Feature-Based Pronunciation Modeling for Automatic Speech Recognition

by

Karen Livescu
S.M., Massachusetts Institute of Technology (1999)
A.B., Princeton University (1996)

Submitted to the Department of
Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2005

Author .....
Department of
Electrical Engineering and Computer Science
August 31, 2005

Certified by .....
James R. Glass
Principal Research Scientist
Thesis Supervisor

Accepted by .....
Arthur C. Smith
Chairman, Department Committee on Graduate Students

Feature-Based Pronunciation Modeling for Automatic Speech Recognition

by

Karen Livescu
S.M., Massachusetts Institute of Technology (1999)
A.B., Princeton University (1996)

Submitted to the Department of
Electrical Engineering and Computer Science
on August 31, 2005, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Electrical Engineering and Computer Science

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. One approach to handling this variation consists of expanding the dictionary with phonetic substitution, insertion, and deletion rules. Common rule sets, however, typically leave many pronunciation variants unaccounted for and increase word confusability due to the coarse granularity of phone units.

We present an alternative approach, in which many types of variation are explained by representing a pronunciation as multiple streams of linguistic 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 to acoustic or perceptual categories. By allowing for asynchrony between features and per-feature substitutions, many pronunciation changes that are difficult to account for with phone-based models become quite natural. Although it is well-known that many phenomena can be attributed to this "semi-independent evolution" of features, previous models of pronunciation variation have typically not taken advantage of this.

In particular, we propose a class of feature-based pronunciation models represented as dynamic Bayesian networks (DBNs). The DBN framework allows us to naturally represent the factorization of the state space of feature combinations into feature-specific factors, as well as providing standard algorithms for inference and parameter learning. We investigate the behavior of such a model in isolation using manually transcribed words. Compared to a phone-based baseline, the feature-based model has both higher coverage of observed pronunciations and higher recognition rate for isolated words. We also discuss the ways in which such a model can be incorporated into various types of end-to-end speech recognizers and present several examples of implemented systems, for both acoustic speech recognition and lipreading tasks.

Thesis Supervisor: James R. Glass
Title: Principal Research Scientist

Acknowledgments

I would like to thank my advisor, Jim Glass, for his guidance throughout the past few years, for his seemingly infinite patience as I waded through various disciplines and ideas, and for allowing me the freedom to work on this somewhat unusual topic. I am also grateful to the other members of my thesis committee, Victor Zue, Jeff Bilmes, and Tommi Jaakkola. I thank Victor for his insightful and challenging questions, for always keeping the big picture in mind, and for all of his meta-advice and support throughout my time in graduate school. I am grateful to Tommi for his comments and questions, and for helping me to keep the "non-speech audience" in mind.

It is fair to say that this thesis would have been impossible without Jeff's support throughout the past few years. Many of the main ideas can be traced back to conversations with Jeff at the 2001 Johns Hopkins Summer Workshop, and since then he has been a frequent source of feedback and suggestions. The experimental work has depended crucially on the Graphical Models Toolkit, for which Jeff has provided invaluable support (and well-timed new features). I thank him for giving generously of his time, ideas, advice, and friendship.

I was extremely fortunate to participate in the 2001 and 2004 summer workshops of the Johns Hopkins Center for Language and Speech Processing, in the projects on Discriminatively Structured Graphical Models for Speech Recognition, led by Jeff Bilmes and Geoff Zweig, and Landmark-Based Speech Recognition, led by Mark Hasegawa-Johnson. The first of these resulted in my thesis topic; the second allowed me to incorporate the ideas in this thesis into an ambitious system (and resulted in Section 5.2 of this document). I am grateful to Geoff Zweig for inviting me to participate in the 2001 workshop, for his work on and assistance with the first version of GMTK, and for his advice and inspiration to pursue this thesis topic; and to Jeff and Geoff for making the project a fun and productive experience. I thank Mark Hasegawa-Johnson for inviting me to participate in the 2004 workshop, and for many conversations that have shaped my thinking and ideas for future work. The excellent teams for both workshop projects made these summers even more rewarding; in particular, my work has benefitted from interactions with Peng Xu, Karim Filali, and Thomas Richardson in the 2001 team and Katrin Kirchhoff, Amit Juneja, Kemal Sonmez, Jim Baker, and Steven Greenberg in 2004. I am indebted to Fred Jelinek and Sanjeev Khudanpur of CLSP for making these workshops an extremely rewarding way to spend a summer, and for allowing me the opportunity to participate in both projects. The members of the other workshop teams and the researchers, students, and administrative staff of CLSP further conspired to make these two summers stimulating, fun, and welcoming. In particular, I have benefitted from interactions with John Blitzer, Jason Eisner, Dan Jurafsky, Richard Sproat, Izhak Shafran, Shankar Kumar, and Brock Pytlik. Finally, I would like to thank the city of Baltimore for providing the kind of weather that makes one want to stay in lab all day, and for all the crab.

The lipreading experiments of Chapter 6 were done in close collaboration with Kate Saenko and under the guidance of Trevor Darrell. The audio-visual speech recognition ideas described in Chapter 7 are also a product of this collaboration.

This work was a fully joint effort. I am grateful to Kate for suggesting that we work on this task, for contributing the vision side of the effort, for making me think harder about various issues in my work, and generally for being a fun person to talk to about work and about life. I also thank Trevor for his support, guidance, and suggestions throughout this collaboration, and for his helpful insights on theses and careers.

In working on the interdisciplinary ideas in this thesis, I have benefitted from contacts with members of the linguistics department and the Speech Communication group at MIT. Thanks to Janet Slifka and Stefanie Shattuck-Hufnagel for organizing a discussion group on articulatory phonology, and to Ken Stevens and Joe Perkell for their comments and suggestions. Thanks to Donca Steriade, Edward Flemming, and Adam Albright for allowing me to sit in on their courses and for helpful literature pointers and answers to my questions. In particular, I thank Donca for several meetings that helped to acquaint me with some of the relevant linguistics ideas and literature.

In the summer of 2003, I visited the Signal, Speech and Language Interpretation Lab at the University of Washington. I am grateful to Jeff Bilmes for hosting me, and to several SSLI lab members for helpful interactions during and outside this visit: Katrin Kirchhoff, for advice on everything feature-related, and Chris Bartels, Karim Filali, and Alex Norman for GMTK assistance.

Outside of the direct line of my thesis work, my research has been enriched by summer internships. In the summer of 2000, I worked at IBM Research, under the guidance of George Saon, Mukund Padmanabhan, and Michael Picheny. I am grateful to them and to other researchers in the IBM speech group—in particular, Geoff Zweig, Brian Kingsbury, Lidia Mangu, Stan Chen, and Mirek Novak—for making this a pleasant way to get acquainted with industrial speech recognition research.

My first speech-related research experience was as a summer intern at AT&T Bell Labs, as an undergraduate in the summer of 1995, working with Richard Sproat and Chilin Shih on speech synthesis. The positive experience of this internship, and especially Richard's encouragement, is probably the main reason I chose to pursue speech technology research in graduate school, and I am extremely thankful for that.

I am grateful to everyone in the Spoken Language Systems group for all of the ways they have made life as a graduate student more pleasant. TJ Hazen, Lee Hetherington, Chao Wang, and Stefanie Seneff have all provided generous research advice and assistance at various points throughout my graduate career. Thanks to Lee and Scott Cyphers for frequent help with the computing infrastructure (on which my experiments put a disproportionate strain), and to Marcia Davidson for keeping the administrative side of SLS running smoothly. Thanks to all of the SLS students and visitors, past and present, for making SLS a lively place to be. A very special thanks to Issam Bazzi and Han Shu for always being willing to talk about speech recognition, Middle East peace, and life. Thanks to Jon Yi, Alex Park, Ernie Pusateri, John Lee, Ken Schutte, Min Tang, and Ghinwa Choueiter for exchanging research ideas and for answering (and asking) many questions over the years. Thanks also to my officemates Ed and Mitch for making the office a pleasant place to be.

Thanks to Nati Srebro for the years of support and friendship, for fielding my machine learning and graphical models questions, and for his detailed comments on

parts of this thesis.

Thanks to Marilyn Pierce and to the rest of the staff of the EECS Graduate Office, for making the rules and requirements of the department seem not so onerous.

Thanks to my family in Israel for their support over the years, and for asking "nu....?" every once in a while. Thanks to Greg, for everything. And thanks to my parents, whom I can't possibly thank in words.

Contents

List of Figures

List of Tables

Nomenclature

AF articulatory feature

ANN artificial neural network

AP articulatory phonology

ASR automatic speech recognition

CPT conditional probability table

DBN dynamic Bayesian network

DCT discrete cosine transform

DF distinctive feature

EM expectation-maximization

GLOTTIS, G glottis opening degree

HMM hidden Markov model

ICSI International Computer Science Institute, UC Berkeley

IPA International Phonetic Alphabet

LIP-LOC, LL lip constriction location

LIP-OPEN, LO lip opening degree

MFCC Mel-frequency cepstral coefficient

PCA principal components analysis

SVM support vector machine

TB-LOC, TBL tongue body constriction location

TB-OPEN, TBO tongue body opening degree

TT-LOC, TTL tongue tip constriction location

TT-OPEN, TTO tongue tip opening degree

VELUM, V velum opening degree

Chapter 1

Introduction

Human speech is characterized by a great deal of variability. Two utterances of the same string of words may produce speech signals that, on arrival at a listener's ear, may differ in a number of respects:

- **Pronunciation**, or the speech sounds that make up each word. 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 (i) speaking style—the same words may be pronounced carefully and clearly when reading but more sloppily in conversational or fast speech; and (ii) the surrounding words—*green beans* may be pronounced "greem beans".

- **Prosody**, or the choice of amplitudes, pitches, and durations of different parts of the utterance. This can give the same sentence different meanings or emphases, and may drastically affect the signal.

- **Speaker-dependent acoustic variation**, or "production noise", due to the speakers' differing vocal tracts and emotional or physical states.

- **Channel and environment effects**. The same utterance may produce one signal at the ear of a listener who is in the same room as the speaker, another signal in the next room, yet other signals for a listener on a land-line or cellular phone, and yet another for a listener who happens to be underwater. In addition, there may be interfering signals in the acoustic environment, such as noise or crosstalk.

The characterization of this variability, and the search for invariant aspects of speech, is a major organizing principle of research in speech science and technology (see, e.g., [PK86, JM97]). Automatic speech recognition (ASR) systems must account for each of these types of variability in some way. This thesis is concerned with variability in pronunciation, and in particular speaker-independent variability. This has been identified in several studies [MGSN98, WTHSS96, FL99] as a main factor in the poor performance of automatic speech recognizers on conversational speech,

which is characterized by a larger degree of variability than read speech. Fosler-Lussier [FL99] found that words pronounced non-canonically, according to a manual transcription, are more likely than canonical productions to be deleted or substituted by an automatic speech recognizer. Weintraub et al. [WTHSS96] 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 speech than for the same sentences read by the same speakers in a dictation style. McAllaster and Gillick [MGSN98] generated synthetic speech with pronunciations matching the canonical dictionary forms, and found that it can be recognized with extremely low error rates of around 5%, compared with around 40% for synthetic speech with the pronunciations observed in actual conversational data, and 47% for real conversational speech.

