Articulatory Feature-Based Pronunciation Modeling

Karen Livescu[1a], Preethi Jyothi[b], Eric Fosler-Lussier[c]

[a]*TTI-Chicago, Chicago, IL, USA*
[b]*Beckman Institute, UIUC, Champaign, IL, USA*
[c]*Department of Computer Science and Engineering, OSU, Columbus, OH, USA*

Abstract

Spoken language, especially conversational speech, is characterized by great variability in word pronunciation, including many variants that differ grossly from dictionary prototypes. This is one factor in the poor performance of automatic speech recognizers on conversational speech, and it has been very difficult to mitigate in traditional phone-based approaches to speech recognition. An alternative approach, which has been studied by ourselves and others, is one based on sub-phonetic features rather than phones. In such an approach, a word's pronunciation is represented as multiple streams of phonological features rather than a single stream of phones. Features may correspond to the positions of the speech articulators, such as the lips and tongue, or may be more abstract categories such as manner and place.

This article reviews our work on a particular type of articulatory feature-based pronunciation model. The model allows for asynchrony between features, as well as per-feature substitutions, making it more natural to account for many pronunciation changes that are difficult to handle with phone-based models. Such models can be efficiently represented as dynamic Bayesian networks. The feature-based models improve significantly over phone-based counterparts in terms of frame perplexity and lexical access accuracy. The remainder of the article discusses related work and future directions.

Keywords: speech recognition, articulatory features, pronunciation modeling, dynamic Bayesian networks

1. Introduction

Human speech is characterized by enormous variability in pronunciation. Two speakers may use different variants of the same word, such as *EE-ther* vs. *EYE-ther*, or they may have different dialectal or non-native accents. There are also speaker-independent causes, such as speaking style–the same words may be pronounced carefully and clearly when reading but more sloppily in conversational or fast speech [44] (e.g. "probably" may be pronounced "probly" or even "prawly"). In this article we are concerned with building a pronunciation model that is a distribution over the possible sub-word sequences that may be produced in uttering a given word; and we focus on building a model that is as accurate as possible for conversational speech. Here we address speaker-independent pronunciation variability, i.e. variability due to speaking style or context, although the methods we describe are applicable to studying dialectal or idiolectal variation as well. There are many possible applications for this work, including in automatic speech recognition (ASR), linguistics, and psycholinguistics. In this work, we are mainly motivated by the ASR application, where pronunciation variation in conversational speech is a significant problem.

Pronunciation variation has long been considered a major factor in the poor performance of ASR systems on conversational speech [72, 63, 94, 25, 91, 59]. Early work on this topic analyzed this effect in various ways. For example, Weintraub *et al.* [94] compared the error rates of a recognizer on identical word sequences recorded in identical conditions but with different styles of speech, and found the error rate to be almost twice higher for spontaneous conversational sentences than for the same sentences read by the same speakers in a dictation style. Fosler-Lussier [25] found that words pronounced non-canonically are more likely than canonical productions to be deleted or substituted by an automatic speech recognizer. McAllaster and Gillick [63] generated synthetic speech with pronunciations

[1]Corresponding author. Tel: +1-773-834-2549; fax: +1-773-834-9881; email: klivescu@ttic.edu

matching canonical dictionary forms, and found that it can be recognized with error rates about eight times lower than for synthetic speech with the pronunciations observed in actual conversational data.

Considering recent advances in speech recognition, one may wonder whether this is still a challenge. Indeed it is: Although error rates in general have gone down dramatically, they are still 50% higher for non-canonically pronounced words in a recent discriminatively trained recognizer [59]. Most speech recognition systems use context-dependent (such as triphone) acoustic models to implicitly capture some of the pronunciation variations. However, this approach may be insufficient for modeling pronunciation effects that involve more than a single phone and its immediate neighbors, such as the rounding of [s] in *strawberry*. The work of Jurafsky *et al.* [45] suggests that triphones are in general mostly adequate for modeling phone substitutions, but inadequate for handling insertions and deletions.

There have been a number of approaches proposed for handling this variability in the context of phone-based speech recognition. One approach, which was studied heavily especially in the 1990s but also more recently, is to start with a dictionary containing canonical pronunciations and add to it those alternative pronunciations that occur often in some database, or that are generated by deterministic or probabilistic phonetic substitution, insertion, and deletion rules (e.g., [88, 80, 95, 91, 25, 83, 38]). Other approaches are based on alternative models of transformations between the canonical and observed pronunciations, such as phonetic edit distance models [42] and log-linear models with features based on canonical-observed phone string combinations [103]. Efforts to use such ideas in ASR systems have produced performance gains, but not of sufficient magnitude to solve the pronunciation variation problem.

One often-cited problem is that with the introduction of additional pronunciations, confusability between words is also introduced [23, 80]. This may be due to the large granularity of phone-level descriptions: An actual pronunciation may contain a sound that is neither a dictionary phone nor an entirely different phone, but rather something intermediate [83], suggesting that a finer-grained level of representation may be needed. One way of introducing a finer-grained representation is to represent words in terms of multiple streams of sub-phonetic features, rather than a single stream of phones. This idea is supported by modern phonological theories such as autosegmental phonology [29] and articulatory phonology [14]. Many authors have pursued such approaches in one form or another [19, 22, 52, 79, 66, 60]. Early approaches used hidden Markov models (HMMs) in which each state corresponds to a combination of articulatory feature values [19, 22, 79]. The HMM state space is constructed by allowing features to evolve asynchronously between phonetic targets, resulting in a very large state space. To control the size of the space, this work involved constraining the models, for example by requiring that the features re-synchronize often (e.g. at phonetic targets).

In the work we review here, we have proposed explicit models of multiple streams of articulatory features, which factor the large state space of articulatory configurations into models of inter-articulator asynchrony and articulatory substitution (e.g., reduction). The factorization of the joint articulatory feature distribution can be represented via graphical models (in particular dynamic Bayesian networks) and requires fewer model parameters than if we "compiled" the model into an HMM. Incorporating such models into complete recognizers is an important research question that we do not address in this paper; instead, here we focus on evaluation of pronunciation models independently of any particular recognizer. In prior work on new phone-based pronunciation models, the models were evaluated through experiments with manual phonetic transcriptions. Similarly, here we describe experiments using manual phonetic transcriptions converted to articulatory features. Section 4 provides more details on the evaluation methods.

In the next section, we provide detailed motivating background and examples. The remainder of the paper presents the proposed model (Section 3), along with a feature set based on articulatory phonology and the model representation in terms of dynamic Bayesian networks; summarizes a series of evaluations using perplexity and lexical access accuracy as performance measures (Section 4); and gives a brief description of model variants and extensions of the basic model that have been explored by ourselves and others (Section 5).

2. Motivation

We motivate our approach by considering in greater detail some prior work and relevant examples. We define a *pronunciation* of a word as a representation of the way the word can be produced by a speaker, in terms of some set of linguistically meaningful sub-word units. We distinguish between a word's (i) *underlying* (or *target* or *canonical*) pronunciations, the ones typically found in a dictionary and represented as strings of phonemes, and its (ii) *surface*

pronunciations, the ways in which a speaker may actually produce the word. While a given word usually has one or a few underlying phonetic pronunciations, the same word may have dozens of surface phonetic pronunciations. Table 1 shows canonical pronunciations for four words, along with all of their surface pronunciations that appear in the portion of the Switchboard conversational database that has been manually transcribed phonetically [34]. The data set is described further in Section 4.3. The surface pronunciations given here are somewhat simplified from the original transcriptions for ease of reading; e.g., [dx] has been transcribed as [d] and [nx] as [n], and vowel nasalization is not shown. The exact transcriptions are to some extent subjective.[2] However, there are a few points about the data in Table 1 that are clear:

- There is a large number of surface pronunciations per word, with most occurring only once in the data.
- The canonical pronunciation rarely appears in the transcriptions: It was not used at all in the two instances of *sense*, eleven of *probably*, and five of *everybody*, and used four times out of 89 instances of *don't*.
- Many observed pronunciations differ grossly from the canonical one, with entire phones or syllables deleted (as in *probably* → [p r ay] and *everybody* → [eh b ah iy]) or inserted (as in *sense* → [s eh n t s]).
- Many observed pronunciations are the same as those of other English words. For example, according to this table, *sense* can sound like *cents* and *sits*; *probably* like *pry*; and *don't* like *doe, own, oh, done, a, new, tow*, and *dote*. In other words, it would seem that all of these word sets should be confusable.

These four words are not outliers: For words spoken at least five times in this database, the mean number of distinct pronunciations is 8.8.[3]

word	*sense*	*probably*	*everybody*	*don't*
canonical	s eh n s	p r aa b ax b l iy	eh v r iy b aa d iy	d ow n t
surface	s eh n t s (1) s ih t s (1)	p r aa b iy (2); p r ay (1); p r aw l uh (1); p r ah b iy (1); p r aa l iy (1); p r aa b uw (1); p aa b uh b l iy (1); p aa ah iy (1)	eh v r ax b ax d iy (1) eh v er b ah d iy (1) eh ux b ax iy (1) eh r uw ay (1) eh b ah iy (1)	d ow n (37); d ow (16); ow n (6); d ow n t (4); d ow t (3); d ah n (3); ow (3); n ax (3); d ax n (2); ax (2); n uw (1); n (1); t ow (1); d ow ax n (1); d el (1); d ao (1); d ah (1); dh ow n (1); d uh n (1); ax ng (1)

Table 1: Canonical and observed surface pronunciations of four words in the phonetically transcribed portion of the Switchboard database [34].[4] The number of times each observed pronunciation appears in the database is given in parentheses.

2.1. Phone-based pronunciation modeling in ASR

One approach used in ASR research for handling this variability is to start with a dictionary containing only canonical pronunciations and add to it those alternate pronunciations that occur often in some database [91]. The alternate pronunciations can be weighted according to the frequencies with which they occur in the data. By limiting the number of pronunciations per word, we can ensure that we have sufficient data to estimate the probabilities, and we can (to some extent) control the degree of confusability between words. However, this does not address the problem of the many remaining pronunciations that do not occur with sufficient frequency to be counted. Perhaps more importantly, for any reasonable-sized vocabulary and database, most words in the vocabulary will only occur a handful of times, and many will not occur at all. For example, of the 29,695-word vocabulary of the Switchboard database, 18,504 words occur fewer than five times. It is therefore infeasible to robustly estimate the probabilities of most words' pronunciations.

