Modeling pronunciation variation with context-dependent articulatory feature decision trees

Sam Bowman[1], *Karen Livescu*[2]

[1]Department of Linguistics, University of Chicago
[2]TTI-Chicago

sbowman@uchicago.edu, klivescu@ttic.edu

Abstract

We consider the problem of predicting the surface pronunciations of a word in conversational speech, using a model of pronunciation variation based on articulatory features. We build context-dependent decision trees for both phone-based and feature-based models, and compare their perplexities on conversational data from the Switchboard Transcription Project. We find that a fully-factored model, with separate decision trees for each articulatory feature, does not perform well, but a feature-based model using a smaller number of "feature bundles" outperforms both the fully-factored model and a phone-based model. The articulatory feature-based decision trees are also much more robust to reductions in training data. We also analyze the usefulness of various context variables.

Index Terms: pronunciation modeling, articulatory features

1. Introduction

Modeling pronunciation variation has long been a challenging issue in automatic speech recognition research. In conversational speech, words often do not adhere to their dictionary pronunciations, resulting in a mismatch with standard baseform dictionaries. This issue has been analyzed in several studies [10, 6], which suggest that extreme variation in conversational speech accounts for a significant amount of the degradation in recognizer performance relative to read speech.

Most approaches to pronunciation modeling are based on predicting a sequence of surface phones from a sequence of underlying phones [13, 5]. This approach often produces an improvement in recognition performance, but not as much as one might hope. A possible reason for this is that pronunciation variation is often not the result of replacement of one entire phone for another, but rather the result of more gradual changes [14]. This has motivated efforts to model this variation on the more fine-grained scale of articulatory features [9, 2, 12].

In this work, we consider an extension of the models in [9]. In particular, we replace their context-independent surface feature distributions with context-dependent ones, modeled using decision trees, and study the predictive power of various types of context dependency by measuring perplexity with respect to a test set of conversational speech.

While this work is motivated by the task of speech recognition, we study the problem of pronunciation modeling in isolation, and note that it is an important problem in its own right (as part of the study of phonology) as well as necessary for other tasks (such as conversational speech synthesis [11]).

2. Feature-based pronunciation modeling

The pronunciation models we consider are based on ideas from autosegmental [3] and articulatory [1] phonology. In such a model, the typical single sequence of phones/phonemes is replaced with multiple sequences of features, which evolve in a semi-independent manner during the course of an utterance. Pronunciation variation, under such a model, is the result of (a) inter-feature *asynchrony*, i.e. the phenomenon of one articulatory feature "getting ahead of" another; and (b) within-feature *substitutions*. For example, feature asynchrony accounts for effects such as vowel nasalization as in "can't" → /k ae_n n t/ and epenthetic stop insertion as in "warmth" → /w ao r m p th/ (in both cases, the velum changes state before the other articulators). Feature substitutions account for, e.g., incomplete stop closures as in "probably" → /p r aa w l iy/.

This type of model can be represented as a dynamic Bayesian network (DBN), as shown in Fig. 1. This figure represents a model with two feature streams, but in general an arbitrary number may be included. Since the features may not be synchronized, at any given time t each feature f has a corresponding underlying phoneme ϕ_t^f that it is currently "aiming for", and a corresponding underlying (target) feature value u_t^f. The degree of asynchrony is controlled by the $async$ variable and state indices i^f (see [7] for more details).

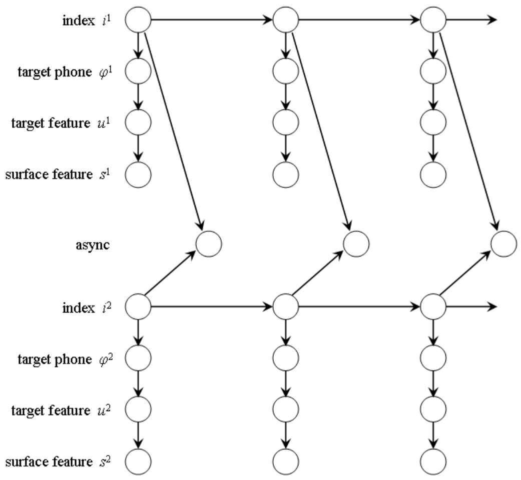

Figure 1: *A dynamic Bayesian network (DBN) corresponding to a feature-based pronunciation model with two features.*

