Feature-based Pronunciation Modeling for Speech Recognition

Karen Livescu and James Glass
MIT Computer Science and Artificial Intelligence Laboratory
Cambridge, MA 02139, USA
{klivescu, glass}@csail.mit.edu

Abstract

We present an approach to pronunciation modeling in which the evolution of multiple linguistic feature streams is explicitly represented. This differs from phone-based models in that pronunciation variation is viewed as the result of feature asynchrony and changes in feature values, rather than phone substitutions, insertions, and deletions. We have implemented a flexible feature-based pronunciation model using dynamic Bayesian networks. In this paper, we describe our approach and report on a pilot experiment using phonetic transcriptions of utterances from the Switchboard corpus. The experimental results, as well as the model's qualitative behavior, suggest that this is a promising way of accounting for the types of pronunciation variation often seen in spontaneous speech.

1 Introduction

Pronunciation variation in spontaneous speech has been cited as a serious obstacle for automatic speech recognition (McAllester et al., 1998). Typical pronunciation models approach this problem by augmenting a phonemic dictionary with additional pronunciations, often resulting from the application of phone substitution, insertion, and deletion rules. By carefully constructing a rule set (Hazen et al., 2002), or by deriving rules or variants from data (Riley and Ljolje, 1996), many phenomena can be accounted for. However, the recognition improvement over a phonemic dictionary is typically modest, and some types of variation remain awkward to represent.

These observations have motivated approaches to speech recognition based on multiple streams of linguistic features rather than a single stream of phones (e.g., King et al. (1998); Metze and Waibel (2002); Livescu et al. (2003)). Most of this work, however, has focused on acoustic modeling, i.e. the mapping between the features and acoustic observations. The pronunciation model is typically still phone-based, limiting the feature values to the target configurations of phones and forcing them to behave as a synchronous "bundle". Some approaches have begun to relax these constraints. For example, Deng et al. (1997) and Richardson et al. (2000) model asynchronous feature trajectories using hidden Markov models (HMMs), with each state corresponding to a vector of feature values. This approach is powerful, but it cannot represent independencies between features. Kirchhoff (1996), in contrast, models the feature streams as independent, except for a requirement that they synchronize at syllable boundaries. As pointed out by Ostendorf (2000), such independence assumptions may allow for too much variability.

In this paper, we propose a more general feature-based pronunciation model implemented using dynamic Bayesian networks (Dean and Kanazawa, 1989), which allow us to take advantage of inter-feature independencies while avoiding overly strong independence assumptions. In the following sections, we describe the model and present proof-of-concept experiments using phonetic transcriptions of utterances from the Switchboard conversational speech corpus (Greenberg et al., 1996).

2 Serval [*sic*] examples

To help ground the discussion, we first present several examples of pronunciation variation. One common phenomenon is the nasalization of vowels preceding nasal consonants. This is a result of asynchrony: The velum is lowered before the oral closure is made. In more extreme cases, the nasal consonant is entirely absent, leaving only a nasalized vowel, as in *can't* → [k ae_n t] [1]. All of the feature values are still correct, although phonetically, this would be described as a deletion.

Another example, taken from the Switchboard corpus, is *several* → [s eh r v ax l]. In this case, the tongue and lips have desynchronized to the point that the tongue

[1] Here and throughout, we use the ARPAbet phonetic symbol set with additional diacritics, such as "_n" for nasalization.

retroflexion for [r] starts and ends before the lip narrowing gesture for [v]. Again, all of the feature streams are produced correctly, but there is an apparent exchange of two phones, which cannot be represented via single-phone confusions conditioned on phonemic context.

A final example from Switchboard is *everybody* → [eh r uw ay]. It is difficult to imagine a set of phonetic transformations that would predict this pronunciation without allowing a host of other impossible pronunciations. However, when viewed in terms of features, the transformation from [eh v r iy bcl b ah dx iy] to [eh r uw ay] is fairly simple. The tongue and lips desynchronize, causing the lips to start to close for the [bcl] during the previous vowel. In addition, the lip constrictions for [bcl] and [v], and the tongue tip gesture for [dx], are reduced. We will return to this example in the sections below.