[2] As noted by Johnson, "Linguists have tended to assume that transcription disagreements indicate ideolectal differences among speakers, or the moral degeneracy of the other linguist." [43]

[3] This is after dropping some diacritics and collapsing similar phone labels, reducing the phonetic label set from 396 to 179 distinct labels. Before collapsing the labels, the mean number of distinct pronunciations for words spoken at least five times is 11.0.

[4] Here and throughout, we use the ARPABET phonetic alphabet [85], with additional diacritics as needed.

However, many pronunciation variants are predictable. For example, we have seen that *sense* can be pronounced [s eh n t s]. In fact, there are many words that show a similar pattern of epenthetic stop insertion: *defense* → [d ih f eh n t s], *prince* → [p r ih n t s], *insight* → [ih n t s ay t], and so on. These can be generated by a phonetic rewrite rule, $\varepsilon \rightarrow$ t / n _ s, read "The empty string (ε) can become *t* in the context of an *n* on the left and *s* on the right." Many pronunciation phenomena are described well by rules of the form $p_1 \rightarrow p_2 \,/\, c_l \_ c_r$, where p_1, p_2, c_l, and c_r are phonetic labels or strings. Such rules have been documented in the linguistics, speech science, and speech technology literature (e.g., [39, 84, 54, 50, 71]) and are the basis for another phone-based pronunciation modeling approach: One or a few main pronunciations are listed for each word, and a bank of rewrite rules is used to generate additional pronunciations. The rules can be pre-specified [38] or learned from data [23], and their firing probabilities can also be learned [86]. A related approach is to learn, for each phoneme, a decision tree representing a detailed set of rules for predicting the phoneme's surface pronunciation depending on context [81, 25]. This approach alleviates the data sparseness issue mentioned above: Instead of observing many instances of each word, we need only observe many instances of words susceptible to the same rules. But data sparseness still remains an issue: There are many possible phonetic contexts to consider, and many of them occur very rarely.

There is also the additional challenge of confusability, as mentioned before and studied by others previously [23, 80, 24]. While word confusions may be disambiguated by a language model, previous work has shown that it is not always enough to alleviate the problem [30]. The issue of confusability could be alleviated by using a finer-grained phonetic labeling of the observed pronunciations. For example, a more detailed transcription of the two instances of *sense* above is [s eh$_n$n t s] and [s ih$_n$t s], indicating that the two vowels were nasalized, a common phenomenon for vowels occurring before nasal consonants. Similarly, *don't* → [d ow t] is more finely transcribed [d ow$_n$t]. This labeling makes it clear that that what seems like a nasal consonant deletion in *don't* is not a deletion at all; rather, the nasalization feature surfaces on the vowel preceeding the nasal consonant. With this labeling, the second instance of *sense* is no longer confusable with *sits*, and *don't* is no longer confusable with *dote*. (The first *sense* token, however, is still confusable with *cents*.)[4] Next we consider an alternative fine-grained representation provided by sub-phonetic features.

2.2. Sub-phonetic feature-based representations

Returning to the two examples of *sense* above, it is useful to consider the process by which the [t] in these productions comes about. In order to produce an [n], the speaker must make an alveolar closure with the tongue tip, as well as lower the velum to allow airflow to the nasal cavity. To produce the following [s], the tongue closure is slightly released, and voicing and nasality are turned off. If the voicing and nasality are turned off before the tongue closure is released, this results in a segment of the speech signal with no voicing or nasality but with complete tongue tip closure; this configuration of articulators happens to be the same one used in producing a [t]. The second example of *sense* is characterized by more extreme asynchrony: Nasality and voicing are turned off even before the complete tongue closure is made, leaving no [n] and only a [t].

Such observations motivate a representation of pronunciations using, rather than a single stream of phonetic labels, multiple streams of sub-phonetic features such as nasality, voicing, and closure degrees. Tables 2–4 show such a representation of the canonical pronunciation of *sense* and of the observed pronunciations [s eh$_n$n t s] and [s ih$_n$t s]. Deviations from the canonical values are marked (*). In Table 3, all of the feature values are produced faithfully, but with some asynchrony in the timing of feature changes. In Table 4, most feature values are produced canonically, except for slightly different amounts of tongue opening accounting for the observed [ih$_n$]. This contrasts with the phonetic representation, in which half of the phones are different from the canonical pronunciation.

This representation allows us to account for the three phenomena seen in these examples—vowel nasalization, [t] insertion, and [n] deletion—with the single mechanism of asynchrony, between voicing and nasality on the one hand and the tongue features on the other. *don't* → [d ow$_n$t] is similarly accounted for, as is the common related phenomenon of epenthetic [p] insertion in words like *warmth* → [w ao r m p th].

The feature-based representation may also allow us to better handle the *sense/cents* confusability. By ascribing the [t] to part of the [n] closure gesture, this analysis predicts that a [t] inserted in this environment will be shorter than

[4] In this discussion we are not directly addressing the issue of which confusability is important to eliminate, and which is already handled by the language model. We are considering all *potential* confusability between words. A study of the detailed interaction between the language model and confusability is outside the scope of this paper.

a "true" [t]. This, in fact, appears to be the case in at least some contexts [99]. This implies that we may be able to distinguish *sense* → [s eh$_n$n t s] from *cents* based on the duration of the [t], without an explicit model of epenthetic [t] duration.

This is an example of the more general idea that we should be able to avoid confusability by using a finer-grained representation of observed pronunciations. The feature-based approach makes it possible to have a fine-grained representation, without the explosion in model size that would result if we used a finer-grained phone set. The epenthetic stop examples also suggest that pronunciation models should be sensitive to timing information. In experiments with manual annotation, it has been found that inter-annotator agreement is higher when labeling features than when labeling phones [57]; presumably this is because, whenever a speech sound is intermediate to two phones, a phone-based representation forces labelers to choose between multiple labels that are equally poor descriptions of the actual sound.

voicing	off	on		off
nasality	off		on	off
lips	open			
tongue body	mid/uvular	mid/palatal	mid/uvular	
tongue tip	critical/alveolar	mid/alveolar	closed/alveolar	critical/alveolar
phone	s	eh	n	s

Table 2: Canonical pronunciation of *sense* in terms of articulatory features.

voicing	off	on	off	
nasality	off	on	off	
lips	open			
tongue body	mid/uvular	mid/palatal	mid/uvular	
tongue tip	critical/alveolar	mid/alveolar	closed/alveolar	critical/alveolar
phone	s	eh$_n$	n	t *	s

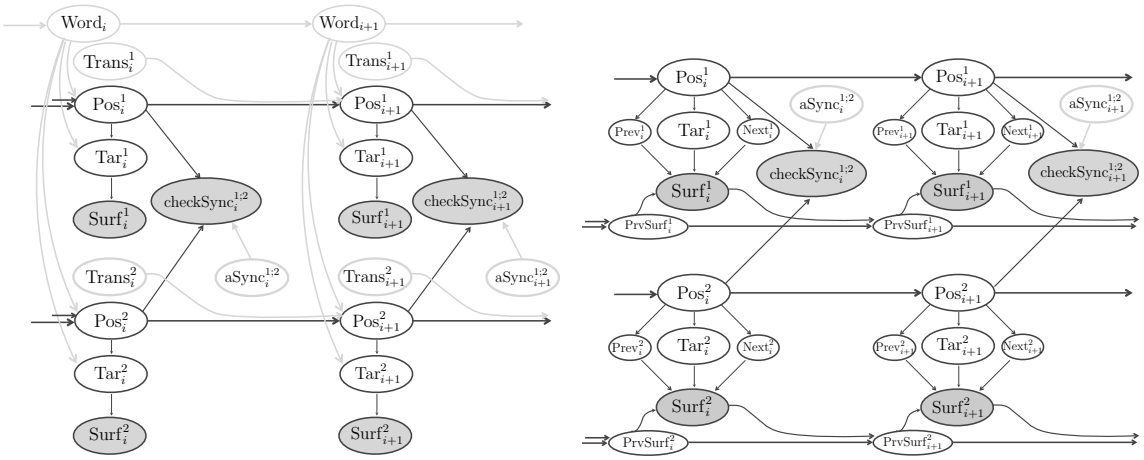
Table 3: Observed pronunciation #1 of *sense* in terms of articulatory features. Asterisks (*) indicate deviations from the baseform.

voicing	off	on	off	
nasality	off	on	off	
lips	open			
tongue body	mid/uvular	mid-narrow/palatal *	mid/uvular	
tongue tip	critical/alveolar	mid-narrow/alveolar *	closed/alveolar	critical/alveolar
phone	s	ih$_n$*	t *	s

Table 4: Observed pronunciation #2 of *sense* in terms of articulatory features. Asterisks (*) indicate deviations from the baseform.

This reasoning is in line with modern ideas in linguistics. The paradigm of a string of phonemes (plus possibly rewrite rules) is characteristic of the generative phonology of the 1960s and 1970s (e.g., [16]). More recent linguistic theories, under the general heading of non-linear or autosegmental phonology [29], have done away with the single-string representation, opting instead for multiple tiers of features. An example of this is articulatory phonology [14], which posits that most or all surface variation results from the relative timings of articulatory gestures, using a representation similar to that of Tables 2–4. Articulatory phonology is a work in progress, but one that motivates ideas in our and others' work [66, 27].

The remaining sections review a line of research using this approach: An initial model with context-independent substitutions [60, 61, 56], a more complex model with context sensitivity [9, 49, 46], evaluation via perplexity and



(a) Context-independent model (b) Context-dependent model

Figure 1: DBNs representing a feature-based pronunciation model with two features and context-independent (left) or context-dependent (right) substitution probabilities. These figures assume for simplicity that there is one baseform per word. Also, in the context-dependent model (b), the word and transition variables have been excluded for visual clarity.

lexical access, and related work that has used our models for other tasks [92, 77, 78, 36, 58, 48].