3. Context-dependent feature-based models

Fig. 1 depicts a model with context-independent feature substitutions; that is, each surface feature distribution is conditioned only on its corresponding target feature value. This model was found in [9] to outperform certain phone-based models on a lexical access task. Here we study the possible benefit of modeling the dependence of each surface feature distribution on other variables, such as the feature's previous value (due to inertia) and its next target (anticipatory coarticulation), as well as the values of other features. In other words, we consider adding edges in the DBN from additional context variables to the surface feature variables, and investigate the predictive power of various combinations of such dependencies.

The context-dependent models we consider here are conditional distributions with potentially many conditioning variables, so it is not feasible to enumerate all of the distributions and learn their parameters by counting frequencies in training data. A natural approach for modeling such distributions is decision trees. Similarly to previous work on phone-based pronunciation modeling [13], we learn one decision tree per underlying value per feature, using data that has been aligned. That is, we first align the data, i.e. we find the most likely values of all of the hidden variables, which are potential context variables. Given the alignments, we then learn decision trees predicting each surface feature value conditioned on the chosen set of context variables in the alignment. In order to align the data, we use the context-independent feature model. Again, this is analogous to previous work on phone-based models, in which phones are aligned to phonemes using a simple context-independent aligner, and the more complex decision trees are learned from the aligned data.

It is not meaningful to compare a traditional, segmental phone-based pronunciation model with a frame-based feature model. For purposes of making a direct comparison, we train phone-based decision trees on the frame level as well; that is, we perform a frame-level phonetic alignment and train decision trees for the surface realization of each phoneme in each frame.

4. Evaluation

There are several ways to evaluate pronunciation models: by incorporating them into a speech recognizer, by using them in a lexical access task as in [9], or by measuring how well they predict a test set of pronunciations as in [13]. Here we follow the third approach, and measure the perplexity of the model on a test set of surface pronunciations. This allows us to quickly experiment with many combinations of context variables.

Since the feature-based model is inherently frame-based, we measure the frame-level perplexity. As mentioned above, this makes it difficult to compare with more traditional models that operate on the segment level. We therefore also implement a frame-based phone model using decision trees.

The goal of our evaluation is to measure how well the pronunciation model predicts the surface forms in a test set, given the underlying words w. Let the surface phone label at time frame t be s_t and the corresponding N-feature vector be $\{s_t^1, \ldots, s_t^N\} = s_t^{1:N}$. Our goal, therefore, is to accurately model $p(s_1, \ldots, s_T|w) = p(s_1^{1:N}, \ldots, s_T^{1:N}|w)$, where T is the number of frames in the test set. Since there are many ways to produce a given surface form, corresponding to different alignments (i.e. different settings of the hidden variables such as the underlying phonemes, feature targets, etc.), we need to sum over all of the possible alignments a_i:

$$\begin{aligned} p(s_1, \ldots, s_T|w) &= p(s_1^{1:N}, \ldots, s_T^{1:N}|w) \\ &= \sum_i p(a_i|w) p(s_1^{1:N}, \ldots, s_T^{1:N}|a_i, w) \end{aligned}$$

We make the typical assumptions that the surface form is independent of the word given the alignment, and that one "correct" alignment a is much more likely than all others (i.e. has probability essentially 1). We further assume that each surface feature value s_t^f is independent of all others given some context variables c_t^f, which are a subset of the information in a. Therefore:

$$\begin{aligned} p(s_1, \ldots, s_T|w) &= p(s_1^{1:N}, \ldots, s_T^{1:N}|a, w) \\ &= \prod_{t=1}^{T} \prod_{f=1}^{N} p(s_t^f|c_t^f) \end{aligned}$$

In other words, we can compute the probability of a collection of frames as the product of the feature-specific probabilities in each frame. Note that the perplexity of the feature-based model with respect to a set of test frames is, therefore, simply the product of perplexities of the individual feature models.

In the context-independent model of [9], the context for each surface feature is just the current underlying (target) value for that feature, u_t^f. In general, the context can be any set of variables in the DBN, as long as we don't induce loops. Our goal in this work is to study which context variables are most useful in predicting surface pronunciations.