3 Approach

A feature-based pronunciation model is one that explicitly models the evolution of multiple underlying linguistic feature streams to predict the allowed realizations of a word and their probabilities. Our approach begins with the usual assumption that each word has one or more target phonemic pronunciations, or baseforms. Each baseform is converted to a table of *underlying* feature values. Table 1 shows what part of this table might look like for the word *everybody*. The table may include "unspecified" values ('*' in the table). More generally, each table entry can be a distribution over the range of feature values. For now, we assume that all of the features go through the same sequence of indices (and therefore the same number of targets) in a given word; e.g., in Table 1, **LIP-OPEN** goes through the same indices as **TT-LOC**, although it has the same target value for indices 2 and 3. In the first time frame of speech, all of the features begin in index 0; in subsequent frames, each feature can either stay in the same index or transition to the next one with some probability.

The *surface* feature values—i.e., the ones that are actually produced by the speaker—can stray from the underlying pronunciation in two ways, typically because of articulatory inertia: *substitution*, in which a feature fails to reach its target underlying value; and *asynchrony*, in which different features proceed through their sequences of indices at different rates. We define the degree of asynchrony between two sets of features as the difference between the average index of one set relative to the average index of the second. The degree of asynchrony is constrained: More "synchronous" configurations are more probable (soft constraints), and we make the further simplifying assumption that there is an upper bound on the degree of asynchrony (hard constraints).

A natural framework for such a model is provided by dynamic Bayesian networks (DBNs), because of their ability to efficiently implement factored state representations. Figure 1 shows one frame of the type of DBN used in our model (simplified somewhat for clarity of presentation). This example DBN assumes a feature set with three features. The variables at time frame t are as follows:

index	0	1	2	3	...
phoneme	eh	v	r	iy	...
LIP-OPEN	wide	critical	wide	wide	...
TT-LOC	alv.	*	ret.	alv.	...
...	...	...	...	...	...

Table 1: *Part of a target pronunciation for* everybody. *In this feature set,* **LIP-OPEN** *is the lip opening degree;* **TT-LOC** *is the location along the palate to which the tongue tip is closest (*alv. = *alveolar;* ret. = *retroflex).*

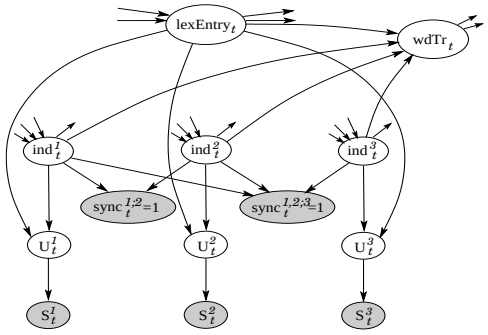

Figure 1: *One frame of a DBN for recognition with a feature-based pronunciation model. Nodes represent variables; shaded nodes are observed. Edges represent dependencies between variables. Edges without parents/children point from/to variables in adjacent frames (see text).*

lexEntry$_t$ – entry in the lexicon corresponding to the current word and baseform. Words with multiple baseforms have one entry per baseform. *lexEntry*$_t$'s parents are *lexEntry*$_{t-1}$ and *wdTr*$_{t-1}$

ind^j_t – index of feature j into the underlying pronunciation, as in Table 1. *ind*$^j_0$ = 0; in subsequent frames *ind*^j_t is conditioned on *lexEntry*$_{t-1}$, *ind*$^j_{t-1}$, and *wdTr*$_{t-1}$ (defined below).

U^j_t – underlying value of feature j. Its distribution $p(U^j_t|lexEntry_t, ind^j_t)$ is determined by the target feature table of *lexEntry*$_t$.

S^j_t – observed surface value of feature j. $p(S^j_t|U^j_t)$ encodes allowed feature substitutions.

wdTr$_t$ – binary variable indicating whether this is the last frame of the current word.

sync$^{A;B}_t$ – binary variable that enforces a synchrony constraint between subsets A and B of the feature set. It is observed with value 1; its distribution is con-

structed in such a way as to force its parent *ind* variables to obey the desired constraint. For example, to enforce a constraint between the average index of features 1 and 2 and the index of feature 3, we would have $P(sync_t^{1,2;3} = 1|ind_t^1, ind_t^2, ind_t^3) = 0$ whenever ind_t^1, ind_t^2, ind_t^3 violate the constraint.

In an end-to-end recognizer, the acoustic observations would depend on the S_t^j, which would be unobserved. However, to facilitate quick experimentation and isolate the pronunciation model, we begin by testing how well we can do when given observed surface feature values.

4 Experiments

We have performed a pilot experiment using the following feature set, based on the vocal tract variables of articulatory phonology (Browman and Goldstein, 1992): degree of lip opening; tongue tip location and opening degree; tongue body location and opening degree; velum state; and glottal (voicing) state. We imposed the following synchrony constraints: (1) All four tongue features are completely synchronized; (2) the lips can desynchronize from the tongue by up to one index; and (3) the glottis and velum are synchronized, and their index must be within 2 of the mean index of the tongue and lips.