3. Feature-based pronunciation modeling

Motivated by the observations of the previous section, this section describes the proposed approach of modeling pronunciation variation in terms of the time course of multiple sub-phonetic features. The main components of the model are (1) a *baseform dictionary*, defining the sequence of target values for each feature, from which the surface realization can stray via the processes of (2) inter-feature *asynchrony*, in which different features proceed through their trajectories at different rates, and (3) *substitutions* of individual feature values, which subsumes both reductions and assimilations. Unlike in many phone-based models, we do not allow deletions or insertions of features, instead accounting for apparent phone insertions or deletions as resulting from feature asynchrony or substitution. We formalize the models as dynamic Bayesian networks (DBNs), a generalization of HMMs that allows for natural and parsimonious representations of multi-stream models. Figure 1 shows DBN representations for our models; they are described in detail in the sections below.

We define an *underlying* pronunciation of a word in the usual way as a string of phonemes, or more generally a string of phonetic units found in a baseform pronuncing dictionary.[5] We define *surface* pronunciations in a somewhat unconventional way. In Section 2.2, we proposed a representation consisting of multiple streams of feature values, as in Tables 2–4. We also mentioned, but did not formalize, the idea that the representation should be sensitive to timing information, so as to take advantage of knowledge such as the tendency of epenthetic stops to be shorter than non-epenthetic stops. To formalize this, we define a surface pronunciation as a *time-aligned* listing of all of the surface feature values produced by a speaker. Such a representation might look like Table 4 with the addition of time stamps. In practice, we assume that time is discretized into short frames, say of 10ms each as is typical in ASR.

3.1. Articulatory feature set

There is no standard feature set used in feature-based ASR research. One commonly used type of feature set is based on the categories used in the International Phonetic Alphabet (IPA) [1] to distinguish phones, such as manner,

[5] Building a useful baseform dictionary is itself an interesting problem [55], and some recent work dispenses with the pre-built baseform dictionary and instead learns all possible pronunciations from acoustic data [64]. For our purposes here we assume the baseforms are given.

place, voicing, nasalization, rounding, and height and frontness for vowels. However, for some types of pronunciation variation, this feature set does not seem well-suited. One example is the reduction of consonants to glides or vowel-like sounds. For example, a /b/ with an incomplete closure may surface as an apparent [w]. Intuitively, there is only one dimension of change, the reduction of the constriction at the lips. In terms of IPA-based features, however, this would involve a large number of substitutions: The manner would change from *stop* to *approximant*, and the vowel front/back and height features would change from *nil* to the appropriate values.

Motivated by such examples, we have used a feature set based on the vocal tract variables of Browman and Goldstein's articulatory phonology (AP) [14], shown in Table 5. We have informally assumed this type of features in examples in previous sections. These features refer to the locations and degrees of constriction of the major articulators in the vocal tract. The state space of this feature set consists of 41,472 combinations of feature values. However, in practice we do not allow all of these combinations, due to constraints that we impose on the asynchrony and substitutions in the model.

This feature set was developed with articulatory phonology as a starting point. However, since neither the entire feature space nor a complete mapping from a phone set to feature values was available in the literature, we have filled in gaps as necessary, using the guideline that the number of feature values should be kept as low as possible, while differentiating between as many phones as possible. In constructing phone-to-feature mappings, we have consulted the articulatory phonology literature [10, 11, 12, 13, 14], phonetics literature [54, 90], and X-ray tracings of speech articulation [74]. Additional details about the features and a phone-to-feature mapping are given in [56].

In preliminary experiments comparing different feature sets in our pronunciation models, these articulatory features outperformed IPA-style features [56].

feature	**description**	**values**
LIP-LOC	lip location	protruded, labial, dental
LIP-OPEN	degree of lip opening	closed, critical, narrow, wide
TT-LOC	tongue tip location	inter-dental, alveolar, palato-alveolar, retroflex
TT-OPEN	degree of tongue tip opening	closed, critical, narrow, mid-narrow, mid, wide
TB-LOC	tongue body location	palatal, velar, uvular, pharyngeal
TB-OPEN	degree of tongue body opening	closed, critical, narrow, mid-narrow, mid, wide
VELUM	state of the velum (nasality)	closed (non-nasal), open (nasal)
GLOTTIS	state of the glottis (voicing)	closed (glottal stop), critical (voiced), wide (voiceless)

Table 5: A feature set based on the vocal tract variables of articulatory phonology.

3.2. A generative recipe

In this section, we describe a procedure for generating all of the possible surface pronunciations of a given word, along with their probabilities. We denote the N features $F^j, 1 \le j \le N$. A T-frame surface pronunciation in terms of these features is denoted Surf_i^j, $1 \le j \le N, 1 \le i \le T$, where Surf_i^j is the surface value of feature F^j in time frame i.

Each word has one or more phonetic baseforms, each of which is converted to a table of underlying feature values using a phone-to-feature mapping table.[6] Dynamic phones consisting of more than one feature configuration are divided into multiple segments: Stops are divided into closure + release, affricates into closure + frication, and diphthongs into the beginning and ending configurations. The mapping from phones to feature values may be non-deterministic in some cases. Table 6 shows a possible baseform for *sense*.

The top row of Table 6 is simply an index into the underlying phone sequence; it will be needed later in the discussion of asynchrony. Note that it is assumed that all features have the same number of targets in a given word. For example, **lips** is assumed to have four targets, although they are all identical. This means that, for each phone in the baseform, and for each feature, there must be a span of time in the production of the word during which the feature is "producing" that phone. This is a basic assumption that, in practice, amounts to a duration constraint and makes it particularly easy to talk about feature asynchrony by referring to index differences.

[6] A table used in our work can be obtained from [56].

index (position)	1	2	3	4
voicing	off	on	on	off
nasality	off	off	on	off
lips	wide	wide	wide	wide
tongue body	mid/uvular	mid/palatal	mid/uvular .5 mid/velar .5	mid/uvular
tongue tip	critical/alveolar	mid/alveolar	closed/alveolar	critical/alveolar
phone	s	eh	n	s

Table 6: A possible baseform and target feature values for the word *sense*. The target value for the feature **tongue body** for an [n] is *mid/velar* or *mid/uvular* with probability 0.5 each, and the remaining feature values are deterministic.

3.2.1. Asynchrony

We assume that in the first frame of a word, all of the features begin in index 1. Denoting the index, or position, of feature j at time i Pos_i^j, this means $\text{Pos}_1^j = 1\ \forall j$.[5] In subsequent frames, each feature can either stay in the same state or transition to the next one with some (possibly phone-dependent) transition probability. Features may transition at different times, and this is what we refer to as feature *asynchrony*. We define the *degree of asynchrony* between two features F^j and F^k in a given time frame i as the absolute difference between their indices in that frame:

$$\text{aSync}_i^{j:k} = |\text{Pos}_i^j - \text{Pos}_i^k|. \tag{1}$$

Similarly, we define the degree of asynchrony between two sets of features F^A and F^B as the difference between the means of their indices, rounded to the nearest integer:

$$\text{aSync}_i^{A:B} = round\left(|mean\left(\text{Pos}_i^A\right) - mean\left(\text{Pos}_i^B\right)|\right), \tag{2}$$

where A and B are subsets of $\{1, \ldots, N\}$ and $F^{\{j_1, j_2, \ldots\}} = \{F^{j_1}, F^{j_2}, \ldots\}$. More "synchronous" configurations may be preferred, and there may be an upper bound on the degree of asynchrony. We express this via a distribution over the degree of asynchrony between features in each frame, $p(\text{aSync}_i^{j:k})$, or feature sets, $p(\text{aSync}_i^{A:B})$. Given the index sequence for each feature, the corresponding frame-by-frame sequence of target feature values, Tar_i^j, $1 \leq i \leq T$, is drawn according to the baseform table (Table 6).

3.2.2. Substitution

A feature may fail to reach its target value, instead substituting another value. This may happen, for example, if a constriction is reduced, or if a given feature value assimilates to neighboring values. One example of substitution is *sense* $\rightarrow$ [s ih$_n$t s]; a frame-by-frame representation is shown in Table 7. We model substitutions with a conditional distribution over each surface feature value in a given frame, given its corresponding underlying value and other context variables, $p(\text{Surf}_i^j|\text{Tar}_i^j, \text{c}_i^j)$, where c_i^j is a vector of context variables such as neighboring feature values. In initial experiments looking at the separate effects of asynchrony and substitutions, we have found that substitutions alone account for more conversational pronunciations than asynchrony alone, but both make a significant contribution [56].

3.2.3. Summary

To summarize the generative recipe, in this model a surface pronunciation of a given word is produced in the following way:

1. Choose a baseform from the baseform dictionary.
2. For the chosen baseform, generate state sequences for all features, with probabilities given by the transition and asynchrony probabilities.

[5] This corresponds to the (strong) assumption that there is no cross-word asynchrony. Relaxing this assumption is in principle straightforward but is a topic for future work.

frame	1	2	3	4	5	6	7	8	9	10
voicing index (Pos)	1	1	2	3	3	3	4	4	4	4
voicing phone	s	s	eh	n	n	n	s	s	s	s
voicing (Tar)	off	off	on	on	on	on	off	off	off	off
voicing (Surf)	off	off	on	on	on	on	off	off	off	off
nasality index (Pos)	1	1	2	3	3	3	4	4	4	4
nasality phone	s	s	eh	n	n	n	s	s	s	s
nasality (Tar)	off	off	off	on	on	on	off	off	off	off
nasality (Surf)	off	off	off	on	on	on	off	off	off	off
tongue body index (Pos)	1	1	2	2	2	3	3	4	4	4
tongue body phone	s	s	eh	eh	eh	n	n	s	s	s
tongue body (Tar)	m/u	m/u	m/p	m/p	m/p	m/u	m/u	m/u	m/u	m/u
tongue body (Surf)	m/u	m/u	**m-n/p**	**m-n/p**	**m-n/p**	m/u	m/u	m/u	m/u	m/u
tongue tip index (Pos)	1	1	2	2	2	3	3	4	4	4
tongue tip phone	s	s	eh	eh	eh	n	n	s	s	s
tongue tip (Tar)	cr/a	cr/a	m/a	m/a	m/a	cl/a	cl/a	cr/a	cr/a	cr/a
tongue tip (Surf)	cr/a	cr/a	**m-n/a**	**m-n/a**	**m-n/a**	cl/a	cl/a	cr/a	cr/a	cr/a
$\text{aSync}^{A:B}$	0	0	0	1	1	0	1	0	0	0
phone	s	s	ih	ih_n	ih_n	n	t	s	s	s

Table 7: Frame-by-frame sequences of index values, corresponding phones, underlying (Tar) and surface (Surf) feature values, and degrees of asynchrony between A = {voicing, nasality} and B = {tongue body, tongue tip}, for a 10-frame production of *sense*. Where the underlying feature value is non-deterministic, only one of the values is shown for ease of viewing. Bold-faced entries indicate feature value substitutions. The lips feature has been left off and is assumed to be "wide" throughout. The bottom row shows the resulting phone transcription corresponding to these feature values.