Under the assumption of a single "correct" alignment, we evaluate our models by first aligning the test data, i.e. finding the settings of all hidden variables in the model given the surface labels and the target word, and then computing the frame-level perplexity of a test set with T labeled frames:

$$perp(s_1, \ldots, s_T) = 2^{\frac{-1}{T} \sum_t \log_2 p(s_t|c_t)}$$

In this expression, $p(s_t|c_t)$ is computed from either the phonetic or the feature-based decision trees, via $p(s_t|c_t) = \prod_{f=1}^{N} p(s_t^f|c_t^f)$. The alignment is done exactly in the same way as for training, i.e. using the context-independent model.

5. Experiments

5.1. Features and data

Our feature-based models use the articulatory features of [7], which are in turn based on the vocal tract variables of Browman and Goldstein's articulatory phonology [1]. We start with seven features, each of which can take on 2-6 discrete values: lip aperture, tongue tip position and aperture, tongue body position and aperture, velum position and glottis aperture. We find that building a separate tree for each of these features does not produce good results: Using the best-performing decision trees, each feature has test set perplexity between 1.1 and 1.5, and the overall test set perplexity is 4.74. In contrast, even a context-independent phone model has perplexity 3.65.

This poor performance is presumably because the assumption of independence between the features is too strong. Instead, we "bundle" the features into three streams that should be more independent: all tongue features (19 possible underlying values, and therefore 19 trees); glottis and velum (3 trees); and lip aperture (4 trees). The context questions used by the trees, however, are still based on the individual features. This factorization of the feature space yields much better results, as we show below.

For the phone-based models, we predict the surface phones given various types of phonetic context, expressed as features, similarly to previous work with phonetic decision trees [13].

We use data from the Switchboard Transcription Project, a subset of the Switchboard database that has been manually labeled at a fine phonetic level, including various diacritics [4]. We drop all diacritics except for nasalization and convert the phone labels to articulatory features. We use a subset containing 90,000 10ms frames of speech, corresponding to about fifteen minutes of conversation from several speakers. We divide the data randomly by frame, retaining 60% for training, 20% for development, and 20% for testing.[1]

5.2. Decision tree learning

All of the models in our experiments are classification trees, built using MATLAB's `classregtree` class. The trees are binary branching, with every non-terminal node representing a decision criterion based on one of the context variables. The decision tree parameters are tuned on the development set. Every leaf node represents a discrete distribution over the possible output values for the tree, based on the frequencies of frames in the training data having the corresponding context.

Fig. 2 shows an example decision tree for the lip aperture feature. This tree has been significantly pruned for ease of viewing. This tree is for the case of lip aperture = 2, corresponding to a medium-open lip opening. According to this tree, if the previous state of the lip aperture was less than medium-open, then the surface value of lip aperture will nearly certainly remain medium-open. Otherwise (i.e., if the previous lip aperture was wide), the surface lip aperture may be either medium or wide, depending on the location and aperture of the tongue.

The decision tree learning often produces distributions that assign zero probability to some output values. In order to avoid assigning zero probability to any test label s, we apply a simple Lidstone's law smoothing at each node n in the tree:

$$p_{\text{smooth}}(s|n) = \frac{count(s \text{ at node } n) + \lambda}{count(s \text{ at any node}) + \lambda M},$$

where M is the number of nodes in the tree. The value of λ is tuned on the development set.

5.3. Context variables

The context variables we use are the current, previous and next *distinct* values of several variables. Table 1 provides a simplified example with only three features and seven frames. The table shows the underlying (target) values of the three features in each frame. At frame 5, the values of (distinct previous, current, distinct next) are as follows:

- Tongue tip position: (1, 0, null)
- Glottis aperture: (null, 1, null)
- Lip aperture: (1, 0, 2)

In order to account for the fact that coarticulation affects nearby frames more than distant ones, we also consider a "distance" context variable, which is the number of frames to the nearest target value transition on either side. In this example, the distances to the previous/next distinct targets are:

- Tongue tip position: (-0, null)
- Glottis aperture: (null, null)
- Lip aperture: (-3, +2)

