acoustic phonological features and related acoustic landmarks directly from data using Multi-Layer Perceptrons (Kirchhoff, 1999), Support Vector Machines (Hasegawa-Johnson et al., 2005; Sharenborg et al., 2006), or Hidden Markov Models (Li and Lee, 2005). These techniques typically assume that acoustic phonological feature events are independent for ease of modeling.

In one study that broke the independence assumption (Chang et al., 2001), the investigators developed *conditional detectors*: MLP detectors of acoustic phonological features that are hierarchically dependent on a different phonological class. In (Rajamanohar and Fosler-Lussier, 2005) it was shown that such a conditional training of detectors tended to have correlated frame errors, and that improvements in detection could be obtained by training joint detectors. For many features, the best detector can be obtained by collapsing MLP phone posteriors into feature classes by marginalizing across phones within a class. This was shown only for frame-level classification rather than phone recognition.

Posterior estimates of phonological feature classes, as in Table 1, particularly those derived from MLPs, have been used as input to HMMs (Launay et al., 2002), Dynamic Bayesian Networks (DBNs) (Frankel et al., 2004; Livescu, 2005), and Conditional Random Fields (CRFs) (Morris and Fosler-Lussier, 2006). Here we evaluate phonological feature detectors created from MLP phone posterior estimators (joint feature models) rather than the independently trained MLP feature detectors used in previous work.

2 Conditional Random Fields

CRFs (Lafferty et al., 2001) are a joint model of a label sequence conditioned on a set of inputs. No independence is assumed among the input; the CRF model discriminates between hypothesized label sequences according to an exponential function of weighted feature functions:

$$P(\mathbf{y}|\mathbf{x}) \propto \exp \sum_i (S(\mathbf{x}, \mathbf{y}, i) + T(\mathbf{x}, \mathbf{y}, i)) \quad (1)$$

Class	Feature Values
SONORITY	Vowel, Obstruent, Sonorant, Syllabic, Silence
VOICE	Voiced, Unvoiced, N/A
MANNER	Fricative, Stop, Stop-Closure, Flap, Nasal, Approximant, Nasalflap, N/A
PLACE	Labial, Dental, Alveolar, Palatal, Velar, Glottal, Lateral, Rhotic, N/A
HEIGHT	High, Mid, Low, Lowhigh, Midhigh, N/A
FRONT	Front, Back, Central, Backfront, N/A
ROUND	Round, Nonround, Roundnonround, Nonroundround, N/A
TENSE	Tense, Lax N/A

Table 1: Phonetic feature classes and associated values

where $P(y|\mathbf{x})$ is the probability of label sequence y given an input frame sequence x , i is the frame index, and S and T are a set of state feature functions and a set of transition feature functions, defined as:

$$S(x, y, i) = \sum_j \lambda_j s_j(y, x, i), \quad \text{and} \quad (2)$$

$$T(x, y, i) = \sum_k \mu_k t_k(y_{i-1}, y_i, x, i) \quad (3)$$

where λ and μ are weights determined by the learning algorithm. In NLP applications, the component feature functions s_j and t_k are typically realized as binary indicator functions indicating the presence or absence of a feature, but in ASR applications it is more typical to utilize real-valued functions, such as those derived from the sufficient statistics of Gaussians (e.g., (Gunawardana et al., 2005)).

We can use posterior estimates of phone classes or phonological feature classes from MLPs as feature functions (inputs) within the CRF model. A more detailed description of this CRF paradigm can be found in (Morris and Fosler-Lussier, 2006), which shows that the results of phone recognition using CRFs is comparable to that of HMMs or Tandem systems, with fewer constraints being imposed on the model. State feature functions in our system are defined such that

$$s_{\phi, f}(y_i, \mathbf{x}, i) = \begin{cases} NN_f(x_i), & \text{if } y_i = \phi \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where the MLP output for feature f at time i is $NN_f(x_i)$. This allows for an association between a phone ϕ and a feature f (even if f is traditionally not associated with ϕ).

In this study, we experiment with different methods of generating these feature functions. In various

experiments, they are generated by training MLP phone detectors, by evaluating the feature information inherent in the MLP phone posteriors, and by training independent MLPs to detect the various features within the classes described. The use of CRFs allows us to explore the dependencies among feature classes, as well as the usefulness of phone posteriors versus feature classes as inputs.