Efforts to model pronunciation variability in ASR systems have often resulted in performance gains, but of a much smaller magnitude than these analyses would suggest (e.g., [RBF+99, WWK+96, SC99, SK04, HHSL05]). In this thesis, we propose a new way of handling this variability, based on modeling the behavior of multiple streams of linguistic *features* rather than the traditional single stream of phones. We now describe the main motivations for such an approach through a brief survey of related research and examples of pronunciation data. We will then outline the proposed approach, the contributions of the thesis, and the remaining chapters.

1.1 Motivations

We are motivated in this work by a combination of (i) the limitations of existing ASR pronunciation models in accounting for pronunciations observed in speech data, (ii) the emergence of feature-based acoustic observation models for ASR with no corresponding pronunciation models, and (iii) recent work in linguistics and speech science that supersedes the linguistic bases for current ASR systems. We describe these in turn, after covering a few preliminaries regarding terminology.

1.1.1 Preliminaries

The term *pronunciation* is a vague one, lacking a standard definition (see, e.g., efforts to define it in [SC99]). For our purposes, we define a *pronunciation* of a word as a representation, in terms of some set of linguistically meaningful sub-word units, of the way the word is or can be produced by a speaker. By a linguistically meaningful representation, we mean one that can in principle differentiate between words: Acoustic variations in the signal caused by the environment or the speaker's vocal tract characteristics are not considered linguistically meaningful; degrees of aspiration of a stop consonant may be. Following convention in linguistics and speech research, we distinguish between a word's (i) *underlying* (or *target* or *canonical*) pronunciations, the ones typically found in an English dictionary, and its (ii) *surface* pronunciations, the ways in which a speaker may actually produce the word. Underlying pronunciations are typically represented as strings of *phonemes*, the basic sub-word units distinguishing

words in a language. For example, the underlying pronunciations for the four words *sense, probably, everybody*, and *don't* might be written[1]

- *sense* → /s eh n s/
- *probably* → /p r aa b ax b l iy/
- *everybody* → /eh v r iy b ah d iy/
- *don't* → /d ow n t/

Here and throughout, we use a modified form of the ARPABET phonetic alphabet [Sho80], described in Appendix A, and use the linguistic convention that phoneme strings are enclosed in "/ /".

While dictionaries usually list one or a few underlying pronunciations for a given word, the same word may have dozens of surface pronunciations. Surface pronunciations are typically represented as strings of *phones*, usually a somewhat more detailed label set, enclosed in square brackets ("[ ]") by convention. Table 1.1 shows all of the surface pronunciations of the above four words that were observed in a set of phonetically-transcribed conversational speech. [2]

1.1.2 The challenge of pronunciation variation

The pronunciations in Table 1.1 are drawn from a set of recorded and manually phonetically transcribed American English conversations consisting of approximately 10,000 spoken word tokens [GHE96]. The exact transcriptions of spoken pronunciations are, to some extent, subjective.[3] However, there are a few clear aspects of the data in Table 1.1 that are worthy of mention:

- There is a large number of pronunciations per word, with most pronunciations 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*.

[1] We note that not all dictionaries agree on the pronunciations of these words. For example, Merriam-Webster's Online Dictionary [M-W] lists the pronunciations for *sense* as /s eh n s/ and /s eh n t s/, and for *probably* as /p r aa b ax b l iy/ and /p r aa (b) b l iy/ (the latter indicating that there may optionally be two /b/s in a row). This appears to be unusual, however: None of the Oxford English Dictionary, Random House Unabridged Dictionary, and American Heritage Dictionary list the latter pronunciations [SW89, RHD87, AHD00].

[2] These phonetic transcriptions are drawn from the phonetically transcribed portion of the Switchboard corpus, described in Chapter 4. The surface pronunciations 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.

[3] 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." [Joh02]

	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
observed	(1) s eh n t s (1) s ih t s	(2) p r aa b iy (1) 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 ow ih (1) p aa iy (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	(37) d ow n (16) d ow (6) ow n (4) d ow n t (3) d ow t (3) d ah n (3) ow (3) n ax (2) d ax n (2) ax (1) 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

Table 1.1: *Canonical and observed pronunciations of four words in the phonetically transcribed portion of the Switchboard database [GHE96]. The number of times each observed pronunciation appears in the database is given in parentheses. Single-character labels are pronounced like the corresponding English letters; the remaining labels are: [ax], as in the beginning of* ***a****bout; [aa], as in* f***a***ther*; [ay], as in* b***ye****; [ah], as in* m***u***d*; [ao], as in* ***aw****e; [el], as in* bott***le****; [ow], as in* l***ow****; [dh], as in* ***th****is; [uh], as in* b***oo***k*; [ih], as in* b***i***d*; [iy], as in* b***e****; [er], as in* b***ir***d*; [ux], as in* t***oo***t*; and [uw], as in* b***oo***m*. See Appendix A for further information on the phonetic alphabet.*

- 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.[4] We will describe and analyze this

[4]After dropping some diacritics and collapsing similar phone labels. This reduced the phonetic label set from 396 to 179 distinct labels. Before collapsing the labels, the mean number of distinct

database further in Chapter 4.

Humans seem to be able to recognize these words in all of their many manifestations. How can an automatic speech recognizer know which are the legal pronunciations for a given word? For a sufficiently small vocabulary, we can imagine recording a database large enough to obtain reliable estimates of the distributions of word pronunciations. In fact, for a very small vocabulary, say tens of words, we may dispense with sub-word units entirely, instead modeling directly the signals corresponding to entire words. This is the approach used in most small-vocabulary ASR systems such as digit recognizers (e.g., [HP00]). However, for larger vocabularies, this is infeasible, especially if we wish to record naturally occurring speech rather than read scripts: In order to obtain sufficient statistics for rare words, the database may be prohibitively large. The standard approach is therefore to represent words in terms of smaller units and to model the distributions of signals corresponding to those units. The problem therefore remains of how to discover the possible pronunciations for each word.

1.1.3 Previous work: 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 [SW96]. 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 reasonably-sized vocabulary and reasonably-sized database, most words in the vocabulary will only occur a handful of times, and many will not occur at all. Consider the Switchboard database of conversational speech [GHM92], from which the above examples are drawn, which is often considered the standard database for large-vocabulary conversational ASR. The database contains over 300 hours of speech, consisting of about 3,000,000 spoken words covering a vocabulary of 29,695 words. Of these 29,695 words, 18,504 occur fewer than five times. The prospects for robustly estimating the probabilities of most words' pronunciations are therefore dim.

However, if we look at a variety of pronunciation data, we notice that many of the 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:

- *defense* → [d ih f eh n t s]
- *prince* → [p r ih n t s]
- *insight* → [ih n t s ay t]
- *expensive* → [eh k s p eh n t s ih v]

pronunciations for words spoken at least five times is 11.0.

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." There are in fact many pronunciation phenomena that are well-described 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. Such rules have been documented in the linguistics, speech science, and speech technology literature (e.g., [Hef50, Sch73, Lad01, Kai85, OZW+75]) and are the basis for another approach that has been used in ASR research for pronunciation modeling: One or a few main pronunciations are listed for each word, and a bank of rewrite rules are used to generate additional pronunciations. The rules can be pre-specified based on linguistic knowledge [HHSL05], or they may be learned from data [FW97]. The probability of each rule "firing" can also be learned from data [SH02]. A related approach is to learn, for each phoneme, a decision tree that predicts the phoneme's surface pronunciation depending on context [RL96].

This approach greatly 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 it does not alleviate it entirely; there are many possible phonetic sequences to consider, and many of them occur very rarely. When such rules are learned from data, therefore, it is still common practice to exclude rarely observed sequences. As we will show in Chapter 4, it is difficult to account for the variety of pronunciations seen in conversational speech with phonetic rewrite rules.

The issue of confusability can also 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 would be

- s eh_n n t s
- s ih_n t s

indicating that the two vowels were *nasalized*. Similarly, *don't* $\rightarrow$ [d ow t] is more finely transcribed [d ow_n t]. Vowel nasalization, in which there is airflow through both the mouth and the nasal cavity, often occurs before nasal consonants (/m/, /n/, and /ng/). 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*.

1.1.4 Feature-based representations

The presence of [t] in the two examples of *sense* might seem a bit mysterious until we consider the mechanism by which it comes about. In order to produce an [n], the

speaker must make a closure with the tongue tip just behind the top teeth, as well as lower the soft palate to allow air to flow to the nasal cavity. To produce the following [s], the tongue closure is slightly released and voicing and nasality are turned off. If these tasks are not done synchronously, new sounds may emerge. In this case, voicing and nasality are turned off before the tongue closure is released, resulting 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].

This observation motivates 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 1.2 and 1.3 show such a representation of the canonical pronunciation of *sense* and of the observed pronunciation [s eh_n n t s], along with the corresponding phonetic string. Deviations from the canonical values are marked (*). The feature set is described more fully in Chapter 3 and in Appendix B. Comparing each row of the canonical and observed pronunciations, we see that all of the feature values are produced faithfully, but with some asynchrony in the timing of feature changes.

Table 1.4 shows a feature-based representation of the second example, [s ih_n t s]. Here again, most of the 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 [p] insertion in words like *warmth* → [w ao r m p th].

In addition, the feature-based representation allows 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 a "true" [t]. This, in fact, appears to be the case in at least some contexts [YB03]. 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 inserted [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. This is supported by Saraçlar and Khudanpur [SK04], who show that pronunciation change typically does not result in an entirely new phone but one that is intermediate in some way to the canonical phone and another phone. The feature-based representation makes it possible to have such a fine-grained representation, without the explosion in training data that would normally be required to train a phone-based pronunciation model with a finer-grained phone set. This also suggests that pronunciation models should be sensitive to timing information.

feature	values			
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 1.2: *Canonical pronunciation of* sense *in terms of articulatory features.*

feature	values				
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 1.3: *Observed pronunciation #1 of* sense *in terms of articulatory features.*

feature	values			
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 1.4: *Observed pronunciation #2 of* sense *in terms of articulatory features.*

1.1.5 Previous work: Acoustic observation modeling

Another motivation for this thesis is provided by recent work suggesting that, for independent reasons, it may be useful to use sub-phonetic features in the *acoustic observation* modeling component of ASR systems [KFS00, MW02, Eid01, FK01]. Reasons cited include the potential for improved pronunciation modeling, but also better use of training data—there will typically be vastly more examples of each feature than of each phone—better performance in noise [KFS00], and generalization to multiple languages [SMSW03].

One issue with this approach is that it now becomes necessary to define a mapping from words to features. Typically, feature-based systems simply convert a phone-based dictionary to a feature-based one using a phone-to-feature mapping, limiting the features to their canonical values and forcing them to proceed synchronously in

phone-sized "bundles". When features stray from their canonical values or evolve asynchronously, there is a mismatch with the dictionary. These approaches have therefore benefited from the advantages of features with respect to data utilization and noise robustness, but may not have reached their full potential as a result of this mismatch. There is a need, therefore, for a mechanism to accurately represent the evolution of features as they occur in the signal.

1.1.6 Previous work: Linguistics/speech research

A final motivation is that the representations of pronunciation used in most current ASR systems are based on outdated linguistics. The paradigm of a string of phonemes plus rewrite rules is characteristic of the generative phonology of the 1960s and 1970s (e.g., [CH68]). More recent linguistic theories, under the general heading of non-linear or autosegmental phonology [Gol90], have done away with the single-string representation, opting instead for multiple tiers of features. The theory of articulatory phonology [BG92] posits that most or all surface variation results from the relative timings of articulatory gestures, using a representation similar to that of Tables 1.2–1.4. Articulatory phonology is a work in progress, although one that we will draw some ideas from. However, the principle in non-linear phonology of using multiple streams of representation for different aspects of speech is now standard practice.