3. Given the generated index sequence, generate underlying feature values by drawing from the baseform feature distributions at each index. (This is a deterministic lookup in the case of a single baseform.)
4. For each underlying feature value, generate a surface feature value by drawing from the substitution distribution.

3.3. Representation via dynamic Bayesian networks

The model we have described can be naturally represented as a DBN. Figure 1(a) shows a DBN representing our model with two feature streams and with context-independent feature substitutions. In this article, we model each word independently, although ultimately we should include context from the previous and following words as well. In the model with context-independent substitutions, the variables at time frame i are as follows:

Word_i – The current word at time i.

Trans_i^j – binary transition variable. $\text{Trans}_i^j = 1$ indicates that frame i is the last frame of the current state for feature j.

Pos_i^j – index of feature j at time i. $\text{Pos}_1^j = 1\ \forall j$; in subsequent frames $\text{Pos}_{i+1}^j = \text{Pos}_i^j$ if $\text{Trans}_i^j = 0$, and $\text{Pos}_i^j + 1$ otherwise.

Tar_i^j – underlying (target) value of feature j.

Surf_i^j – surface value of feature j. $p(\text{Surf}_i^j|\text{Tar}_i^j)$ encodes allowed feature substitutions.

$\text{aSync}_i^{A:B}$ and $\text{checkSync}_i^{A:B}$ represent the asynchrony model. $\text{aSync}_i^{A:B}$ is drawn from an (unconditional) distribution over the integers, while $\text{checkSync}_i^{A:B}$ checks that the degree of asynchrony between A and B is in fact equal to $\text{aSync}_i^{A:B}$. To enforce this constraint, $\text{checkSync}_i^{A:B}$ is always observed with value 1 and is given deterministically by its parents' values: $\text{checkSync}_i^{A:B} = 1 \iff \mathit{round}\left(\left|\mathit{mean}\left(\text{Pos}_i^A\right) - \mathit{mean}\left(\text{Pos}_i^B\right)\right|\right) = \text{aSync}_i^{A:B}$

If we assume that the features always synchronize at the end of a word, then in the final frame we must enforce that all features are exiting their final state. Cross-word asynchrony (as in "green beans" → [g r iy m b iy n z]) and cross-word coarticulation are also important, and can be incorporated into such a model, but as mentioned previously, we limit the presentation to the case where asynchrony and conditioning contexts are restricted to be within word boundaries.

We can use standard DBN inference algorithms to answer such questions as:

- **Decoding:** Given a surface pronunciation (set of surface feature value sequences $\text{Surf}_{1:T}^{1:N}$), what is the most likely word that generated them?
- **Parameter learning:** Given a training set of words and corresponding surface pronunciations, what are the best settings of the conditional probability tables in the DBN?[7]
- **Alignment:** Given a word and a corresponding surface pronunciation, what are the most likely values of the hidden variables Pos_i^j and Tar_i^j? This is useful for analysis or qualitative evaluation.

Parameter learning can be done with a maximum-likelihood criterion via the Expectation-Maximization algorithm [17], or with discriminative criteria (e.g., [75]).

3.4. Modeling context-dependent feature substitutions

Figure 1(b) shows a version of the model where the surface features depend on additional context variables besides the target feature values. In this case the context variables are the previous and next target values of the feature and the previous surface value (denoted Prev_i^j, Next_i^j and PrvSurf_i^j respectively), but many other context variables are possible (refer to Section 4.2 for some examples). Dependencies between surface feature values can encode smoothness constraints in the motion of articulators; and conditioning surface feature values on past or future target values can model assimilation effects that are not accounted for by asynchrony. Analogously to previous work on phone-based pronunciation models [80, 25], we can use decision trees to represent the context-dependent surface feature distributions. In such a setup, we have a decision tree for every target feature value, with questions about the context variables determining splits in the decision trees.

In addition to standard decision trees, we have also considered maximum-entropy models [5] of the surface feature distributions [46]. These may be more robust since, unlike decision trees, their training is a convex optimization problem and involves fewer tuning parameters. In this approach, we learn maximum-entropy predictors of the surface feature values given the target and context variables, using target/context variable values as feature functions in the predictors:

$$p(\text{Surf}|\text{Tar}, \text{c}; \lambda) = \frac{1}{Z(\text{Tar}, \text{c})} \exp\left[\sum_{k=1}^{K} \lambda_k f_k(\text{Tar}, \text{c}, \text{Surf})\right], \tag{3}$$

where c refers to all of the context variables, $f_k(\cdot)$ are feature functions (in our case they are properties of the context variables), λ_k are weights learned to maximize the conditional likelihood of the training data, and $Z(\text{Tar}, \text{c})$ is the partition function.

In order to train either the decision trees or the maximum-entropy predictors, we need training data with all context variables observed. For this purpose we have used alignments generated using a trained context-independent model, similarly to the training of phone-based models such as those of Riley *et al.* [80].

3.5. Related work

Besides the earlier motivating work mentioned in Section 1 [19, 22, 52, 79], the most closely related work to ours is the recent work of Mitra *et al.* [66] on noise robustness with articulatory models. In their work, Mitra *et al.* define a DBN with multiple streams of articulatory variables based on articulatory phonology. In their models, however, there is no explicit modeling of inter-feature asynchrony or substitutions.[6] Markov *et al.* [62] also use DBNs with

[7] This assumes training data in the form of labeled surface feature values. If these are not available, but only acoustics are, training can be done by combining the model with an acoustic observation model; this is outside the scope of this article.

[6] For this reason we have not compared directly against this model in experiments. Since our experiments are based on manual transcriptions, while their model accounts for a large amount of variation at the acoustic level, that model would perform unfairly poorly in our experiments.

articulatory variables, but with no modeling of inter-frame dependencies. Some authors have explored DBN models for the task of recognition of asynchronous articulatory features [96]. Other work has explored the use of multiple asynchronous streams of variables other than sub-phonetic features, such as different streams of acoustic observations or acoustic observations with the addition of an auxiliary variable [70, 89, 102, 101]. Finally, in linguistics and speech science there have been several efforts to formalize models of multiple asynchronous tiers [41, 98] and a simulation of articulatory phonology itself has now been implemented in a toolkit [68].

There has also been a great deal of work in speech recognition on *acoustic observation models* based on sub-phonetic features and on feature classification [53, 65, 21, 26, 97, 51, 15, 67, 87]. Such methods can be combined with articulatory feature-based pronunciation models to build complete speech recognizers. The outputs of feature classifiers can be used as observations for surface feature variables, or alternatively the feature variables can be kept hidden and inferred (or marginalized out) in recognition. In this article we focus on pronunciation modeling alone, although the two topics are linked.

4. Evaluation using manual phonetic transcriptions

There are multiple ways of evaluating the proposed approach. For purposes of ASR, we could of course embed the model in a complete recognizer. Sub-phonetic feature-based pronunciation models have indeed been evaluated in complete ASR systems [18, 79, 66, 58], but always using a simplified pronunciation model for computational or other reasons. It is also difficult to discern in such ASR experiments how much of the performance differences are attributable to the new pronunciation model versus the corresponding new observation models. In the work reviewed in this article, therefore, we have chosen to use "intrinsic" evaluation, external to any ASR system.

In the past, the value of new pronunciation models has been evaluated in a few "intrinsic" ways. For example, an extensive study by Riley *et al.* [80] used the perplexity of the surface phone strings in a test set as a performance criterion. Bates *et al.* [3] also used phone perplexity, as well as phonetic error rate and distance. We cannot use the same phone perplexity measure, since our models do not produce a probability per surface phone label. Our models do produce a per-frame probability of the surface form, so we use frame perplexity as one measure of performance.

A second type of evaluation we use is via a lexical access task, that is classification of a word's identity given its surface phonetic transcription. We compared our feature-based models to a phonetic baseline model based on the one of Riley *et al.* [80], described below.

4.1. Data

In the experiments described in this section, we used data from the Switchboard Transcription Project (STP) [34], a subset of the Switchboard conversational speech database [28] that has been manually labeled at a fine phonetic level, including various diacritics, and phonetically time-aligned. This dataset provides a set of examples of conversational surface pronunciations, which we convert to frame-by-frame articulatory feature transcriptions via a mapping from the STP phone set to our feature set. All of our experiments have been done on the "train-ws96-i" subset of the STP transcriptions, excluding partial words, words whose transcriptions contain non-speech noise, and words whose baseforms are four phones or shorter (where stops, affricates, and diphthongs are considered two phones each). The length restriction is intended to exclude words that are so short that most of their pronunciation variation is caused by neighboring words. The resulting data set contains 3343 single-word tokens excised from continuous conversational utterances.

Since the surface pronunciations are transcribed phonetically in this data, we may not be able to exploit the full power of the feature-based model in this evaluation. For example, partially nasalized vowels, incomplete stop closures, and anticipatory rounding are handled by the model but are missing from the phonetic transcriptions. This may be addressed in the future by either manual articulatory transcription efforts [57] or automatic forced articulatory transcription [77].