We used the Graphical Models Toolkit (Bilmes and Zweig, 2002) to implement the model. The distributions $p(S_t^j|U_t^j)$ were constructed by hand based on linguistic considerations, e.g. that features tend to go from more "constricted" values to less constricted ones, but not vice versa. $p(U_t^j|lexEntry_t, ind_t^j)$ was derived from manually-constructed phoneme-to-feature-probability mappings. For these experiments, no parameter learning has been done.

The task was to recognize an isolated word, given a set of observed surface feature sequences S_t^j. To create the observations, we used the detailed phonetic transcriptions created at ICSI for the Switchboard corpus (Greenberg et al., 1996). For each word, we converted its transcription to a sequence of feature vectors, one vector per 10 ms frame. For this purpose, we divided diphthongs and stops into pairs of feature configurations. Given the input feature sequences, we computed a Viterbi score for each lexical entry in a 3000+-word (5500+-lexEntry) vocabulary, by "observing" the $lexEntry$ variable and finding the most likely settings of all remaining variables. The most likely variable settings can be thought of as a multistream alignment between the surface and underlying feature streams. Finally, we output the word corresponding to the highest-scoring lexical entry.

We performed this procedure on a development set of 165 word transcriptions, which was used to tune settings such as synchronization constraints, and a test set of 236 transcriptions [2]. We compared the performance of several models, measured in terms of word error rate (WER) and failure rate (FR), the percentage of inputs that had no Viterbi alignment with the correct word. To get a sense of the effect of feature asynchrony, we compared our asynchronous model with a version in which all features are forced to be synchronized, so that only feature substitution is allowed. This uses the same DBN, but with degenerate distributions for the synchronization variables. Also, since the S^j values are derived from phonetic transcriptions, and are therefore constant over several frames at a time, we also built a variant of the DBN in which S^j is allowed to change value with non-zero probability only when ind^j changes (by adding parents ind_t^j, ind_{t-1}^j, S_{t-1}^j to S_t^j); we refer to this DBN as "segment-based", and to the original as "frame-based". We compared four variants, differing along the "synchronous vs. asynchronous" and "frame-based vs. segment-based" dimensions. The variant which is both synchronous and segment-based is similar to a phone-based pronunciation model with only context-independent phone substitutions.

	dev set		test set	
model	**WER**	**FR**	**WER**	**FR**
baseforms only	63.6	61.2	69.5	66.9
phonological rules	50.3	47.9	59.7	55.5
sync. seg.-based	38.2	24.8	43.2	35.2
sync. fr.-based	35.2	23.0	46.2	31.4
async. seg.-based	32.7	19.4	41.1	31.4
async. fr.-based	29.7	16.4	42.7	26.3

Table 2: *Results of Switchboard ranking experiment.*

Table 2 shows the performance of these four models, as well as of two "baseline" models: one allowing only the baseform pronunciations (on average 1.7 per word), and another including all pronunciations produced by an extensive set of context-dependent phonological rules (about 4 per word), with no feature substitutions or asynchrony in either case. The phonological rules are the "full rule set" described in Hazen et al. (2002). We note that they were not designed with Switchboard in mind.

The models that allow asynchrony outperform the ones that do not, in terms of both WER and FR. Looking more closely at the performance on the development set, the inputs on which the synchronous models failed but the asynchronous models succeeded were in fact the kinds of pronunciations that we expect to arise from feature asynchrony, including: nasals replaced by nasalization on a preceding vowel; a /t r/ sequence realized as /ch/; and *everybody* → [eh r uw ay]. The relative merits of the frame-based and segment-based models is less clear, as

[2] We required that words in the development and test sets have phonemic pronunciations with at least 4 phonemes, so as to limit context effects from adjacent words.

they have opposite relative performance on the development and test sets. For 27 (16.4%) development utterances, none of the models was able to find an alignment with the correct word. Most of these were due to apparent gesture deletions and context-dependent feature changes, which are not yet included in the model.

Figure 2 shows a part of the Viterbi alignment of *everybody* with [eh r uw ay], produced by the segment-based, asynchronous model. Using this model, *everybody* was the top-ranked word. As expected, the asynchrony is manifested in the [uw] region, and the lips do not close but reach only a narrow (glide-like) configuration.