1.2 Proposed approach

Motivated by these observations, this thesis proposes a probabilistic approach to pronunciation modeling based on representing the time course of multiple streams of linguistic features. In this model, features may stray from the canonical representation in two ways:

- **Asynchrony**, in which different features proceed through their trajectories at different rates.
- **Substitution** of values of individual features.

Unlike in phone-based models, we will not make use of deletions or insertions of features, instead accounting for apparent phone insertions or deletions as resulting from feature asynchrony or substitution. The model is defined probabilistically, enabling fine tuning of the degrees and types of asynchrony and substitution and allowing these to be learned from data. We formalize the model as a dynamic Bayesian network (DBN), a generalization of hidden Markov models that allows for natural and parsimonious representations of multi-stream models.

1.3 Contributions

The main contributions of this thesis are:

- Introduction of a feature-based model for pronunciation variation, formalizing some aspects of current linguistic theories and addressing limitations of phone-based models.
- Investigation of this model, along with a feature set based on articulatory phonology, in a lexical access task using manual transcriptions of conversational speech. In these experiments, we show that the proposed model outperforms a phone-based one in terms of coverage of observed pronunciations and ability to retrieve the correct word.
- Demonstration of the model's use in several end-to-end recognition systems for both acoustic speech recognition and lipreading applications.

1.4 Thesis outline

The remainder of the thesis is structured as follows. In Chapter 2, we describe the relevant background: the prevailing generative approach to ASR (which we follow), several threads of previous research, and related work in linguistics and speech science. Chapter 3 describes the proposed model, its implementation as a dynamic Bayesian network, and several ways in which it can be incorporated into a complete ASR system. Chapter 4 presents experiments done to test the pronunciation model in isolation, by recognizing individual words excised from conversational speech based on their detailed manual transcriptions. Chapter 5 describes the use of the model in two types of acoustic speech recognition systems. Chapter 6 describes how the model can be applied to lipreading and presents results showing improved performance using a feature-based model over a more traditional viseme-based one. Finally, Chapter 7 discusses future directions and conclusions.

Chapter 2

Background

This chapter provides some background on the statistical formulation of automatic speech recognition (ASR); the relevant linguistic concepts and principles; dynamic Bayesian networks and their use in ASR; and additional description of previous work beyond the discussion of Chapter 1.

2.1 Automatic speech recognition

In this thesis, we are concerned with pronunciation modeling not for its own sake, but in the specific context of automatic speech recognition. In particular, we will work within the prevailing *statistical, generative* formulation of ASR, described below. The history of ASR has seen both non-statistical approaches, such as the knowledge-based methods prevalent until around the mid-1970s (e.g., [Kla77]), and non-generative approaches, including most of the knowledge-based systems but also very recent non-generative statistical models [RSCJ04, GMAP05]. However, the most widely used approach, and the one we assume, is the generative statistical one. We will also assume that the task at hand is *continuous* speech recognition, that is, that we are interested in recognition of word strings rather than of isolated words. Although much of our experimental work is in fact isolated-word, we intend for our approach to apply to continuous speech recognition and formulate our presentation accordingly.

In the standard formulation [Jel98], the problem that a continuous speech recognizer attempts to solve is: For a given input speech signal $\mathbf{s}$, what is the most likely string of words $\mathbf{w}^* = \{w_1, w_2, \ldots, w_M\}$ that generated it? In other words,[1]

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{s}), \tag{2.1}$$

where $\mathbf{w}$ ranges over all possible word strings $\mathcal{W}$, and each word w_i is drawn from a finite vocabulary $\mathcal{V}$. Rather than using the raw signal $\mathbf{s}$ directly, we assume that all of the relevant information in the signal can be summarized in a set of *acoustic*

[1] We use the notation $p(x)$ to indicate either the probability mass function $P_X(x) = P(X = x)$ when X is discrete or the probability density function $f_X(x)$ when X is continuous.

observations[2] $\mathbf{o} = \{o_1, o_2, \ldots, o_T\}$, where each o_i is a vector of measurements computed over a short time frame, typically 5ms or 10ms long, and T is the number of such frames in the speech signal[3]. The task is now to find the most likely word string corresponding to the acoustic observations:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} p(\mathbf{w}|\mathbf{o}), \tag{2.2}$$

Using Bayes' rule of probability, we may rewrite this as

$$\begin{aligned} \mathbf{w}^* &= \arg\max_{\mathbf{w}} \frac{p(\mathbf{o}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{o})} && (2.3) \\ &= \arg\max_{\mathbf{w}} p(\mathbf{o}|\mathbf{w})p(\mathbf{w}), && (2.4) \end{aligned}$$

where the second equality arises because $\mathbf{o}$ is fixed and therefore $p(\mathbf{o})$ does not affect the maximization. The first term on the right-hand side of 2.4 is referred to as the *acoustic model* and the second term as the *language model*. $p(\mathbf{o}|\mathbf{w})$ is also referred to as the *likelihood* of the hypothesis $\mathbf{w}$.

2.1.1 The language model

For very restrictive domains (e.g., digit strings, command and control tasks), the language model can be represented as a finite-state or context-free grammar. For more complex tasks, the language model can be factored using the chain rule of probability:

$$p(\mathbf{w}) = \prod_{i=1}^{M} p(w_i|w_1, \ldots, w_{i-1}) \tag{2.5}$$

and it is typically assumed that, given the history of the previous $n-1$ words (for $n = 2, 3$, or perhaps 4), each word is independent of the remaining history. That is, the language model is an n-gram model:

$$p(\mathbf{w}) = \prod_{i=1}^{M} p(w_i|w_{i-n+1}, \ldots, w_{i-1}) \tag{2.6}$$

2.1.2 The acoustic model

For all but the smallest-vocabulary isolated-word recognition tasks, we cannot hope to model $p(\mathbf{o}|\mathbf{w})$ directly; there are too many possible $\mathbf{o}, \mathbf{w}$ combinations. In general, the acoustic model is further decomposed into multiple factors, most commonly using hidden Markov models. A hidden Markov model (HMM) is a modified finite-state

[2]These are often referred to as acoustic *features*, using the pattern recognition sense of the term. We prefer the term *observations* so as to not cause confusion with linguistic features.

[3]Alternatively, acoustic observations may be measured at non-uniform time points or over segments of varying size, as in segment-based speech recognition [Gla03]. Here we are working within the framework of frame-based recognition as it is more straightforward, although our approach should in principle be applicable to segment-based recognition as well.

machine in which states are not observable, but each state emits an observable output symbol with some distribution. An HMM is characterized by (a) a distribution over initial state occupancy, (b) probabilities of transitioning from a given state to each of the other states in a given time step, and (c) state-specific distributions over output symbols. The output "symbols" in the case of speech recognition are the (usually) continuous acoustic observation vectors, and the output distributions are typically mixtures of Gaussians. For a more in-depth discussion of HMMs, see [Jel98, RJ93].

For recognition of a limited set of phrases, as in command and control tasks, each allowable phrase can be represented as a separate HMM, typically with a chain-like state transition graph and with each state intended to represent a "steady" portion of the phrase. For example, a whole-phrase HMM may have as many states as phones in its baseform; more typically, about three times as many states are used, in order to account for the fact that the beginnings, centers, and ends of phone segments typically have different distributions. For somewhat less constrained tasks such as small-vocabulary continuous speech recognition, e.g. digit string recognition, there may be one HMM per word. To evaluate the acoustic probability for a given hypothesis $\mathbf{w} = \{w_1, \ldots, w_M\}$, the word HMMs for $w_1, \ldots, w_M$ can be concatenated to effectively construct an HMM for the entire word string. Finally, for larger vocabularies, it is infeasible to use whole-word models for all but the most common words, as there are typically insufficient training examples of most words; words are further broken down into sub-word units, most often phones, each of which is modeled with its own HMM. In order to account for the effect of surrounding phones, different HMMs can be used for phones in different contexts. Most commonly, the dependence on the immediate right and left phones is modeled using *triphone* HMMs.

The use of context-dependent phones is a way of handling some pronunciation variation. However, some pronunciation effects involve more than a single phone and its immediate neighbors, such as the rounding of [s] in *strawberry*. Jurafsky et al. [JWJ+01] show that triphones are in general adequate for modeling phone substitutions, but inadequate for handling insertions and deletions.

2.1.3 Decoding

The search for the most likely word string $\mathbf{w}$ is referred to as decoding. With the hypothesis $\mathbf{w}$ represented as an HMM, we can rewrite the speech recognition problem, making several assumptions (described below), as

$$\begin{aligned}
\mathbf{w}^* &= \arg\max_{\mathbf{w}} p(\mathbf{o}|\mathbf{w})p(\mathbf{w}) && (2.7)\\
&= \arg\max_{\mathbf{w}} \sum_{\mathbf{q}} p(\mathbf{o}|\mathbf{w},\mathbf{q})p(\mathbf{q}|\mathbf{w})p(\mathbf{w}) && (2.8)\\
&\approx \arg\max_{\mathbf{w}} \sum_{\mathbf{q}} p(\mathbf{o}|\mathbf{q})p(\mathbf{q}|\mathbf{w})p(\mathbf{w}) && (2.9)\\
&\approx \arg\max_{\mathbf{w}} \max_{\mathbf{q}} p(\mathbf{o}|\mathbf{q})p(\mathbf{q}|\mathbf{w})p(\mathbf{w}) && (2.10)\\
&\approx \arg\max_{\mathbf{w}} \max_{\mathbf{q}} \prod_{t=1}^{T} p(o_t|q_t)p(\mathbf{q}|\mathbf{w})p(\mathbf{w}). && (2.11)
\end{aligned}$$

(2.12)

where q_t denotes the HMM state in time frame t. Eq. 2.9 is simply a re-writing of Eq. 2.8, summing over all of the HMM state sequences $\mathbf{q}$ that are possible realizations of the hypothesis $\mathbf{w}$. In going from Eq. 2.9 to Eq. 2.10, we have made the assumptions that the acoustics are independent of the words given the state sequence. To obtain Eq. 2.11, we have assumed that there is a single state sequence that is much more likely than all others, so that summing over $\mathbf{q}$ is approximately equivalent to maximizing over $\mathbf{q}$. This allows us to perform the search for the most probable word string using the Viterbi algorithm for decoding [BJM83]. Finally, Eq. 2.12 arises directly from the HMM assumption: Given the current state q_t, the current observation o_t is independent of all other states and observations. We refer to $p(o_t|q_t)$ as the *observation model*.[4]

2.1.4 Parameter estimation

Speech recognizers are usually trained, i.e. their parameters are estimated, using the maximum likelihood criterion,

$$\Theta^* = \arg\max_{\theta} p(\mathbf{w}, \mathbf{o}|\Theta),$$

where Θ is a vector of all of the parameter values. Training data typically consists of pairs of word strings and corresponding acoustics. The start and end times of words, phones, and states in the training data are generally unknown; in other words, the training data are *incomplete*. Maximum likelihood training with incomplete data is done using the Expectation-Maximization (EM) algorithm [DLR77], an iterative algorithm that alternates between finding the expected values of all unknown variables and re-estimating the parameters given these expected values, until some criterion of convergence is reached. A special case of the EM algorithm for HMMs is the Baum-Welch algorithm [BPSW70].