4.2. Evaluation via frame perplexity [9, 49]

Since the feature-based model is inherently frame-based, we measure the frame-level perplexity. We therefore also trained frame-level phonetic decision trees as a baseline for this evaluation. In these experiments, the STP excised word data described above is split into a 2942-word (~90,000-frame) train set, 165-word development set,

and 236-word test set (see [56, 9] for more details). The goal is to measure how well a model predicts a test set of surface pronunciations. Let the surface phone label at time frame i be P_i and the corresponding N-feature vector be $\{\text{Surf}_i^1, \ldots, \text{Surf}_i^N\} = \text{Surf}_i^{1:N}$. To evaluate our models, we align the test data (using the context-independent model), i.e. we find the most probable values of all hidden variables given the word and the surface labels, and then compute the frame-level perplexity of the test set:

$$perp(P_1, \ldots, P_T) = 2^{-\frac{1}{T}\sum_t \log_2 \; p(P_i|\mathrm{c}_i)} \tag{4}$$

where there are T frames in the test data and c_i refers to the context variables. In this expression, $p(P_i|\mathrm{c}_i)$ is computed from either phonetic or feature-based decision trees, via $p(P_i|\mathrm{c}_i) = \prod_{j=1}^{N} p(\text{Surf}_i^j|\mathrm{c}_i^j)$.

Type of context model	**Phone-based**	**Feature-based**
Context-independent	3.65	2.57
Basic context-dependent	2.52	2.15
Basic context-dep. + previous distinct surface value	1.64	1.79
Basic context-dep. + prev. surf. + distance	2.08	1.69
Cross-feature context-dep. + prev. surf.	N/A	1.57
Cross-feat. context-dep. + prev. surf. + phone context	N/A	1.52

Table 8: Test set perplexities for various models (from [9]). In the context-independent model, the surface label depends only on its current target value. *Basic context-dependent* also includes a dependency on the next and previous distinct target values. *Distance* refers to the distance in frames to the next/previous distinct target value, providing a model of inertia. In *cross-feature* models, the context includes the target and surface values for all features, not just the one being predicted. Context dependency across feature streams (the last two rows) can not be applied to phone-based models, corresponding to the "N/A" entries in the table.

Building a separate decision tree for each articulatory feature produces poor results, 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, which yields much better results: all tongue features (19 possible underlying values, and therefore 19 trees); glottis and velum (3 trees); and lip opening (4 trees). These bundles were chosen because there is typically very little asynchrony among the features within each bundle, but more so in between bundles. Figure 4.2 shows a (highly pruned) example decision tree for the Glottis-Velum feature tier, corresponding to the case where the target value is voiced (VD) and non-nasal (N-N). This tree shows that the current sound is more likely to be nasalized if the following target is a nasal than if it is a non-nasal, and even more likely to be nasalized if the previous surface sound was also nasal.

Table 8 shows the frame perplexities for various models. For most sets of context variables, the feature-based models outperform the phone-based ones. The best perplexities are obtained by including cross-feature context. It can be difficult to interpret the importance of perplexity differences, so next we consider evaluation through lexical access performance.

4.3. Evaluation via lexical access [60, 61, 56, 46]

In the lexical access task we address the question: If we knew the true sequences of surface feature values $\text{Surf}_i^j \forall i, j$ for a word, how well could we guess the identity of the word? In this case, the "true" surface feature values are derived from the STP phonetic transcriptions, by assuming a deterministic mapping from surface phones to surface feature values. The task then consists of introducing these surface feature values as observations of $S = \text{Surf}_i^j \forall i, j$ in the DBN for each word, and finding the word with maximum posterior probability, $p(w_k|S), 1 \leq k \leq V$, where V is the vocabulary size. Given a test set of words with their observed surface feature values, we measure the lexical access error rate of the model as the percentage of the test words that are predicted incorrectly.

In these experiments the parameter learning is done via maximum likelihood using the EM algorithm, given the training set of observed word/surface feature pairs. All DBN inference and parameter learning is done using the Graphical Models Toolkit (GMTK) [7, 6]. The maximum-entropy models are trained using the Maximum Entropy Modeling Toolkit [100]. We enforce two hard constraints on asynchrony:

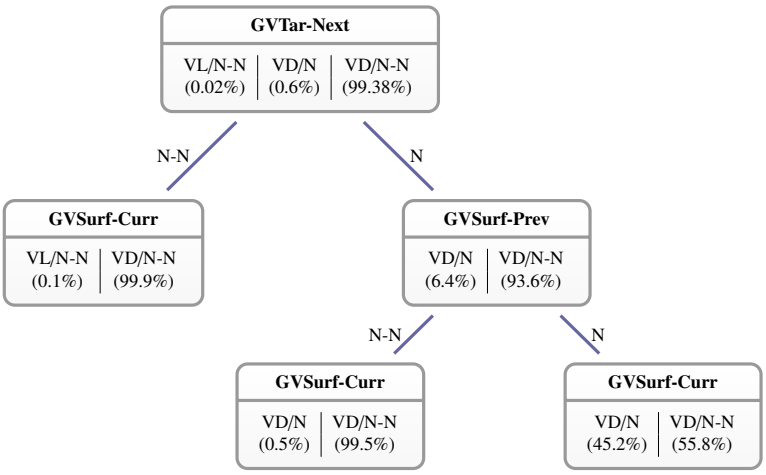


Figure 2: An example decision tree learned for the Glottis-Velum feature tier, when the current target value is VD/N-N. The Glottis tier takes values of voiced (VD) or voiceless (VL) and the Velum tier's values are either non-nasal (N-N) or nasal (N).

1. The lips can desynchronize from the tongue by up to one state: $p(\text{aSync}_i^{L:T} > 1) = 0$. Lip-tongue asynchrony can account for effects such as vowel rounding in the context of a labial consonant. We ignore for now longer-distance lip-tongue asynchrony effects, such as the rounding of [s] in *strawberry*.
2. The glottis/velum must be within 2 states of the tongue and lips: $p(\text{aSync}_i^{L,T;GV} > 2) = 0$. This accounts for longer-distance effects, such as long-range nasalization as in *trying* → [t r ay$_n$n] (where the nasalized [ay] is two phones away from the nasal in the underlying pronunciation).

In addition, we set many of the substitution probabilities to zero, based on the assumption that features will not stray too far from their intended values.

We compare our models to a baseline phonetic pronunciation model built according to the specification in [80]. The phonemic transcriptions were aligned with the hand-labeled phonetic transcriptions from STP to give phoneme-to-phone correspondences which were then used to build phonetic decision tree models. The phonemes were each represented as a six-element feature vector: (type of phoneme (vowel,consonant,silence), consonant-manner, consonant-place, vowel-manner, vowel-place, nasalized or not nasalized). The decision tree questions were based on these features. As in [80], we also allowed for deletion of phonemes in context. The context for the decision trees includes three neighbouring phonemes on either side and the distance of the phoneme from the word boundary on either side, as in [80].

The pronunciation models in these experiments used a 3328-word vocabulary, consisting of the 3500 most likely words in the "Switchboard I" training set [28], excluding partial words, non-speech, and words for which we did not have baseform pronunciations.

Figure 3 shows the lexical access error rates of the context-dependent phonetic baseline and of several feature-based models, in a 5-fold experimental setup with 2000 words in each fold's training set and about 670 words in the development and test sets.

The main result here is that the context-dependent feature-based models outperform the context-dependent phonetic baseline, despite the fact that the phonetic baseline uses a longer context window (three previous/following phones) and additional context variables, and despite the fact that the phonetic labeling of the data favors a phonetic model. These improvements in performance are statistically significant according to McNemar's test [20] at $p < 0.01$. A context-independent feature-based model, however, is insufficient to beat the phonetic baseline. Finally, there is less variability in error rates across folds for the context-dependent feature-based models than for the phone-based model. Context-dependent models based on decision trees and maximum-entropy models perform comparably–the decision tree models slightly outperform the maximum-entropy models, but not significantly so–but the maximum-entropy models require much less tuning to obtain good performance so they may be preferable overall.

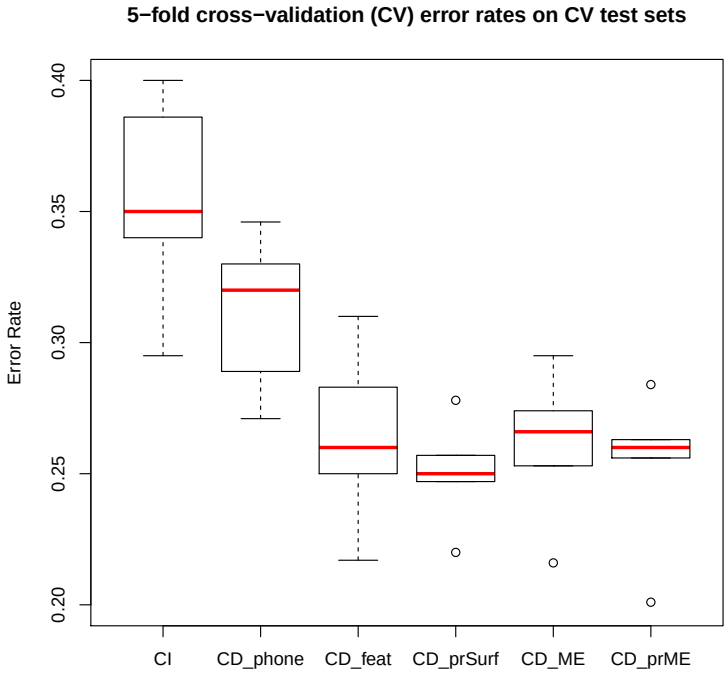
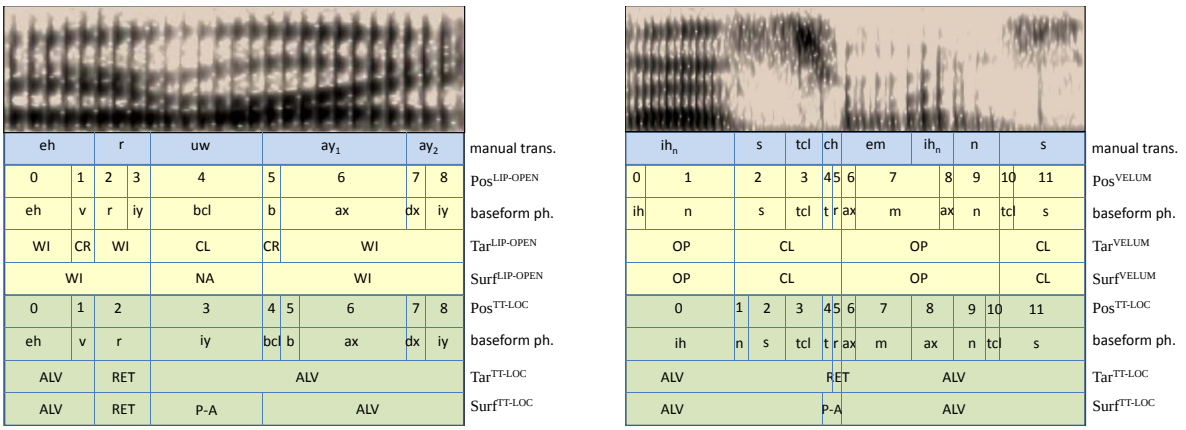


Figure 3: Lexical access error rates. *CI* is a context-independent feature-based model. *CD_Phone* is the context-dependent phone baseline model. *CD_Feat* is a basic context-dependent feature-based model using decision trees, where the context variables are only the previous and next target values of each feature. *CD_prSurf* also includes the previous distinct surface value as a context variable. *CD_ME* and *CD_prME* are the corresponding maximum entropy models without and with the previous surface feature value context.