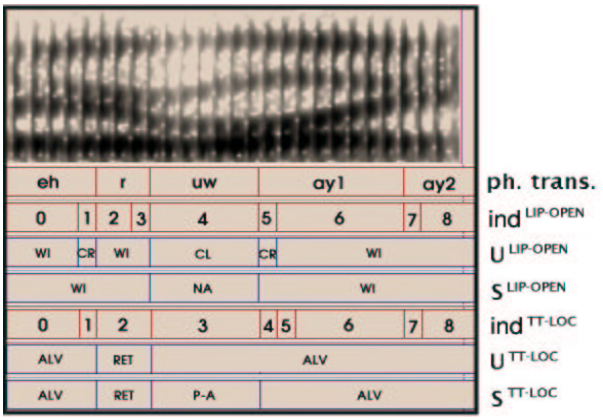

Figure 2: *Spectrogram, transcription, and partial Viterbi alignment, including the lip opening and tongue tip location variables. Indices are relative to the underlying pronunciation* /eh v r iy bcl b ah dx iy/. *Adjacent frames with equal values have been merged for easier viewing.* **WI** = wide*;* **NA** = narrow*;* **CR** = critical*;* **CL** = closed*;* **ALV** = alveolar*;* **P-A** = palato-alveolar*;* **RET** = retroflex.

5 Discussion

We have motivated our pronunciation model as part of an overall strategy of feature-based speech recognition. One way in which this model could fit into a complete recognizer is, as mentioned above, by adding a variable A representing the acoustic observations, with the S^j as its parents. The modeling of $p(A|S^1, \ldots, S^M)$ (where M is the number of features) is a significant problem in its own right. Alternatively, as this study suggests, there may be some benefit to this type of model even if the acoustic model is phone-based. One possible setup would be to use a phonetic recognizer to produce a phone lattice, then convert the phones into features and proceed as in our Switchboard experiments.

Thus far we have not trained the variable distributions. With the exception of the $sync$ variables, these can be trained from feature transcriptions (i.e. S^j observations) using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977). In the absence of actual feature transcriptions, they can be approximated by converting detailed phonetic transcriptions, as we have done in our decoding experiments above. The $sync$ distributions cannot be trained via EM, since they are always observed with value 1. They can either be treated as experimental parameters or trained discriminatively. We are currently working on a new formulation in which the synchronization constraints can be trained via EM.

In addition, we are currently investigating extensions to the model, including context-dependent feature substitutions. We also plan to extend this study to a larger data set and to multi-word utterances.

References

J. Bilmes and G. Zweig, "The Graphical Models Toolkit: An open source software system for speech and time-series processing," *ICASSP*, Orlando, 2002.

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

T. Dean and K. Kanazawa, "A model for reasoning about persistence and causation," *Computational Intelligence*, **5**:142–150, 1989.

A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society*, **39**:1–38,1977.

L. Deng, G. Ramsay, and D. Sun, "Production models as a structural basis for automatic speech recognition," *Speech Communication*, **33**:93–111, 1997.

S. Greenberg, J. Hollenback, and D. Ellis, "Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus," *ICSLP*, Philadelphia, 1996.

T. J. Hazen, I. L. Hetherington, H. Shu, and K. Livescu, "Pronunciation modeling using a finite-state transducer representation," *ITRW PMLA*, Estes Park, CO, 2002.

S. King, T. Stephenson, S. Isard, P. Taylor, and A. Strachan, "Speech recognition via phonetically featured syllables," *ICSLP*, Sydney, 1998.

K. Kirchhoff, "Syllable-level desynchronisation of phonetic features for speech recognition," *ICSLP*, Philadelphia, 1996.

K. Livescu, J. Glass, and J. Bilmes, "Hidden feature models for speech recognition using dynamic Bayesian networks," *Eurospeech*, Geneva, 2003.

D. McAllester, L. Gillick, F. Scattone, and M. Newman, "Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch," *ICSLP*, Sydney, 1998.

F. Metze and A. Waibel, "A flexible stream architecture for ASR using articulatory features," *ICSLP*, Denver, 2002.

M. Ostendorf, "Incorporating linguistic theories of pronunciation variation into speech-recognition models," *Phil. Trans. R. Soc. Lond. A*, **358**:1325-1338, 2000.

M. Richardson, J. Bilmes, and C. Diorio, "Hidden-articulator Markov models for speech recognition," *ITRW ASR2000*, Paris, 2000.

M. D. Riley and A. Ljolje, "Automatic generation of detailed pronunciation lexicons," in C.-H. Lee, F. K. Soong, and K. K. Paliwal (eds.), *Automatic Speech and Speaker Recognition*, Kluwer Academic Publishers, Boston, 1996.