2.2 Pronunciation modeling for ASR

We refer to the factor $p(\mathbf{q}|\mathbf{w})$ of Eq. 2.12 as the *pronunciation model*. This is a nonstandard definition: More typically, this probability is expanded as

$$\begin{aligned} p(\mathbf{q}|\mathbf{w}) &= \sum_{\mathbf{u}} p(\mathbf{q}|\mathbf{u}, \mathbf{w}) p(\mathbf{u}|\mathbf{w}) && (2.13) \\ &\approx \max_{\mathbf{u}} p(\mathbf{q}|\mathbf{u}) p(\mathbf{u}|\mathbf{w}) && (2.14) \\ & && (2.15) \end{aligned}$$

where $\mathbf{u} = \{u_1, u_2, \ldots, u_L\}$ is a string of sub-word units, usually phones or phonemes, corresponding to the word sequence $\mathbf{w}$, and the summation in Eq. 2.14 ranges over all

[4] This, rather than $p(o|\mathbf{w})$, is sometime referred to as the acoustic model.

possible phone/phoneme strings. To obtain Eq. 2.15, we have made two assumptions: That the state sequence $\mathbf{q}$ is independent of the words $\mathbf{w}$ given the sub-word unit sequence $\mathbf{u}$, and that, as before, there is a single sequence $\mathbf{u}$ that is much more likely than all other sequences, so that we may maximize rather than sum over $\mathbf{u}$. The second assumption allows us to again use the Viterbi algorithm for decoding.

In the standard formulation, $p(\mathbf{u}|\mathbf{w})$ is referred to as the pronunciation model, while $p(\mathbf{q}|\mathbf{u})$ is the duration model and is typically given by the Markov statistics of the HMMs corresponding to the u_i. In Chapter 1, we noted that there is a dependence between the choice of sub-word units and their durations, as in the example of short epenthetic [t]. For this reason, we do not make this split between sub-word units and durations, instead directly modeling the state sequence $\mathbf{q}$ given the words $\mathbf{w}$, where in our case $\mathbf{q}$ will consist of feature value combinations (see Chapter 3).

2.3 Dynamic Bayesian networks for ASR

Hidden Markov models are a special case of dynamic Bayesian networks (DBNs), a type of *graphical model*. Recently there has been growing interest in the use of DBNs (other than HMMs) for speech recognition (e.g., [Zwe98, BZR+02, Bil03, SMDB04], and we use them in our proposed approach. Here we give a brief introduction to graphical models in general, and DBNs in particular, and describe how DBNs can be applied to the recognition problem. For a more in-depth discussion of graphical models in ASR, see [Bil03].

2.3.1 Graphical models

Probabilistic graphical models [Lau96, Jor98] are a way of representing a joint probability distribution over a given set of variables. A graphical model consists of two components. The first is a graph, in which a node represents a variable and an edge between two variables means that some type of dependency between the variables is allowed (but not required). The second component is a set of functions, one for each node or some subset of nodes, from which the overall joint distribution can be computed.

2.3.2 Dynamic Bayesian networks

For our purposes, we are interested in *directed, dynamic* graphical models, also referred to as dynamic Bayesian networks [DK89, Mur02]. A *directed* graphical model, or Bayesian network, is one in which the graph is directed and acyclic, and the function associated with each node is the conditional probability of that variable given its parents in the graph. The joint probability of the variables in the graph is given by the product of all of the variables' conditional probabilities:

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i|pa(x_i)), \tag{2.16}$$

where x_i is the value of a variable in the graph and $pa(x_i)$ are the values of x_i's parents.

A *dynamic* directed graphical model is one that has a repeating structure, so as to model a stochastic process over time (e.g., speech) or space (e.g., images). We refer to the repeating part of the structure as a *frame*. Since the number of frames is often not known ahead of time, a dynamic model can be represented by specifying only the repeating structure and any special frames at the beginning or end, and then "unrolling" the structure to the necessary number of frames. An HMM is a simple DBN in which each frame contains two variables (the state and the observation) and two dependencies (one from the state to the observation, and one from the state in the previous frame to the current state).

One of the advantages of representing a probabilistic model as a Bayesian network is the availability of standard algorithms for performing various tasks. A basic "subroutine" of many tasks is *inference*, the computation of answers to queries of the form, "Given the values of the set of variables X_A (*evidence*), what are the distributions or most likely values of variables in set X_B?" This is a part of both decoding and parameter learning. Algorithms exist for doing inference in Bayesian networks in a computationally efficient way, taking advantage of the factorization of the joint distribution represented by the graph [HD96]. There are also approximate inference algorithms [JGJS99, McK99], which provide approximations to the queried distributions, for the case in which a model is too complex for exact inference. Viterbi decoding and Baum-Welch training of HMMs are special cases of the corresponding generic DBN algorithms [Smy98].

Zweig [Zwe98] demonstrated how HMM-based speech recognition can be represented as a dynamic Bayesian network. Figure 2-1 shows three frames of a phone HMM-based decoder represented as a DBN. This is simply an encoding of a typical HMM-based recognizer, with the hidden state factored into its components (word, phone state, etc.). Note that the model shown is intended for *decoding*, which corresponds to finding the highest-probability settings of all of the variables and then reading off the value of the word variable in each frame. For *training*, slightly different models with additional variables and dependencies are required to represent the known word string.

Several extensions to HMMs have been proposed for various purposes, for example to make use of simultaneous speech and video [GPN02] or multiple streams of acoustic observations [BD96]. Viewed as modifications of existing HMM-based systems, such extensions often require developing modified algorithms and new representations. Viewed as examples of DBNs, they require no new algorithms or representations, and can stimulate the exploration of a larger space of related models. It is therefore a natural generalization to use DBNs as the framework for investigations in speech recognition. Bilmes [Bil99, Bil00] developed an approach for discriminative learning of aspects of the structure of a DBN, and used this to generate extensions of an HMM-based speech recognizer with additional learned dependencies between observations.

There have been several investigations into using models similar to that of Figure 2-1 with one or two additional variables to encode articulatory information [SMDB04,

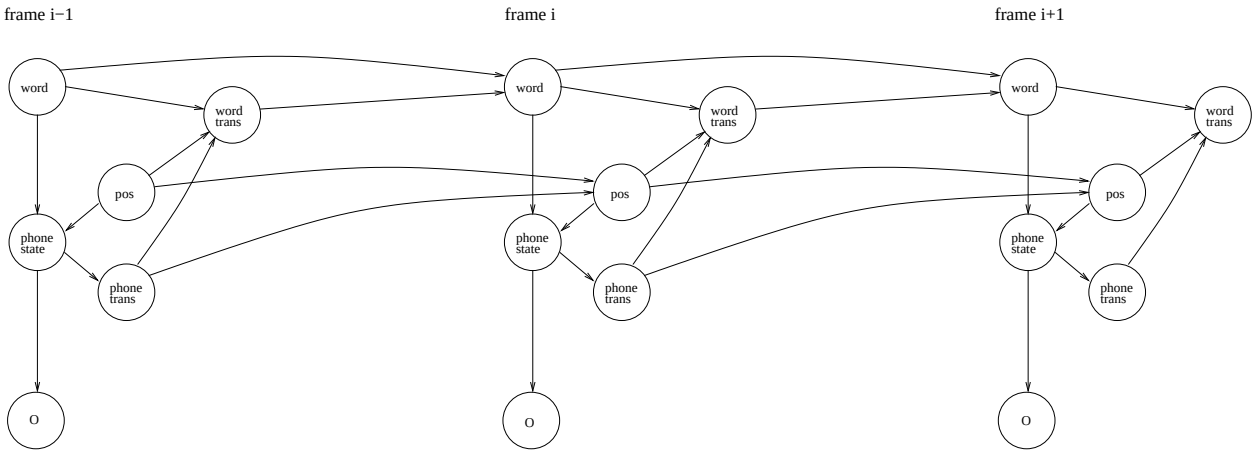

Figure 2-1: *A phone-state HMM-based DBN for speech recognition. The* word *variable is the identity of the word that spans the current frame;* word trans *is a binary variable indicating whether this is the last frame in the current word;* phone state *is the phonetic state that spans the current frame (there are several states per phone);* pos *indicates the phonetic position in the current word,* i.e. *the current phone state is the* pos^{th} *phone state in the word;* phone trans *is the analogue of* word trans *for the phone state; and* O *is the current acoustic observation vector.*

Zwe98]. In [Zwe98], Zweig also suggested, but did not implement, a DBN using a full set of articulatory features, shown in Figure 2-2. A related model, allowing for a small amount of deviation from canonical feature values, was used for a noisy digit recognition task in [LGB03].

2.4 Linguistic background

We now briefly describe the linguistic concepts and theories relevant to current practices in ASR and to the ideas we propose. We do not intend to imply that recognition models should aim to faithfully represent the most recent (or any particular) linguistic theories, and in fact the approach we will propose is far from doing so. However, ASR research has always drawn on knowledge from linguistics, and one of our motivations is that there are many additional ideas in linguistics to draw on than have been used in recognition to date. Furthermore, recent linguistic theories point out flaws in older ideas used in ASR research, and it is worthwhile to consider whether these flaws merit a change in ASR practice.

2.4.1 Generative phonology

Much of the linguistic basis of state-of-the-art ASR systems originates in the generative phonology of the 1960s and 1970s, marked by the influential *Sound Pattern of English* of Chomsky and Halle [CH68]. Under this theory, phonological represen-

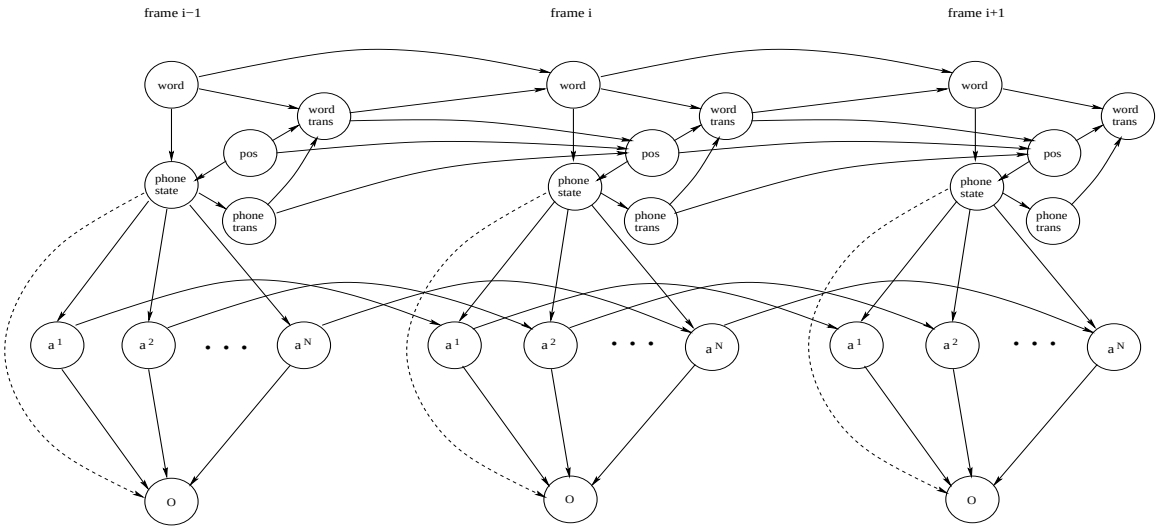

Figure 2-2: *An articulatory feature-based DBN for speech recognition suggested by Zweig [Zwe98]. a^i are articulatory feature values; remaining variables are as in Figure 2-1.*

tations consist of an underlying (phonemic) string, which is transformed via a set of rules to a surface (phonetic) string. Speech segments (phonemes and phones) are classified with respect to a number of binary features, such as voicing, nasality, tongue high/low, and so on, many of which are drawn from the features of Jakobson, Fant, and Halle [JFH52]. Rules can refer to the features of the segments they act on; for example, a vowel nasalization rule may look like

- x → x_n / __ [+nasal]

However, features are always part of a "bundle" corresponding to a given segment and act only as an organizing principle for categorizing segments and rules. In all cases, the phonological and phonetic representation of an utterance is a single string of symbols. For this reason, this type of phonology is referred to as *linear* phonology.