For a more detailed look, it is also informative to consider not only the error rate but also the rank of the correct word. The correct word may not be top-ranked because of true confusability with other words; it is then instructive to compare different systems as to their relative rankings of the correct word. In a real-world connected speech recognition scenario, confusable words may be disambiguated based on the linguistic and phonetic context. The role of the pronunciation model is to give as good an estimate as possible of the goodness of fit of each word to the observed signal. Figure 5 shows the cumulative distribution function of the correct word's rank for the same feature-based and phone-based models as in Figure 3, for one of the five folds. This plot is equivalent to the r-best oracle accuracy for varying r; the previously reported error rate is given by the first point on the plot for each model. This figure demonstrates that, not only are the context-dependent feature-based models better than the phonetic baseline in terms of 1-best accuracy, they are also markedly better in terms of r-best accuracy for $r > 1$. In fact, the phone-based model tapers off at <85% r-best accuracy even for very high r, while the best feature-based models quickly reach about 95% r-best accuracies. Note that, due to zero probabilities in each model, there are some words that fall outside the r-best list for any r, no matter how large. For example, the articulatory models have limits on asynchrony and allowed substitutions, and the phone-based models do not allow all substitutions. However, this does not detract from the overall trend in relative performance of the models, which is visible even for very low r where this "ceiling effect" is not reached.

4.4. Examples

In order to get a qualitative sense of whether a model is behaving reasonably, we can look at the most likely settings for the hidden variables given a word and its surface realization, which we refer to as an alignment. This is the multi-stream analogue of a phonetic alignment, and is the model's best guess for how the surface pronunciation



(a) *everybody* → [eh r uw ay]

(b) *instruments* → [ih$_n$s tcl ch em ih$_n$n s]

Figure 4: Spectrogram, phonetic transcription, and partial alignment for two example words from the STP data set. Each "baseform phone" tier gives the baseform pronunciation aligned to the *Pos* sequence above it. Abbreviations of feature values are as given in Table 5.

was generated. Figures 4(a) and 4(b) show spectrograms and the most likely sequences of some of the model variables for two example words, *everybody* → [eh r uw ay] and *instruments* → [ih$_n$s tcl ch em ih$_n$n s]. Multiple frames with identical variable values have been merged for visual clarity.

Considering first the analysis of *everybody*, it suggests that (i) the deletion of the [v] is caused by the substitution **critical → wide** in the **LIP-OPEN** feature, and (ii) the [uw] comes about through a combination of asynchrony and substitution: The lips begin to form the closure for the [b] while the tongue is still in position for the [iy], and the lips do not fully close but reach only a narrow constriction. Turning to the example of *instruments*, the apparent deletion of the first [n] and nasalization of both [ih]s is explained by asynchrony between the velum and other features. The replacement of /t r/ with [ch] is described as a substitution of a palato-alveolar **TT-LOC** for the underlying alveolar and retroflex values. We cannot be certain of the correct analyses, but these analyses seem reasonable given the phonetic transcriptions.

It is also instructive to consider examples that different models recognize correctly/incorrectly. For example, the context-independent feature-based model fails to recognize the examples *favorite* → [f ey v er t] and *twenty* → [t w eh n iy]. In the case of *twenty*, the canonical pronunciation is /t w eh n t iy/. In this surface realization, the stop is nasalized, and in this case there is insufficient oral pressure built up to cause a burst when the stop is released. This is captured by the context-dependent distribution of the tongue tip feature, which goes directly from a closure to the following vowel in the context of a nasal.

5. Extensions

The models described in this article, or variations on these models, have been extended in various ways and applied to other tasks. First, the experimental results in this article were obtained with generative models trained via maximum likelihood. In related work, we have developed an approach for discriminative training of the models, by converting the DBNs to finite-state transducers and learning their arc weights discriminatively, for improved lexical access performance [47]. Another discriminative approach is to use the models to define feature functions in a log-linear classifier. This approach has been applied in a discriminative whole-word model for lexical access, leading to large improvements over the lexical access results reviewed here [92]; however, such whole-word models are less straightforward to extend to end-to-end recognition since they cannot be incorporated into frame-based recognition models. A related model has also been used to define feature functions in a discriminative keyword spotter, showing improvement over a phone-based model in conversational speech in a low-data setting [78].

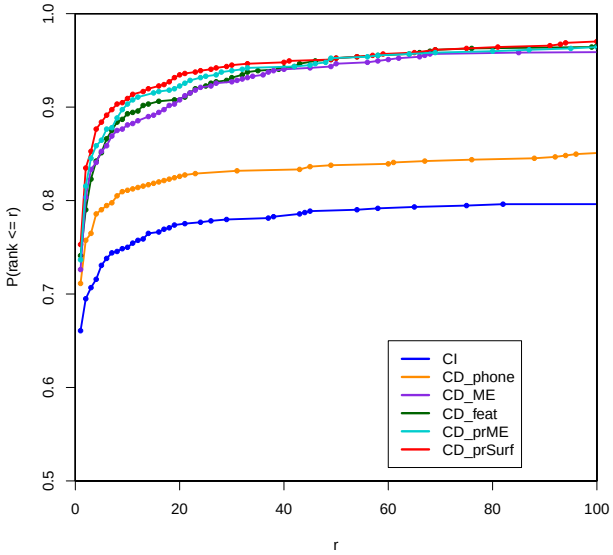


Figure 5: Empirical cumulative distribution functions of the correct word's rank. The model name abbreviations are as in Figure 3.

As mentioned previously, this type of model has been used in end-to-end speech recognition experiments, although with highly constrained versions of the models (with greatly restricted or eliminated substitution models). Extension to full speech recognition could be done by first predicting articulatory feature values and then using them as observations in the pronunciation model. Such hard decisions are likely to introduce additional errors. Instead, the articulatory features can be kept as hidden variables, and the distribution of the acoustic observations can be modeled as a Gaussian mixture [58] or indirectly in a hybrid approach [35]. With the dramatic recent improvements in hybrid models using deep neural networks [40], this approach should be revisited.

The same kind of model has been used in visual and audio-visual speech recognition [36, 82], with improvements found in lipreading performance. In the case of audio-visual speech, the model provides an account of the well-known phenomenon of audio-video asynchrony [69, 32, 31, 37] as the result of asynchrony between visible (e.g., lip opening) and non-visible (e.g., nasality, back articulations of the tongue) articulatory gestures. For example, anticipatory lip rounding can make it appear that the visual signal is "ahead" of the acoustic signal, whereas in fact it is the lips that are ahead of the other articulators.

Finally, pronunciation models can be used for a variety of other applications. One example is the development of "word neighborhood density" measures, which are used in psycholinguistics work to predict human word recognition performance [2] and to analyze speech recognition performance [30]. In recent work we have used the articulatory feature-based models presented here, converted to finite-state transducers, to define word neighborhoods by finding the nearest neighbors of a given word in terms of transducer edit distance. Word neighborhoods defined in this way outperform phone-based neighborhood measures as predictors of ASR errors [48].

6. Conclusion

This article has reviewed our work on a flexible model class that explicitly accounts for articulatory asynchrony and substitution effects in pronunciation variation. We have argued, and have found in initial experiments, that the proposed articulatory feature set (as opposed to, for example, IPA-based features) is not an arbitrary choice but an important one for modeling pronunciation variation. The graphical model formulation–as opposed to "compiling" the model into an HMM–allows us to implement such models with minimal assumptions and to take advantage of the

parsimony of the factored state space. Our main findings from frame perplexity and lexical access experiments are that

- Feature-based pronunciation models of the type we have developed outperform comparable phone-based models in most cases, in terms of both frame perplexity and lexical access performance.
- Context-dependent surface feature models outperform context-independent ones. Maximum-entropy and decision tree-based models perform comparably, but maximum-entropy models are easier to train and tune.
- In context-dependent models, it is helpful to combine the feature tiers into "bundles" of highly interdependent feature subsets; without this we do not obtain improved perplexities over the phonetic baselines.

We have also reviewed extensions of our models to allow for discriminative training and for application to speech recognition, lipreading and audio-visual speech recognition, spoken term detection, and speech recognition error prediction. There are still many opportunities for future work on properly integrating such a pronunciation model into a speech recognizer or other downstream tasks. For example, the relationship between the pronunciation model and acoustic model requires further study: Should the acoustic model be factored into multiple terms for the different articulators? Should there be a separate factor for each bundle? How much of the variation should be accounted for by the acoustic model versus the pronunciation model? Discriminative training is likely to be a big part of such integration, considering that we do not know what the "perfect" feature set and model structure are.

The models presented in this paper can be improved in a number of ways. Studies of pronunciation variation in the ASR literature suggest some useful contextual factors that we have not yet used in our models, such as prosodic and position-dependent factors. For example, Greenberg *et al.* have studied pronunciation variation as a function of the position of a phone within a syllable, and found that the codas of syllables are far more likely to be realized non-canonically [33]. Other prior work [73] has suggested that there is a "hidden mode" or speaking style that may vary during the course of an utterance, which may be useful to model as well. Finally, a number of authors have shown that pronunciation variants depend on contextual factors such as nearby disfluencies (e.g., hesitations), word predictability, utterance position, and speaking rate [4, 24, 25]. In addition, incorporating cross-word context is important and in principle straightforward in our models, but we have thus far ignored cross-word effects on both asynchrony and substitutions.