3 Experimental Setup

We use the TIMIT speech corpus for all training and testing (Garofolo et al., 1993). The acoustic data is manually labeled at the phonetic level, and we propagate this phonetic label information to every frame of data. For the feature analyses, we employ a lookup table that defines each phone in terms of 8 feature classes, as shown in Table 1. We extract acoustic features in the form of 12th order PLP features plus delta coefficients. We then use these as inputs to several sets of neural networks using the ICSI QuickNet MLP neural network software (Johnson, 2004), with the 39 acoustic features as input, a varying number of phone or feature class posteriors as output, and 1000 hidden nodes.

4 Joint Phone Posteriors vs. Independent Feature Posteriors

The first experiment contrasts joint versus independent feature modeling within the CRF system. We compare a set of phonological feature probabilities derived from the phone posteriors (a joint model) with MLP phone posteriors and with independently trained MLP phonological feature posteriors.

The inputs to the first CRF are sets of 61 state feature functions from the phonemic MLP posteriors, each function is an estimate of the posterior proba-

Input Type.	Phn. Accuracy	Phn. Correct
Phones	67.27	68.77
Features	65.25	66.65
Phn. → Feat.	66.45	67.94

Table 2: Results for Exp. 1: Phone and feature posteriors as input to the CRF phone recognition

bility of one phone. The inputs to the second CRF model are sets of 44 functions corresponding to the phonological features listed in Table 1. The CRF models are trained to associate these feature functions with phoneme labels, incorporating the patterns of variation seen in the MLPs.

The results show that phone-based posteriors produce better phone recognition results than independently-trained phonological features. This could be due in part to the larger number of parameters in the system, but it could also be due to the joint modeling that occurs in the phone classifier.

In order to equalize the feature spaces, we use the output of the phoneme classifier to derive phonological feature posteriors. In each frame we sum the MLP phone posteriors of all phones that contain a given feature. For instance, in the first frame, for the feature LOW, we sum the posterior estimates attributed to the phones *aa*, *ae* and *ao*. This is repeated for each feature in each frame. The CRF model is trained on these data and tested accordingly. The results are significantly better ($p \leq .001$) than the previous features model, but are significantly worse than the phone posteriors ($p \leq .005$).

The results of Experiment 1 confirm the hypothesis of (Rajamanohar and Fosler-Lussier, 2005) that joint modeling using several types of feature information is superior to individual modeling in phone recognition, where only phoneme information is used. The difference between the phone posteriors and individual feature posteriors seems to be related both to the larger CRF parameter space with larger input, and the joint modeling provided by phone posteriors.

5 Phonological Feature Class Analysis

In the second experiment, we examine the influence of each feature class on the accuracy of the recognizer. We iteratively remove the set of state feature functions corresponding to each feature class

Class Removed	Feats.	Phn. Acc.	Phn. Corr.
None	44	65.25	66.65
Sonority	39	65.15	66.58
Voice	41	63.60*	65.03*
Manner	36	58.92*	60.60*
Place	35	53.22*	55.13*
Height	38	62.58*	64.07*
Front	39	64.51*	65.95*
Round	39	65.19	66.64
Tense	41	64.20*	65.65*

* $p \leq .05$, different from no features removed

Table 3: Results of Exp. 2: Removing feature classes from the input

from the input to the CRF. The original functions are the output of the independently-trained feature class MLPs. The phone recognition accuracy for the CRF having removed each class is shown in Table 3. In Table 4 we show how removing each feature class affects the labeling of vowels and consonants.

Manner provides an example of the influence of a single feature class. Both the Accuracy and Correctness scores decrease significantly when features associated with Manner are removed. Manner features distinguish consonants but not vowels, so the effect is concentrated on the recognition of consonants.

The results of Experiment 2 show that certain feature classes are redundant from the point of view of phone recognition. In English, Round is correlated with Front. When we remove Round, the phonemes remain uniquely identified by the other classes. The same is true for the Sonority class. The results show that the inclusion of these redundant features is not detrimental to the recognition accuracy. Accuracy and Correctness improve non-significantly when the redundant features are included.