In ASR research, these ideas form the basis of (i) the string-of-phones representation of words, (ii) clustering HMM states according to binary features of the current/neighboring segments, and (iii) modeling pronunciation variation using rules for the substitution, insertion, and deletion of segments.

2.4.2 Autosegmental phonology

In the late 1970s, Goldsmith introduced the theory of autosegmental phonology [Gol76, Gol90]. According to this theory, the phonological representation no longer consists of a single string of segments but rather of multiple strings, or *tiers*, corresponding to different linguistic features. Features can be of the same type as the Jakobson, Fant, and Halle features, but can also include additional features such as tone. This theory was motivated by the observation that some phenomena of feature spreading

are more easily explained if a single feature value is allowed to span (what appears on the surface to be) more than one segment. Autosegmental phonology posits some relationships (or *associations*) between segments in different tiers, which limit the types of transformations that can occur. We will not make use of the details of this theory, other than the motivation that features inherently lie in different tiers of representation.

2.4.3 Articulatory phonology

In the late 1980s, Browman and Goldstein proposed articulatory phonology [BG86, BG92], a theory that differs from previous ones in that the basic units in the lexicon are not abstract binary features but rather articulatory gestures. A gesture is essentially an instruction to the vocal tract to produce a certain degree of constriction at a given location with a given set of articulators. For example, one gesture might be "narrow lip opening", an instruction to the lips and jaw to position themselves so as to effect a narrow opening at the lips. Figure 2-3 shows the main articulators of the vocal tract to which articulatory gestures refer. We are mainly concerned with the lips, tongue, glottis (controlling voicing), and velum (controlling nasality).

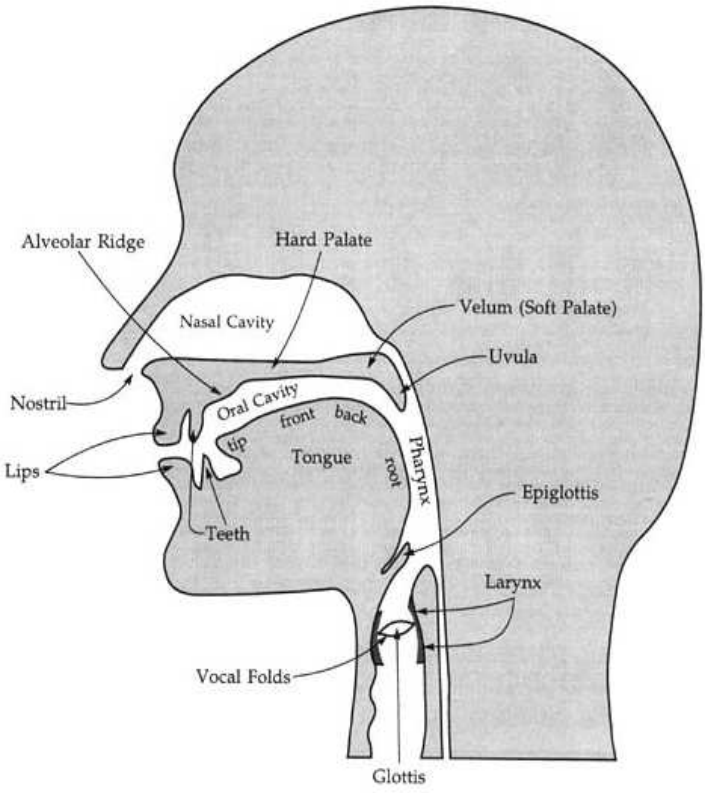

Figure 2-3: *A midsagittal section showing the major articulators of the vocal tract, reproduced from [oL04].*

The degrees of freedom in articulatory phonology are referred to as *tract variables* and include the locations and constriction degrees of the lips, tongue tip, and tongue body, and the constriction degrees of the glottis and velum. The tract variables and the articulators to which each corresponds are shown in Figure 2-4.

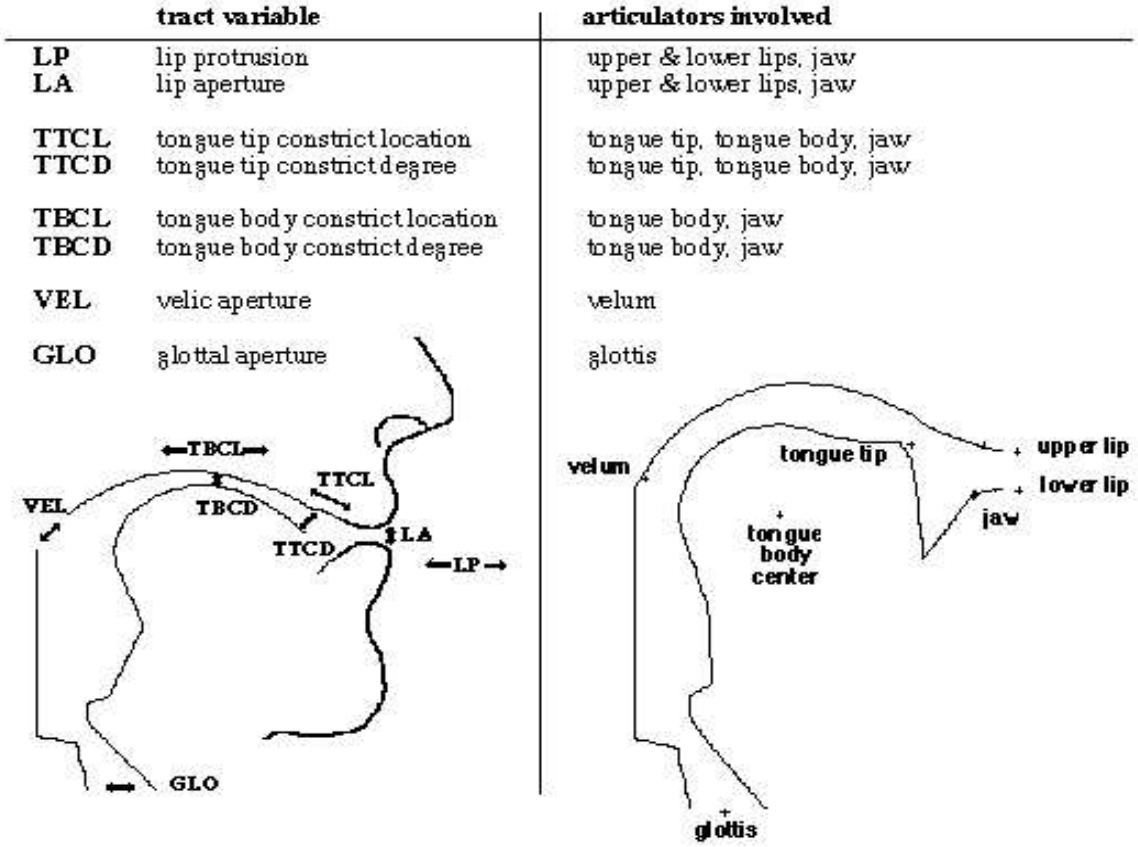

Figure 2-4: *Vocal tract variables and corresponding articulators used in articulatory phonology. Reproduced from [BG92].*

In this theory, the underlying representation of a word consists of a gestural *score*, indicating the gestures that form the word and their relative timing. Examples of gestural scores for several words are given in Figure 2-5. These gestural targets may be modified through gestural *overlap*, changes in timing of the gestures so that a gesture may begin before the previous one ends; and gestural *reduction*, changes from more extreme to less extreme targets. The resulting modified gestural targets are the input to a *task dynamics* model of speech production [SM89], which produces the actual trajectories of the tract variables using a second-order damped dynamical model of each variable.

Browman and Goldstein argue in favor of such a representation on the basis of fast speech data of the types we have discussed, as well as articulatory measurements showing that underlying gestures are often produced faithfully even when overlap prevents some of the gestures from appearing in their usual acoustic form. For example,

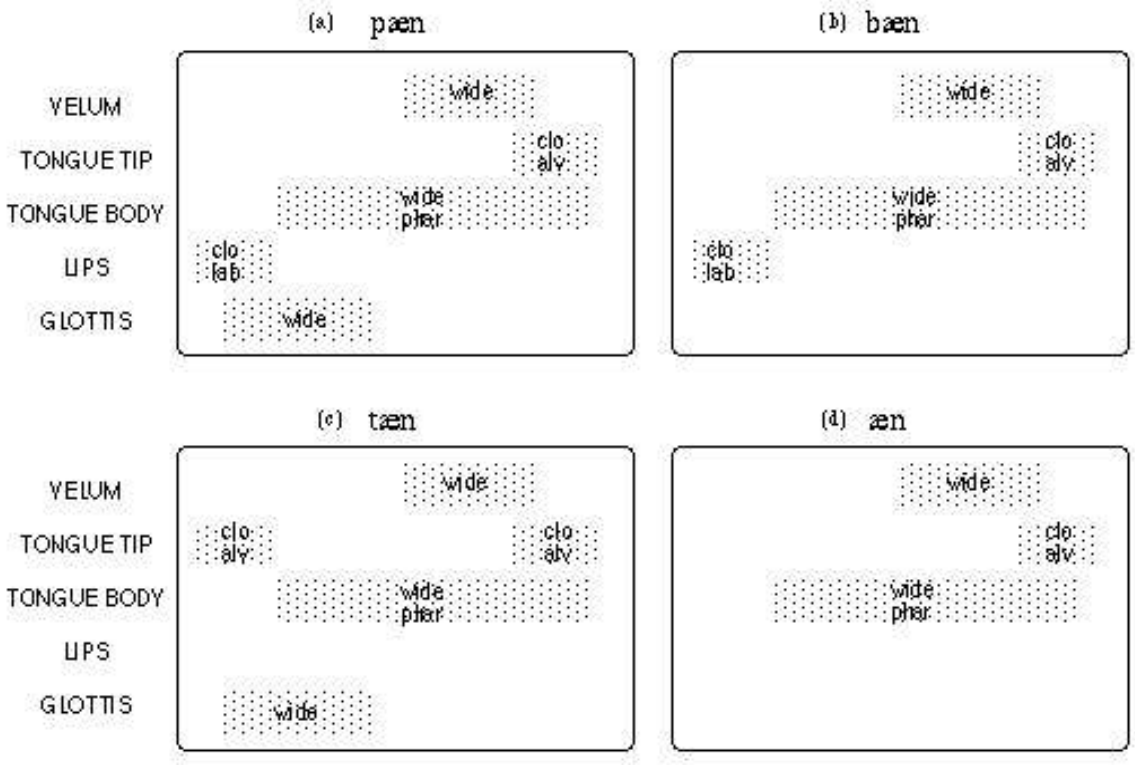

Figure 2-5: *Gestural scores for several words. Reproduced from http://www.haskins.yale.edu/haskins/MISC/RESEARCH/GesturalModel.html.*

they cite X-ray evidence from a production of the phrase *perfect memory*, in which the articulatory motion for the final [t] of *perfect* appeared to be present despite the lack of an audible [t].