We have also assumed that the distribution of asynchrony is symmetrical: The probability of feature i being ahead of feature j by a certain amount is the same as that of j being ahead of i. There are common examples of variation that cause us to doubt this assumption. For example, pre-nasalization of vowels appears to be more common than post-nasalization [14]. The existence of attested phenomena such as *football* → [f uh b ao l] [34] but not, as far as we know, of *haptic* → [h ae p ih k] implies that tongue-lip asynchrony may also be asymmetric. Asymmetric asynchrony can be added to the model by allowing the aSync variables to take on both positive and negative values. This has begun to be explored in some related work [76, 93] but deserves more attention in future work.

Browman and Goldstein use evidence from linguistic and articulatory data to devise specific ordering constraints among their vocal tract variables [14]. We have thus far also used such considerations in deciding on the constraints used in our experiments. However, in the absence of conclusive data on all variables of interest, it would be useful to investigate automatically learning certain aspects of the model, such as the feature bundles and asynchrony constraints. For example, the feature bundles could be automatically discovered from forced alignments with an "unbundled" model. The optimal structure of the model for ASR purposes may differ from a linguistically faithful model. For this purpose discriminative training approaches [8, 47], or using our models to define features in discriminative log-linear models [78, 92], may be fruitful.

The type of model we have proposed may also have applications in speech analysis, for both scientific exploration and more immediate applications. One possible use of the model would be to make automatic articulatory transcriptions of large amounts of recorded speech, to allow the study of articulation phenomena on a larger scale than is possible with existing corpora [77]. For example, we could learn the relative timing of articulatory gestures, therefore contributing to the theory of articulatory phonology.

7. Acknowledgments

A number of colleagues have contributed in various ways to the work reviewed in this paper, including Sam Bowman, Jim Glass, Joseph Keshet, Rohit Prabhavalkar, and Hao Tang. This paper was improved by comments from the anonymous reviewers. 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.

References

[1] Albright, R. W., 1958. The International Phonetic Alphabet: Its background and development. International Journal of American Linguistics 24 (1, part 3).

[2] Bailey, T. M., Hahn, U., 2001. Determinants of wordlikeness: Phonotactics or lexical neighborhoods? Journal of Memory and Language 44 (4), 568–591.

[3] Bates, R., Ostendorf, M., Wright, R. A., 2007. Symbolic phonetic features for modeling of pronunciation variation. Speech Communication 49 (2), 83–97.

[4] Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M., Gildea, D., 2003. Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. Journal of the Acoustical Society of America 113 (2), 1001–1024.

[5] Berger, A. L., Pietra, V. J. D., Pietra, S. A. D., 1996. A maximum entropy approach to natural language processing. Computational Linguistics 22 (1), 39–71.

[6] Bilmes, J., 2002. GMTK: The Graphical Models Toolkit. `http://ssli.ee.washington.edu/ bilmes/gmtk/`.

[7] Bilmes, J., Zweig, G., May 2002. The Graphical Models Toolkit: An open source software system for speech and time-series processing. In: Proceedings of ICASSP.

[8] Bilmes, J., Zweig, G., Richardson, T., Filali, K., Livescu, K., Xu, P., Jackson, K., Brandman, Y., Sandness, E., Holtz, E., Torres, J., Byrne, B., May 2002. Structurally discriminative graphical models for automatic speech recognition. In: Proceedings of ICASSP.

[9] Bowman, S., Livescu, K., 2010. Modeling pronunciation variation with context-dependent articulatory feature decision trees. In: Proceedings of Interspeech.

[10] Browman, C. P., Goldstein, L., 1986. Towards an articulatory phonology. Phonology Yearbook 3, 219–252.

[11] Browman, C. P., Goldstein, L., 1989. Articulatory gestures as phonological units. Phonology 6, 201–251.

[12] Browman, C. P., Goldstein, L., 1990. Gestural specification using dynamically-defined articulatory structures. Journal of Phonetics 18, 299–320.

[13] Browman, C. P., Goldstein, L., 1990. Tiers in articulatory phonology, with some implications for casual speech. In: Kingston, J., Beckman, M. E. (Eds.), Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech. Cambridge University Press, Cambridge, UK, pp. 341–376.

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

[15] Cetin, O., Kantor, A., King, S., Bartels, C., Magimai-Doss, M., Frankel, J., Livescu, K., 2007. An articulatory feature-based tandem approach and factored observation modeling. In: Proceedings of ICASSP.

[16] Chomsky, N., Halle, M., 1968. The Sound Pattern of English. Harper & Row, New York, NY.

[17] Dempster, A. P., Laird, N. M., Rubin, D. B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society 39 (1), 1–38.

[18] Deng, L., Ramsay, G., Sun, D., 1997. Production models as a structural basis for automatic speech recognition. Speech Communication 33, 93–111.

[19] Deng, L., Sun, D. X., 1994. A statistical approach to automatic speech recognition using the atomic speech units constructed from overlapping articulatory features. Journal of the Acoustical Society of America 95 (5), 2702–2719.

[20] Dietterich, T. G., 1998. Approximate statistical tests for comparing supervised classification learning algorithms. Neural Computation 10 (7), 1895–1924.

[21] Eide, E., 2001. Distinctive features for use in an automatic speech recognition system. In: Proceedings of Eurospeech.

[22] Erler, K., Freeman, G. H., 1994. An articulatory-feature-based HMM for automated speech recognition. Journal of the Acoustical Society of America 95 (5), 2873.

[23] Finke, M., Waibel, A., 1997. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In: Proceedings of Eurospeech.

[24] Fosler-Lussier, E., Morgan, N., 1999. Effects of speaking rate and word frequency on pronunciations in conversational speech. Speech Communication 29 (2), 137–158.

[25] Fosler-Lussier, J. E., 1999. Dynamic pronunciation models for automatic speech recognition. PhD dissertation, U. C. Berkeley, Berkeley, CA.

[26] Frankel, J., King, S., 2001. ASR – Articulatory Speech Recognition. In: Proceedings of Eurospeech.

[27] Ghosh, P. K., Narayanan, S., 2011. Automatic speech recognition using articulatory features from subject-independent acoustic-to-articulatory inversion. Journal of the Acoustical Society of America 130 (4), 251–257.

[28] Godfrey, J. J., Holliman, E. C., McDaniel, J., 1992. SWITCHBOARD: Telephone speech corpus for research and development. In: Proceedings of ICASSP.

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

[30] Goldwater, S., Jurafsky, D., Manning, C. D., 2010. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. Speech Communication 52 (3), 181–200.

[31] Gowdy, J. N., Subramanya, A., Bartels, C., Bilmes, J., 2004. DBN-based multi-stream models for audio-visual speech recognition. In: Proceedings of ICASSP.
[32] Gravier, G., Potamianos, G., Neti, C., 2002. Asynchrony modeling for audio-visual speech recognition. In: Proceedings of Human Language Technology Conf.
[33] Greenberg, S., 1999. Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation. Speech Communication 29 (2), 159–176.
[34] Greenberg, S., Hollenback, J., Ellis, D., 1996. Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. In: International Conference on Spoken Language Processing.
[35] Hasegawa-Johnson, M., Baker, J., Borys, S., Chen, K., Coogan, E., Greenberg, S., Juneja, A., Kirchoff, K., Livescu, K., Mohan, S., Muller, J., Sonmez, K., Wang, T., 2005. Landmark-based speech recognition: Report of the 2004 John Hopkins summer workshop. In: Proceedings of ICASSP.
[36] Hasegawa-Johnson, M., Livescu, K., Lal, P., Saenko, K., 2007. Audiovisual speech recognition with articulator positions as hidden variables. In: Proc. International Congress of Phonetic Sciences (ICPhS).
[37] Hazen, T. J., 2006. Visual model structures and synchrony constraints for audio-visual speech recognition. IEEE Trans. on Audio, Speech and Language Processing 14 (3), 1082–1089.
[38] Hazen, T. J., Hetherington, I. L., Shu, H., Livescu, K., 2005. Pronunciation modeling using a finite-state transducer representation. Speech Communication 46 (2), 189–203.
[39] Heffner, R. M. S., 1950. General Phonetics. Foundations of Modern Linguistics. The University of Wisconsin Press, Madison, WI.
[40] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., Kingsbury, B., 2012. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine 29 (6), 82–97.
[41] Huckvale, M. A., 1994. Word recognition from tiered phonological models. In: Proc. Inst. of Acoustics Conf. Speech and Hearing.
[42] Hutchinson, B., Droppo, J., 2011. Learning non-parametric models of pronunciation. In: Proceedings of ICASSP.
[43] Johnson, K., 2002. Abstract, Speech Communication Group Seminar.
[44] Johnson, K., 2004. Massive reduction in conversational American English. In: Spontaneous speech: Data and analysis. Proc. 1^{st} Session of the 10th Int'l Symp. pp. 29–54.
[45] Jurafsky, D., Ward, W., Jianping, Z., Herold, K., Xiuyang, Y., Sen, Z., 2001. What kind of pronunciation variation is hard for triphones to model? In: Proceedings of ICASSP.
[46] Jyothi, P., 2013. Discriminative and articulatory feature-based pronunciation models for conversational speech recognition. PhD dissertation, The Ohio State University, Columbus, OH.
[47] Jyothi, P., Fosler-Lussier, E., Livescu, K., 2012. Discriminatively training factorized finite state pronunciation models from dynamic Bayesian networks. In: Proceedings of Interspeech.
[48] Jyothi, P., Livescu, K., 2014. Revisiting word neighborhoods for speech recognition. In: Proc. ACL MORPHFSM Workshop.
[49] Jyothi, P., Livescu, K., Fosler-Lussier, E., 2011. Lexical access experiments with context-dependent articulatory feature-based models. In: Proceedings of ICASSP.
[50] Kaisse, E. M., 1985. Connected Speech. Academic Press, Orlando, FL.
[51] King, S., Frankel, J., Livescu, K., McDermott, E., Richmond, K., Wester, M., 2007. Speech production knowledge in automatic speech recognition. Journal of the Acoustical Society of America 121 (2), 723–742.
[52] Kirchhoff, K., 1996. Syllable-level desynchronisation of phonetic features for speech recognition. In: International Conference on Spoken Language Processing.
[53] Kirchhoff, K., Fink, G. A., Sagerer, G., 2002. Combining acoustic and articulatory feature information for robust speech recognition. Speech Communication 37 (3), 303–319.
[54] Ladefoged, P., 2001. A Course in Phonetics, 4th Edition. Harcourt, Brace, Jovanovich.
[55] Lamel, L., Adda, G., 1996. On designing pronunciation lexicons for large vocabulary continuous speech recognition. In: International Conference on Spoken Language Processing.
[56] Livescu, K., 2005. Feature-based pronunciation modeling for automatic speech recognition. PhD dissertation, MIT, Cambridge, MA.
[57] Livescu, K., Bezman, A., Borges, M., Yung, L., Cetin, O., Frankel, J., King, S., Xhi, X., Lavoie, L., 2007. Manual transcription of conversational speech at the articulatory feature level. In: Proceedings of ICASSP.
[58] Livescu, K., Cetin, O., Hasegawa-Johnson, M., King, S., Bartels, C., Borges, N., Kantor, A., Lal, P., Yung, L., Bezman, A., Dawson-Haggerty, S., Woods, B., Frankel, J., Magimai-Doss, M., Saenko, K., 2007. Articulatory feature-based methods for acoustic and audio-visual speech recognition: Summary from the 2006 JHU Summer Workshop. In: Proceedings of ICASSP.
[59] Livescu, K., Fosler-Lussier, E., Metze, F., 2012. Subword modeling for automatic speech recognition: Past, present, and emerging approaches. IEEE Signal Processing Magazine 29 (6), 44–57.
[60] Livescu, K., Glass, J. R., 2004. Feature-based pronunciation modeling for automatic speech recognition. In: Proceedings of HLT-NAACL.
[61] Livescu, K., Glass, J. R., 2004. Feature-based pronunciation modeling with trainable asynchrony probabilities. In: International Conference on Spoken Language Processing.
[62] Markov, K., Dang, J., Nakamura, S., 2006. Integration of articulatory and spectrum features based on the hybrid HMM/BN modeling framework. Speech Communication 48 (2), 161–175.
[63] McAllaster, D., Gillick, L., Scattone, F., Newman, M., 1998. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. In: International Conference on Spoken Language Processing.
[64] McGraw, I., Badr, I., Glass, J. R., 2013. Learning lexicons from speech using a pronunciation mixture model. IEEE Trans. on Audio, Speech and Language Processing 21 (2), 357–366.
[65] Metze, F., Waibel, A., 2002. A flexible stream architecture for ASR using articulatory features. In: International Conference on Spoken Language Processing.
[66] Mitra, V., Nam, H., Espy-Wilson, C., Saltzman, E., Goldstein, L., 2012. Recognizing articulatory gestures from speech for robust speech