Frame	1	2	3	4	5	6	7
Tongue tip position	0	0	1	1	0	0	0
Glottis aperture	1	1	1	1	1	1	1
Lip aperture	1	1	0	0	0	0	2

Table 1: *Example values for three features.*

	Phone-based	Feature-based
1. *Context-independent:*		
	3.6483	2.5672
2. *Basic context-dependent:*		
	2.5189	2.1511
3. *Context-dep. + previous surface value*		
	1.6449	1.7861
4. *Context-dep. + prev. surf. + distance:*		
	2.0784	1.6850
5. *Cross-feature context-dep. + prev. surf.:*		
	N/A	1.5664
6. *Cross-feat. context-dep. + prev. surf. + phone context:*		
	N/A	1.5248
7. *Cross-feat. context-dep. + prev. surf. + cross-phone context:*		
	N/A	1.5237

Table 2: *Test set perplexities for various models.*

We tested decision tree models conditioned on the following sets of context variables. Except where noted, we used the analogous context variables to test both phone-based and feature-based models.

1. *Context-independent*: The context is only the target value for the phone/feature at the current frame.
2. *Basic context-dependent*: The target value at the current frame, and the next/previous distinct target value.
3. *Context-dependent + previous surface value*: Same as above, plus the last distinct observed value of the phone/feature.
4. *Context-dependent + previous surface value + distance*: Same as above, plus the distance in frames to the previous/next distinct target value.

   We also tested the following versions of the feature-based model, which use context information not applicable to a phone-based model:

5. *Cross-feature context-dependent + previous surface value*: Same as (3), but with target and actual values for all features, not just the one being predicted.
6. *Cross-feature context-dependent + previous surface value + phone context*: Same as (5), plus the target phones corresponding to the current, previous, and next targets of the current feature.
7. *Cross-feature context-dependent + previous surface value + cross-phone context*: Same as (6), plus the current, previous, and next target phones corresponding to all other features.

In addition, we also tested several models with varying amounts of training data. The results are shown in Figure 3.

[1] The random division causes some labels to be divided between the test and training sets, which may lead to some over-fitting; however, this should only bias the results against our feature-based models by giving an advantage to the model with more parameters, i.e. the phone model.

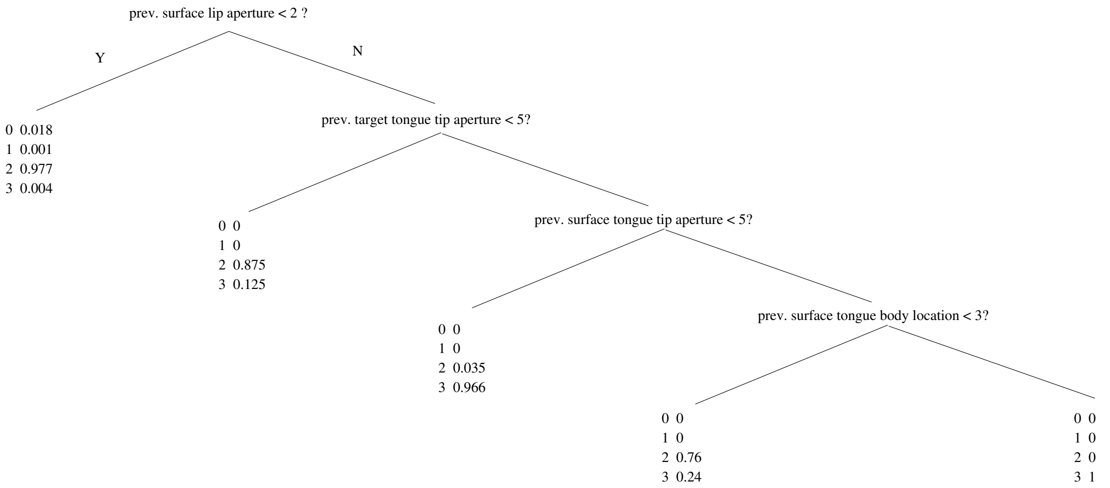
prev. surface lip aperture < 2 ?
Y
N
0 0.018
1 0.001
2 0.977
3 0.004
prev. target tongue tip aperture < 5?
0 0
1 0
2 0.875
3 0.125
prev. surface tongue tip aperture < 5?
0 0
1 0
2 0.035
3 0.966
prev. surface tongue body location < 3?
0 0
1 0
2 0.76
3 0.24
0 0
1 0
2 0
3 1