Articulatory phonology is under active development, as is a system for speech synthesis based on the theory [BGK+84]. We draw heavily on ideas from articulatory phonology in our own proposed model in Chapter 3. We note that in the sense in which we use the term "feature", Browman and Goldstein's tract variables can be considered a type of feature set (although they do not consider these to be the basic unit of phonological representation). Indeed, the features we will use correspond closely to their tract variables.

2.5 Previous ASR research using linguistic features

The automatic speech recognition literature is rife with proposals for modeling speech at the sub-phonetic feature level. Rose et al. [RSS94] point out that the primary articulator for a given sound often displays less variation than other articulators, suggesting that while phone-based models may be thrown off by the large amount of *overall* variation, the critical articulator may be more easily detected and used to improve the robustness of ASR systems. Ostendorf [Ost99, Ost00] notes the large

distance between current ASR technology and recent advances in linguistics, and suggests that ASR could benefit from a tighter coupling.

Although linguistic features have not yet found their way into mainstream, state-of-the-art recognition systems, they been used in various ways in ASR research. We now briefly survey the wide variety of related research. This survey covers work at various stages of maturity and is at the level of ideas, rather than of results. The goal is to get a sense for the types of models that have been proposed and used, and to demonstrate the need for a different approach.

An active area of work has been feature detection and classification, either as a stand-alone task or for use in a recognizer [FK01, Eid01, WFK04, KFS00, MW02, WR00]. Different types of classifiers have been used, including Gaussian mixture [Eid01, MW02] and neural network-based classifiers [WFK04, KFS00]. In [WFK04], asynchronous feature streams are jointly recognized using a dynamic Bayesian network that models possible dependencies between features. In almost all cases in which the outputs of such classifiers have been used in a complete recognizer [KFS00, MW02, Eid01], it has been assumed that the features are synchronized to the phone and with values given deterministically by a phone-to-feature mapping.

There have been a few attempts to explicitly model the asynchronous evolution of features. Deng et al. [DRS97], Erler and Freeman [EF94], and Richardson et al. [RBD00] used HMMs in which each state corresponds to a combination of feature values. They constructed the HMM feature space by allowing features to evolve asynchronously between phonetic targets, while requiring that the features re-synchronize at phonetic (or bi-phonetic) targets. This somewhat restrictive constraint was necessary to control the size of the state space. One drawback to this type of system is that it does not take advantage of the factorization of the state space, or equivalently the conditional independence properties of features.

In [Kir96], on the other hand, Kirchhoff models the feature streams independently in a first pass, then aligns them to syllable templates with the constraint that they must synchronize at syllable boundaries. Here the factorization into features is taken advantage of, although the constraint of synchrony at syllable boundaries is perhaps a bit strong. Perhaps more importantly, however, there is no control over the asynchrony within a syllable, for example by indicating that a small amount of asynchrony may be preferable to a large amount.

An alternative approach, presented by Blackburn [Bla96, BY01], is analysis by synthesis: A baseline HMM recognizer produces an N-best hypothesis for the input utterance, and an articulatory synthesizer converts each hypothesis to an acoustic representation for matching against the input.

Bates [Bat04] uses the idea of factorization into feature streams in a model of phonetic substitutions, in which the probability of a surface phone s_t, given its context c_t (typically consisting of the underlying phoneme, previous and following phonemes, and some aspect of the word context), is the product of probabilities corresponding to each of the phone's N features $f_{t,i}, i = 1..N$:

$$P(s_t|c_t) = \prod_{i=1}^{N} P(f_{t,i}|c_t) \tag{2.17}$$

Bates also considers alternative formulations where, instead of having a separate factor for each feature, the feature set is divided into groups and there is a factor for each group. This model assumes that the features or feature groups are independent given the context. This allows for more efficient use of sparse data, as well as feature combinations that do not correspond to any canonical phone. Each of the probability factors is represented by a decision tree learned over a training set of transcribed pronunciations. Bates applies a number of such models to manually transcribed Switchboard pronunciations using a set of distinctive features, and finds that, while these models do not improve phone perplexity or prediction accuracy, they predict surface forms with a smaller feature-based distance from ground truth than does a phone-based model. In this work, the *values* of features are treated as independent, but their *time course* is still dependent, in the sense that features are constrained to organize into synchronous, phoneme-sized segments.

In addition, several models have been proposed in which linguistic and speech science theories are implemented more faithfully. Feature-based representations have been, for a long time, used in the landmark-based recognition approach of Stevens [Ste02]. In this approach, recognition starts by hypothesizing the locations of *landmarks*, important events in the speech signal such as stop bursts and extrema of glides. Various cues, such as voice onset times or formant trajectories, are then extracted around the landmarks and used to detect the values of distinctive features such as voicing, stop place, and vowel height, which are in turn matched against feature-based word representations in the lexicon.

Recent work by Tang [Tan05] combines a landmark-based framework with a previous proposal by Huttenlocher and Zue [HZ84] of using sub-phonetic features as a way of reducing the lexical search space to a small cohort of words. Tang et al. use landmark-based acoustic models clustered according to place and manner features to obtain the relevant cohort, then perform a second pass using the more detailed phonetic landmark-based acoustic models of the MIT SUMMIT recognizer [Gla03] to obtain the final hypothesis. In this work, then, features are used as a way of defining broad phonetic classes.

In Lahiri and Reetz's [LR02] featurally underspecified lexicon (FUL) model of human speech perception, the lexicon is underspecified with respect to some features. Speech perception proceeds by (a) converting the acoustics to feature values and (b) matching these values against the lexicon, allowing for a no-mismatch condition when comparing against an underspecified lexical feature. Reetz [Ree98] describes a knowledge-based automatic speech recognition system based on this model, involving detailed acoustic analysis for feature detection. This system requires an error correction mechanism before matching features against the lexicon.

Huckvale [Huc94] describes an isolated-word recognizer using a similar two-stage strategy. In the first stage, a number of articulatory features are classified in each frame using separate multi-layer perceptrons. Each feature stream is then separately aligned with each word's baseform pronunciations and an N-best list is derived for each. Finally, the N-best lists are combined heuristically, using the N-best lists corresponding to the more reliable features first. As noted in [Huc94], a major drawback of this type of approach is the inability to jointly align the feature streams with the

baseforms, thereby potentially losing crucial constraints. This is one problem that our proposed approach corrects.

2.6 Summary

This chapter has presented the setting in which the work in this thesis has come about. We have described the generative statistical framework for ASR and the graphical modeling tools we will use, as well as the linguistic theories from which we draw inspiration and previous work in speech recognition using ideas from these theories. One issue that stands out from our survey of previous work is that there has been a lack of computational frameworks that can combine the information from multiple feature streams in a principled, flexible way: Models based on conventional ASR technology tend to ignore the useful independencies between features, while models that allow for more independence typically provide little control over this independence. Our goal, therefore, is to formulate a general, flexible model of the joint evolution of linguistic features. The next chapter presents such a model, formulated as a dynamic Bayesian network.

Chapter 3

Feature-based Modeling of Pronunciation Variation

This chapter describes the proposed approach of modeling pronunciation variation in terms of the joint evolution 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*, controlled via soft constraints, and (3) *substitutions* of individual feature values.

In Chapter 1, we defined a pronunciation of a word as a representation, in terms of a set of sub-word units, of the way the word is or can be produced by a speaker. Section 3.1 defines the representations we use to describe underlying and surface pronunciations. We next give a detailed procedure—a "recipe"—by which the model generates surface feature values from an underlying dictionary (Section 3.2). This is intended to be a more or less complete description, requiring no background in dynamic Bayesian networks. In order to use the model in any practical setting, of course, we need an implementation that allows us to (a) query the model for the relative probabilities of given surface representations of words, for the most likely word given a surface representation, or for the best analysis of a word given its surface representation; and to (b) learn the parameters of the model automatically from data. Section 3.3 describes such an implementation in terms of dynamic Bayesian networks. Since automatic speech recognition systems are typically presented not with surface feature values but with a raw speech signal, Section 3.4 describes the ways in which the proposed model can be integrated into a complete recognizer. Section 3.5 relates our approach to some of the previous work described in Chapter 2. We close in Section 3.6 with a discussion of the main ideas of the chapter and consider some aspects of our approach that may bear re-examination.

3.1 Definitions

We define an *underlying* pronunciation of a word in the usual way, as a string of phonemes. Closely related are *baseform* pronunciations, the ones typically stored in

an ASR pronouncing dictionary. These are canonical pronunciations represented as strings of phones of various levels of granularity, depending on the degree of detail needed in a particular ASR system.[1] We will typically treat baseforms as our "underlying" representations, from which we derive surface pronunciations, rather than using true phonemic underlying pronunciations.

We define *surface* pronunciations in a somewhat unconventional way. In Section 1.1.4, we proposed a representation consisting of multiple streams of feature values, as in this example:

feature	values			
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 3.1: *An example observed pronunciation of* sense *from Chapter 1.*

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 [t]s inserted in a [n] _ [s] context to be short. To formalize this, then, we define a surface pronunciation as a *time-aligned* listing of all of the surface feature values produced by a speaker. Referring to the discussion in Chapter 2, this means that we define $\mathbf{q}_t$ as the vector of surface feature values at time t. Such a representation might look like the above, with the addition of time stamps (using some abbreviations for feature values):

voi.	off	.1s	on	.2s	off	.5s
nas.	off	.1s	on	.2s	off	.5s
lips	open					.5s
t. body	m/u	.1s	m-n/p	.2s	m/u	.5s
t. tip	cr/a	.1s	m-n/a	.2s	cl/a .35s	cr/a .5s

Table 3.2: *Time-aligned surface pronunciation of* sense.

In practice, we will assume that time is discretized into short *frames*, say of 10ms each. Therefore, for our purposes a surface pronunciation will be represented as in Table 3.3. This representation is of course equivalent to the one in Table 3.2 when the time stamps are discretized to multiples of the frame size.

[1]For example, a baseform for *tattle* may differentiate between the initial plosive [t] and the following flap [dx]: [t ae dx el], although both are phonemically /t/; but the baseform for *ninth* may not differentiate between the two nasals, although the first is typically alveolar while the second is dental.

frame	1	2	3	4	5	6	7	8	9	10	11	...
voi.	off	off	off	off	off	off	off	off	off	off	on	...
nas.	off	off	off	off	off	off	off	off	off	off	on	...
lips	op	op	op	op	op	op	op	op	op	op	op	...
t. body	m/u	m/u	m/u	m/u	m/u	m/u	m/u	m/u	m/u	m/u	m-n/p	...
t. tip	cr/a	cr/a	cr/a	cr/a	cr/a	cr/a	cr/a	cr/a	cr/a	cr/a	m-n/a	...

Table 3.3: *Frame-by-frame surface pronunciation of* sense.

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 the relative likelihoods of the different pronunciations. We denote the feature set, consisting of N features, $F^i, 1 \leq i \leq N$. A T-frame surface pronunciation in terms of these features is denoted $S_t^i, 1 \leq i \leq N, 1 \leq t \leq T$, where S_t^i is the surface value of feature F^i in time frame t.