recognition. Journal of the Acoustical Society of America 131 (3), 2270–2287.
[67] Morris, J., Fosler-Lussier, E., 2008. Conditional random fields for integrating local discriminative classifiers. IEEE Trans. on Audio, Speech and Language Processing 16 (3), 617–628.
[68] Nam, H., Goldstein, L., Saltzman, E., Byrd, D., 2004. Tada: An enhanced, portable task dynamics model in MATLAB. Journal of the Acoustical Society of America 115 (5), 2430–2430.
[69] Nefian, A. V., Liang, L., Pi, X., Xiaoxiang, L., Mao, C., Murphy, K., 2002. A coupled HMM for audio-visual speech recognition. In: Proceedings of ICASSP.
[70] Nock, H. J., Ostendorf, M., 2003. Parameter reduction schemes for loosely coupled HMMs. Computer Speech & Language 17 (2), 233–262.
[71] Oshika, B., Zue, V., Weeks, R., Neu, H., Aurbach, J., 1975. The role of phonological rules in speech understanding research. IEEE Trans. on Acoustics, Speech and Signal Proc. 23 (1), 104–112.
[72] Ostendorf, M., 1999. Moving beyond the 'beads-on-a-string' model of speech. In: Proceedings of IEEE ASRU Workshop.
[73] Ostendorf, M., Byrne, B., Bacchiani, M., Finke, M., Gunawardana, A., Ross, K., Roweis, S., Shriberg, E., Talkin, D., Waibel, A., Wheatley, B., Zeppenfeld, T., 1996. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In: International Conference on Spoken Language Processing. (supplementary paper).
[74] Perkell, J. S., 1969. Physiology of speech production: results and implications of a quantitative cineradiographic study. MIT Press, Cambridge, MA.
[75] Pernkopf, F., Bilmes, J., 2005. Discriminative versus generative parameter and structure learning of Bayesian network classifiers. In: Proceedings of International Conference on Machine Learning. ACM, pp. 657–664.
[76] Prabhavalkar, R., 2013. Discriminative articulatory feature-based pronunciation models with application to spoken term detection. PhD dissertation, The Ohio State University, Columbus, OH.
[77] Prabhavalkar, R., Fosler-Lussier, E., Livescu, K., 2011. A factored conditional random field model for articulatory feature forced transcription. In: Proceedings of IEEE ASRU Workshop.
[78] Prabhavalkar, R., Livescu, K., Fosler-Lussier, E., Keshet, J., 2013. Discriminative articulatory models for spoken term detection in low-resource conversational settings. In: Proceedings of ICASSP.
[79] Richardson, M., Bilmes, J., Diorio, C., 2003. Hidden-articulator Markov models for speech recognition. Speech Communication 41 (2), 511–529.
[80] Riley, M., Byrne, W., Finke, M., Khudanpur, S., Ljolje, A., McDonough, J., Nock, H., Saraçlar, M., Wooters, C., Zavaliagkos, G., 1999. Stochastic pronunciation modelling from hand-labelled phonetic corpora. Speech Communication 29 (2–4), 209–224.
[81] Riley, M. D., Ljolje, A., 1996. Automatic generation of detailed pronunciation lexicons. In: Lee, C.-H., Soong, F. K., Paliwal, K. K. (Eds.), Automatic Speech and Speaker Recognition. Kluwer Academic Publishers, Boston, pp. 285–302.
[82] Saenko, K., Livescu, K., Glass, J., Darrell, T., 2009. Multistream articulatory feature-based models for visual speech recognition. IEEE Trans. on Pattern Analysis and Machine Intelligence 31 (9), 1700–1707.
[83] Saraçlar, M., Khudanpur, S., Oct. 2004. Pronunciation change in conversational speech and its implications for automatic speech recognition. Computer Speech and Language 18 (4), 375–395.
[84] Schane, S. A., 1973. Generative Phonology. Foundations of Modern Linguistics. Prentice-Hall, Inc., Englewood Cliffs, NJ.
[85] Shoup, J. E., 1980. Phonological aspects of speech recognition. In: Trends in Speech Recognition. Prentice Hall, Englewood Cliffs, NJ, pp. 125–138.
[86] Shu, H., Hetherington, I. L., 2002. EM training of finite-state transducers and its application to pronunciation modeling. In: International Conference on Spoken Language Processing.
[87] Siniscalchi, S. M., Yu, D., Deng, L., Lee, C.-H., 2013. Exploiting deep neural networks for detection-based speech recognition. Neurocomputing 106, 148–157.
[88] Sloboda, T., Waibel, A., 1996. Dictionary learning for spontaneous speech recognition. In: International Conference on Spoken Language Processing.
[89] Stephenson, T. A., Magimai-Doss, M., Bourlard, H., 2004. Speech recognition with auxiliary information. IEEE Trans. on Audio, Speech and Language Processing 12 (3), 189–203.
[90] Stevens, K. N., 1998. Acoustic Phonetics. MIT Press, Cambridge, MA.
[91] Strik, H., Cucchiarini, C., 1999. Modeling pronunciation variation for ASR: A survey of the literature. Speech Communication 29 (2–4), 225–246.
[92] Tang, H., Keshet, J., Livescu, K., 2012. Discriminative pronunciation modeling: A large-margin, feature-rich approach. In: Proceedings of ACL.
[93] Terry, L., Livescu, K., Pierrehumbert, J., Katsaggelos, A., 2010. Audio-visual anticipatory coarticulation modeling by human and machine. In: Proceedings of Interspeech.
[94] Weintraub, M., Taussig, K., Hunicke-Smith, K., Snodgrass, A., 1996. Effect of speaking style on LVCSR performance. In: International Conference on Spoken Language Processing.
[95] Weintraub, M., Wegmann, S., Kao, Y.-H., Khudanpur, S., Galles, C., Fosler, E., Saraçlar, M., Oct. 1996. Automatic learning of word pronunciation from data. In: Proc. ICSLP. Philadelphia, PA.
[96] Wester, M., Frankel, J., King, S., 2004. Asynchronous articulatory feature recognition using dynamic bayesian networks. In: Proc. IEICI Beyond HMM Workshop.
[97] Wester, M., Frankel, J., King, S., 2004. Asynchronous articulatory feature recognition using dynamic Bayesian networks. In: Proceedings of IEICI Beyond HMM Workshop.
[98] Wiebe, B., 1992. Modelling autosegmental phonology with multi-tape finite state transducers. MS dissertation, Simon Frasier University.
[99] Yoo, I. W., Blankenship, B., 2003. Duration of epenthetic [t] in polysyllabic American English words. Journal of the International Phonetic Association 33 (2), 153–164.
[100] Zhang, L., 2004. Maximum entropy modeling toolkit. `http://homepages.inf.ed.ac.uk/lzhang10/maxent_toolkit.html`.
[101] Zhang, Y., Diao, Q., Huang, S., Hu, W., Bartels, C. D., Bilmes, J., 2003. DBN based multi-stream models for speech. In: Proceedings of

ICASSP.
[102] Zweig, G., 1998. Speech recognition using dynamic Bayesian networks. PhD dissertation, U. C. Berkeley, Berkeley, CA.
[103] Zweig, G., Nguyen, P., 2009. A segmental CRF approach to large vocabulary continuous speech recognition. In: Proceedings of IEEE ASRU Workshop.