Clearly, the “independent” phonological feature streams are not truly independent. Otherwise, performance would decrease overall as we removed each feature class, assuming predictiveness.

Removal of Place causes a slight worsening of recognition of vowels. This is surprising, because Place does not characterize vowels. An analysis of the MLP activations showed that the detector for Place=N/A is a stronger indicator for vowels than is the Sonority=Vowel detector. This is especially true for the vowel *ax*, which is frequent in the data,

Class Removed	Percent Correct:	
	Vowels	Consonants
None	62.68	68.91
Sonority	62.18	69.08
Voice	62.39	66.53*
Manner	61.84	59.89*
Place	60.77*	51.94*
Height	55.92*	68.69
Frontness	60.80*	68.87
Roundness	62.25	69.13
Tenseness	60.15*	68.76
* $p \leq .05$, different from no features removed		

Table 4: Effect of removing each feature class on recognition accuracy of vowels and consonants

thus greatly influences the vowel recognition statistic. Removing the Place detectors leads to a loss in vowel vs. consonant information. This results in an increased number of consonant for vowel substitutions (from 560 to 976), thus a decrease in vowel recognition accuracy.

Besides extending the findings in (Rajamanohar and Fosler-Lussier, 2005), this provides a cautionary tale for incorporating redundant phonological feature estimators into ASR: these systems need to be able to handle correlated input, either by design (as in a CRF), through full or semi-tied covariance matrices in HMMs, or by including the appropriate statistical dependencies in DBNs.

6 Summary

We have shown the effect of using joint modeling of phonetic feature information in conjunction with the use of CRFs as a discriminative classifier. Phonetic posteriors, as joint models of phonological features, provide superior phone recognition performance over independently-trained phonological feature models. We also find that redundant features are often modeled well within the CRF framework.

7 Acknowledgments

The authors thank the International Computer Science Institute for providing the neural network software. The authors also thank four anonymous reviewers. This work was supported by NSF ITR grant IIS-0427413; the opinions and conclusions expressed in this work are those of the authors and not of any funding agency.

References

- S. Chang, S. Greenberg, and M. Wester. 2001. An elitist approach to articulatory-acoustic feature classification. In *Interspeech*.
- J. Frankel, M. Wester, and S. King. 2004. Articulatory feature recognition using dynamic bayesian networks. In *ICSLP*.
- J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren. 1993. DARPA TIMIT acoustic-phonetic continuous speech corpus. Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, February. Speech Data published on CD-ROM: NIST Speech Disc 1-1.1, October 1990.
- A. Gunawardana, M. Mahajan, A. Acero, and J. Platt. 2005. Hidden conditional random fields for phone classification. In *Interspeech*.
- M. Hasegawa-Johnson et al. 2005. Landmark-based speech recognition: Report of the 2004 Johns Hopkins Summer Workshop. In *ICASSP*.
- D. Johnson. 2004. ICSI Quicknet software package. <http://www.icsi.berkeley.edu/Speech/qn.html>.
- K. Kirchhoff. 1999. *Robust Speech Recognition Using Articulatory Information*. Ph.D. thesis, University of Bielefeld.
- J. Lafferty, A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning*.
- B. Launay, O. Siohan, A. C. Surendran, and C.-H. Lee. 2002. Towards knowledge based features for large vocabulary automatic speech recognition. In *ICASSP*.
- J. Li and C.-H. Lee. 2005. On designing and evaluating speech event detectors. In *Interspeech*.
- K. Livescu. 2005. *Feature-Based Pronunciation Modeling for Automatic Speech Recognition*. Ph.D. thesis, MIT.
- J. Morris and E. Fosler-Lussier. 2006. Combining phonetic attributes using conditional random fields. In *Interspeech*.
- M. Rajamanohar and E. Fosler-Lussier. 2005. An evaluation of hierarchical articulatory feature detectors. In *IEEE Automatic Speech Recognition and Understanding Workshop*.
- O. Sharenborg, V. Wan, and R.K. Moore. 2006. Capturing fine-phonetic variation in speech through automatic classification of articulatory features. In *ITRW on Speech Recognition and Intrinsic Variation*.