Our approach begins with the usual assumption that each word has one or more baseforms. Each baseform is then converted to a table of *underlying*, or target, feature values, using a phone-to-feature mapping table[2]. For this purpose, dynamic phones consisting of more than one feature configuration are divided into multiple segments: Stops are divided into a closure and a release; affricates into a closure and a frication portion; and diphthongs into the beginning and ending configurations. More precisely, the mapping from phones to feature values may be probabilistic, giving a distribution over the possible values for a given feature and phone. Table 3.4 shows what a baseform for *sense* and the corresponding underlying feature distributions might look like. For the purposes of our example, we are assuming a feature set based on the locations and opening degrees of the articulators, similarly to the vocal tract variables of articulatory phonology [BG92]; however, our approach will not assume a particular feature set. In the following experimental chapters, we will give fuller descriptions of the feature sets we use.

The top row of Table 3.4 is simply an index into the underlying phone sequence; it will be needed in the discussion of asynchrony. This is not to be confused with the frame number, as in Table 3.3: The index says nothing about the amount of time spent in a particular feature configuration. Note that it is assumed that all features go through the same sequence of indices (and therefore 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. Alternatively, we could have a single index value for identical successive targets, and a different way of measuring asynchrony (see below).

[2]Here we are abusing terminology slightly, as the underlying features do not necessarily correspond to an underlying (phonemic) pronunciation.

index	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 mid/uvular
tongue tip	critical/alveolar	mid/alveolar	closed/alveolar	critical/alveolar
phone	s	eh	n	s

Table 3.4: *A possible baseform and target feature distributions for the word* sense. *Expressions of the form "f_1 p_1" give the probabilities of different feature values; for example, the target value for the feature* **tongue body** *for an [n] is* mid/velar *or* mid/uvular *with probability 0.5 each. When no probability is given for a feature value, it is assumed to be 1.*

This is an issue that warrants future re-examination.

The baseform table does not tell us anything about the amount of time each feature spends in each state; this is our next task.

3.2.1 Asynchrony

We assume that in the first time frame of speech, all of the features begin in index 1, $ind_1^i = 1 \; \forall i$. In subsequent frames, each feature can either stay in the same state or transition to the next one with some *transition probability*. The transition probability may depend on the phone corresponding to the feature's current index: Phones with longer intrinsic durations will tend to have higher transition probabilities. Features may transition at different times. This is what we refer to as feature *asynchrony*. We define the *degree of asynchrony* between two features F^i and F^j in a given time frame t as the absolute difference between their indices in that frame:

$$async_t^{i:j} = |ind_t^i - ind_t^j|. \tag{3.1}$$

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:

$$async_t^{A:B} = round\left(|mean\left(ind_t^A\right) - mean\left(ind_t^B\right)|\right), \tag{3.2}$$

where A and B are subsets of $\{1, \ldots, N\}$ and $F^{\{i_1, i_2, \ldots\}} = \{F^{i_1}, F^{i_2}, \ldots\}$. For example, Tables 3.5 and 3.6 show two possible sets of trajectories for the feature indices in *sense*, assuming a 10-frame utterance.

The degree of asynchrony may be constrained: More "synchronous" configurations may be more probable (soft constraints), and there may be an upper bound on the degree of asynchrony (hard constraints). For example, the sequence of asynchrony values in Table 3.5 may be preferable to the one in Table 3.6. We express this by imposing a distribution over the degree of asynchrony between features in each frame,

frame	1	2	3	4	5	6	7	8	9	10
voi. index	1	1	2	3	3	3	4	4	4	4
voi. phone	s	s	eh	n	n	n	s	s	s	s
voicing	off	off	on	on	on	on	off	off	off	off
nas. index	1	1	2	3	3	3	4	4	4	4
nas. phone	s	s	eh	n	n	n	s	s	s	s
nasality	off	off	off	on	on	on	off	off	off	off
t.b. index	1	1	2	2	2	3	3	4	4	4
t.b. phone	s	s	eh	eh	eh	n	n	s	s	s
t. body	m/u	m/u	m/u	m/p	m/p	m/u	m/u	m/u	m/u	m/u
t.t. index	1	1	2	2	2	3	3	4	4	4
t.t. phone	s	s	eh	eh	eh	n	n	s	s	s
t. tip	cr/a	cr/a	cr/a	m/a	m/a	cl/a	cl/a	cr/a	cr/a	cr/a
$async^{A:B}$	0	0	0	1	1	0	1	0	0	0
phone	s	s	eh	eh_n	eh_n	n	t	s	s	s

Table 3.5: *Frame-by-frame sequences of index values, corresponding phones, underlying feature values, and degrees of asynchrony between {voicing, nasality} and {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. The lips feature has been left off, as its index sequence does not make a difference to the surface pronunciation. The bottom row shows the resulting phone transcription corresponding to these feature values, assuming they are produced canonically.*

$p(async_t^{i:j})$, or feature sets, $p(async_t^{A:B})$. One of the design choices in using such a model is which features or sets of features will have such explicit constraints.[3]

3.2.2 Substitution

Given the index sequence for each feature, the corresponding frame-by-frame sequence of underlying feature values, $u_t^i, 1 \leq t \leq T$, is drawn according to the feature distributions in the baseform table (Table 3.4). However, a feature may fail to reach its target value, instead substituting another value. This may happen, for example, if the speaker fails to make a constriction as extreme as intended, or if a given feature value assimilates to neighboring values. One example of substitution is *sense* → [s ih_n t s]; a frame-by-frame representation is shown in Table 3.7. Table 3.8 shows what might happen if the alveolar closure of the [n] is not made, i.e. if the tongue tip value of *closed/alveolar* is substituted with *mid/alveolar*. The result will be a surface pronunciation with a nasalized vowel but no closure, which might be transcribed phonetically as [s eh_n s]. This is also a common effect in words with post-vocalic nasals [Lad01].

We model substitution phenomena with a distribution over each surface feature value in a given frame given its corresponding underlying value, $p(s_t^i|u_t^i)$. For the

[3]There may also be some implicit constraints, e.g. the combination of a constraint on $async_t^{i:j}$ and another constraint on $async_t^{j:k}$ will result in an implicit constraint on features i and k.

frame	1	2	3	4	5	6	7	8	9	10
voi. index	1	2	3	3	3	4	4	4	4	4
voi. phone	s	eh	n	n	n	s	s	s	s	s
voicing	off	on	on	on	on	off	off	off	off	off
nas. index	1	2	3	3	3	4	4	4	4	4
nas. phone	s	eh	n	n	n	s	s	s	s	s
nasality	off	off	off	on	on	off	off	off	off	off
t.b. index	1	1	1	2	2	3	3	4	4	4
t.b. phone	s	s	s	eh	eh	n	n	s	s	s
t. body	m/u	m/u	m/u	m/p	m/p	m/u	m/u	m/u	m/u	m/u
t.t. index	1	1	1	2	2	3	3	4	4	4
t.t. phone	s	s	s	eh	eh	n	n	s	s	s
t. tip	cr/a	cr/a	cr/a	m/a	m/a	cl/a	cl/a	cr/a	cr/a	cr/a
$async^{A:B}$	0	1	2	1	1	1	1	0	0	0
phone	s	z	z_n	eh_n	eh_n	t	t	s	s	s

Table 3.6: *Another possible set of frame-by-frame sequences for* sense.

time being we model substitutions context-independently: Each surface feature value depends only on the corresponding underlying value in the same frame. However, it would be fairly straightforward to extend the model with substitutions that depend on such factors as preceding and following feature values, stress, or syllable position.

3.2.3 Summary

To summarize the generative recipe, we can generate all possible surface pronunciations of a given word in the following way:

1. List the baseforms in terms of underlying features.
2. For each baseform, generate all possible combinations of index sequences, with probabilities given by the transition and asynchrony probabilities.
3. For each such generated index sequence, generate all possible underlying feature values by drawing from the feature distributions at each index.
4. For each underlying feature value, generate the possible surface feature values according to $p(s_t^i|u_t^i)$.

3.3 Implementation using dynamic Bayesian networks

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 3-1 shows one frame of the type of DBN used in our model. For our purposes, we will assume an isolated-word setup; i.e. we will only be recognizing one

frame	1	2	3	4	5	6	7	8	9	10
voi. index	1	1	2	3	3	3	4	4	4	4
voi. phone	s	s	eh	n	n	n	s	s	s	s
voicing (U)	off	off	on	on	on	on	off	off	off	off
voicing (S)	off	off	on	on	on	on	off	off	off	off
nas. index	1	1	2	3	3	3	4	4	4	4
nas. phone	s	s	eh	n	n	n	s	s	s	s
nasality (U)	off	off	off	on	on	on	off	off	off	off
nasality (S)	off	off	off	on	on	on	off	off	off	off
t.b. index	1	1	2	2	2	3	3	4	4	4
t.b. phone	s	s	eh	eh	eh	n	n	s	s	s
t. body (U)	m/u	m/u	m/p	m/p	m/p	m/u	m/u	m/u	m/u	m/u
t. body (S)	m/u	m/u	m-n/p	m-n/p	m-n/p	m/u	m/u	m/u	m/u	m/u
t.t. index	1	1	2	2	2	3	3	4	4	4
t.t. phone	s	s	eh	eh	eh	n	n	s	s	s
t. tip (U)	cr/a	cr/a	m/a	m/a	m/a	cl/a	cl/a	cr/a	cr/a	cr/a
t. tip (S)	cr/a	cr/a	m-n/a	m-n/a	m-n/a	cl/a	cl/a	cr/a	cr/a	cr/a
$async^{A:B}$	0	0	0	1	1	0	1	0	0	0
phone	s	s	ih	ih_n	ih_n	n	t	s	s	s

Table 3.7: *Frame-by-frame sequences of index values, corresponding phones, underlying (U) and surface (S) feature values, and degrees of asynchrony between* {*voicing, nasality*} *and* {*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. 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, assuming they are produced canonically.*

word at a time. However, similar processes occur at word boundaries as do word-internally so that the same type of model could be used for multi-word sequences.

This example assumes a feature set with three features, and a separate DBN for each word. The variables at time frame t are as follows:

$baseform_t$ – The current baseform at time t. For $t = 1$, its distribution is given by the probability of each variant in the baseform dictionary; in subsequent frames, its value is copied from the previous frame.

ind_t^j – index of feature j at time t. $ind_0^j = 0\ \forall j$; in subsequent frames ind_t^j is conditioned on ind_{t-1}^j, and $phTr_{t-1}^j$ (defined below).

ph_t^j – canonical phone corresponding to position ind_t^j of the current word and baseform. Deterministic.

$phTr_t^j$ – binary variable indicating whether this is the last frame of the current phone.

U_t^j – underlying value of feature j. Has a (typically) sparse distribution given ph_t^j.