Figure 2: *A sample decision tree, predicting the surface lip aperture given that the target lip aperture is 2 (medium-open). The "yes" branch of each question node is always the left branch.*

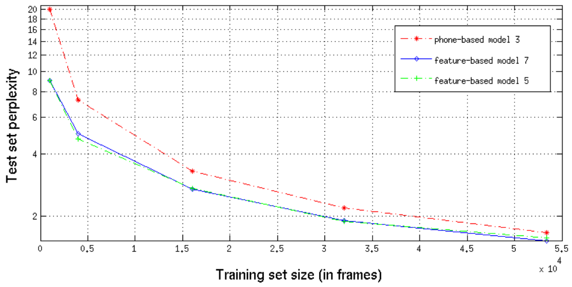

Figure 3: *Test set perplexity vs. training set size.*

6. Discussion and conclusions

We have found that context-dependent articulatory feature decision trees outperform context-independent trees, as well as frame-level phone-based trees, for predicting manual phonetic labels in the Switchboard Transcription Project data. In addition, the best feature-based model (model 7) degrades much more slowly than the best phone-based model (model 3) as the amount of training data is reduced. In fact, a simpler feature-based model (model 5), which is slightly worse than the best when using the full training set, performs slightly better when training data is reduced.

We note that, in these experiments, the full power of the feature-based model is not exploited, because the surface pronunciations are transcribed phonetically. For example, it may be beneficial to model the fact that nasalized vowels are typically nasalized only for a portion of the vowel. We have also ignored some effects such as incomplete stop closures and anticipatory rounding. Such effects could be handled in a feature-based acoustic model, but are missing from the phonetic transcription. We are currently collecting finer-grained feature-level transcriptions of Switchboard utterances (an extension of the transcription effort described in [8]), so that we may carry out more detailed experiments in the future. Such data collections are extremely laborious, so an additional direction for future research is to explore the tradeoff between amount and quality/level of detail of the training labels.

Our next steps are to test the most successful models from these experiments via lexical access experiments as in [9] and end-to-end recognition experiments.

7. References

[1] Browman, C. P. and Goldstein, L, "Articulatory phonology: An overview", *Phonetica*, 49:155–180, 1992.

[2] Deng, L., et al., "Production models as a structural basis for automatic speech recognition." *Speech Communication* 33:93–111, 1997.

[3] Goldsmith, J. A. *Autosegmental and metrical phonology*. Basil Blackwell, Oxford, UK, 1990.

[4] Greenberg, S., et al., "Insights into spoken language gleaned from phonetic transcription of the switchboard corpus", In *Proc. ICSLP*, Philadelphia, October 1996.

[5] Hazen, T. J., et al., "Pronunciation modeling using a finite-state transducer representation." *Speech Communication* 46(2):189–203, June 2005.

[6] Jurafsky, D., et al., "What kind of pronunciation variation is hard for triphones to model?" In *Proc. ICASSP*, 2001.

[7] Livescu, K., *Feature-Based Pronunciation Modeling for Automatic Speech Recognition*, Ph.D. dissertation, MIT, 2005.

[8] Livescu, K., et al., "Manual transcription of conversational speech at the articulatory feature level." In *Proc. ICASSP*, 2007.

[9] Livescu, K., and Glass, J., "Feature-based pronunciation modeling with trainable asynchrony probabilities." In *Proc. ICSLP*, 2004.

[10] McAllaster, D., et al., "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch." In *Proc. ICSLP*, 1998.

[11] Prahallad, K., et al., "Sub-phonetic modeling for capturing pronunciation variations for conversational speech synthesis." In *Proc. ICASSP*, 2006.

[12] Richardson, M., et al., "Hidden-articulator Markov models for speech recognition." In *Proc. ITRW ASR*, 2000.

[13] Riley, M., et al., "Stochastic pronunciation modelling from hand-labelled phonetic corpora", *Speech Communication*, 29(2–4):209–224, November 1999.

[14] Saraçlar, M., *Pronunciation Modeling for Conversational Speech Recognition*. Ph.D. dissertation, Johns Hopkins U., 2000.