S_t^j – surface value of feature j. $p(S_t^j|U_t^j)$ encodes allowed feature substitutions.

frame	1	2	3	4	5	6	7	8	9	10
voi. index	1	1	2	3	3	3	3	4	4	4
voi. phone	s	s	eh	n	n	n	n	s	s	s
voicing (U)	off	off	on	on	on	on	on	off	off	off
voicing (S)	off	off	on	on	on	on	off	off	off	off
nas. index	1	1	2	3	3	3	3	4	4	4
nas. phone	s	s	eh	n	n	n	n	s	s	s
nasality (U)	off	off	off	on	on	on	on	off	off	off
nasality (S)	off	off	off	on	on	on	off	off	off	off
t.b. index	1	1	2	2	2	3	3	4	4	4
t.b. phone	s	s	eh	eh	eh	n	n	s	s	s
t. body (U)	m/u	m/u	m/p	m/p	m/p	m/u	m/u	m/u	m/u	m/u
t. body (S)	m/u	m/u	m/p	m/p	m/p	m/p	m/p	m/u	m/u	m/u
t.t. index	1	1	2	2	2	3	3	4	4	4
t.t. phone	s	s	eh	eh	eh	n	n	s	s	s
t. tip (U)	cr/a	cr/a	m/a	m/a	m/a	cl/a	cl/a	cr/a	cr/a	cr/a
t. tip (S)	cr/a	cr/a	m/a	m/a	m/a	m/a	m/a	cr/a	cr/a	cr/a
$async^{A:B}$	0	0	0	1	1	0	1	0	0	0
phone	s	s	eh	eh_n	eh_n	eh_n	eh_n	s	s	s

Table 3.8: *Another possible set of frame-by-frame sequences for* sense*, resulting in* sense → *[s eh_n s] (using the convention that [eh_n] refers to either a completely or a partially nasalized vowel).*

$wdTr_t$ – binary variable indicating whether this is the last frame of the word. Deterministic and equal to one if all ind^j are at the maximum for the current baseform and all $\text{phTr}_t^j = 1$.

$async_t^{A;B}$ and $checkSync_t^{A;B}$ are responsible for implementing the asynchrony constraints. $async_t^{A;B}$ is drawn from an (unconditional) distribution over the integers, while $checkSync_t^{A;B}$ checks that the degree of asynchrony between A and B is in fact equal to $async_t^{A;B}$. To enforce this constraint, $checkSync_t^{A;B}$ is always observed with value 1 and is given deterministically by its parents' values, via the distribution[4]

$$\begin{aligned} P\left(checkSync_t^{A;B}{=}1 | async_t^{A;B}, ind_t^A, ind_t^B\right) &= 1 \\ \iff round\left(|mean\left(ind_t^A\right) - mean\left(ind_t^B\right)|\right) &= async_t^{A;B}, \end{aligned}$$

or, equivalently,

$$checkSync_t^{A;B} = 1 \iff round\left(|mean\left(ind_t^A\right) - mean\left(ind_t^B\right)|\right) = async_t^{A;B} \quad (3.3)$$

[4]We note that the asynchrony constraints could also be represented more concisely as *undirected* edges among the corresponding *ind* variables. We represent them in this way to show how these constraints can be implemented and their probabilities learned within the framework of DBNs.

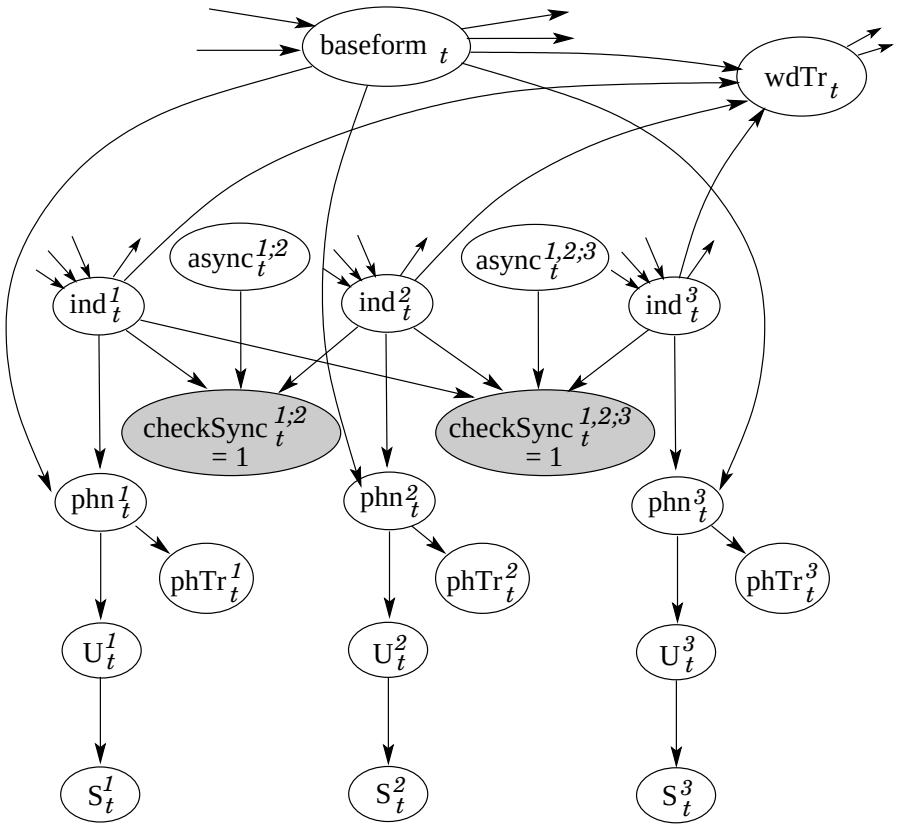

Figure 3-1: *DBN implementing a feature-based pronunciation model with three features and two asynchrony constraints. Edges without parents/children point from/to variables in adjacent frames (see text).*

Once we have expressed the model as a DBN, we can use standard DBN inference algorithms to answer such questions as:

- **Decoding:** Given a set of surface feature value sequences, what is the most likely word that generated them?
- **Parameter learning:** Given a database of words and corresponding surface feature values, what are the best settings of the conditional probability tables (CPT) in the DBN?
- **Alignment:** Given a word and a corresponding surface pronunciation, what is the most likely way the surface pronunciation came about, i.e. what are the most likely sequences of ind_t^i and U_t^i?

There are many interesting issues in inference and learning for DBNs. Perhaps the most important is the choice of optimization criterion in parameter learning. These are, however, general questions equally applicable to any probabilistic model one might use for ASR, and are an active area of research both within ASR [McD00, DB03] and in the graphical models area in general [GGS97, LMP01, GD04]. These questions

are outside the scope of this thesis. For all experiments in this thesis, we will assume a maximum-likelihood learning criterion and will use the Expectation-Maximization algorithm [DLR77] for parameter learning. The observed variables during training can be either the surface feature values, if a transcribed training set is available, or the acoustic observations themselves (see Section 3.4). One issue that would be particularly useful to pursue in the context of our model is that of learning aspects of the DBN *structure*, such as the groupings of features for asynchrony constraints and possible additional dependencies between features. This, too, is a topic for future work.

3.4 Integrating with observations

The model as stated in the previous section makes no assumptions about the relationship between the discrete surface feature values and the (generally continuous-valued) acoustics, and therefore is not a complete recognizer. We now describe a number of ways in which the DBN of Figure 3-1 can be combined with acoustic models of various types to perform end-to-end recognition. Our goal is not to delve deeply into these methods or endorse one over the others; we merely point out that there are a number of choices available for using this pronunciation model in a complete recognizer.

We assume that the acoustic observations are frame-based as in most conventional ASR systems; that is, they consist of a list of vector of acoustic measurements corresponding to contiguous, typically equal-sized, segments of the speech signal. We do not address the integration of this model into segment-based recognizers such as the MIT SUMMIT system [Gla03], in which the acoustic observations are defined as a graph, rather than a list, of vectors.

The integration method most closely related to traditional HMM-based ASR would be to add a single additional variable corresponding to the acoustic observations (say, Mel-frequency cepstral coefficients (MFCCs)) in each frame, as a child of the surface feature values, with a Gaussian mixture distribution conditioned on the feature values. This is depicted in Figure 3-2. This is similar in spirit to the models of [DRS97, RBD00], except that we factor the state into multiple streams for explicit modeling of asynchrony and substitution. In addition, we allow for asynchrony between features throughout the course of a word, while [DRS97, RBD00] require that features synchronize at the target value for each sub-word unit (phone or bi-phone).

It is likely that different features affect different aspects of the acoustic signal; for example, features related to degrees of constriction may be closely associated with the amplitude and degree of noise throughout the signal, whereas nasality may affect mainly the lower frequencies. For this reason, it may be useful to extract different acoustic measurements for different features, as in Figure 3-3.

Figure 3-3 also describes a related scenario in which separate classifiers are independently trained for the various features, whose outputs are then converted (perhaps heuristically) to scaled likelihoods, $\propto p(obs_t^i|s^i)$, for use in a generative ASR system. As mentioned in Chapter 1, this has been suggested as a promising approach for feature-based ASR because of the more efficient use of training data and apparent

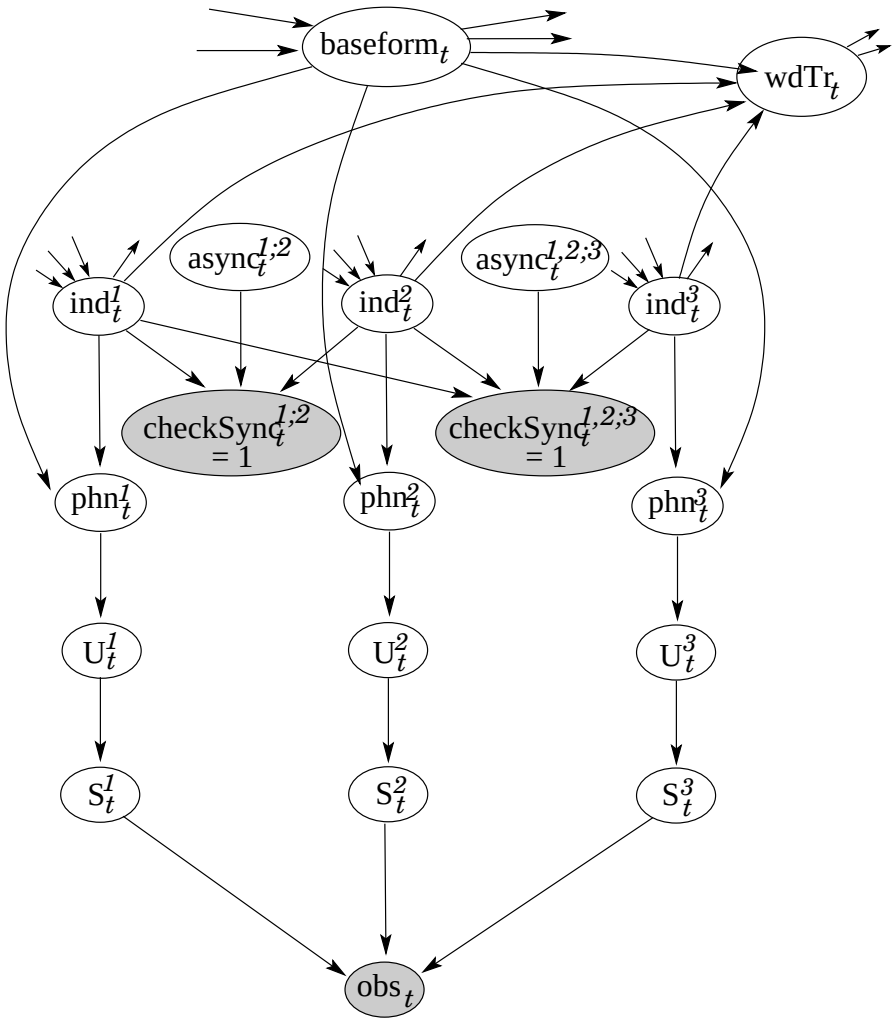

Figure 3-2: *One way of integrating the pronunciation model with acoustic observations.*

robustness to noise [KFS00]. One of our goals is to provide a means to combine the information from the feature classifier outputs in such a system without making overly strong assumptions about the faithful production of baseform pronunciations. Scaled likelihoods can be incorporated via the mechanism of *soft* or *virtual* evidence [Pea88, Bil04]. In the language of DBNs, this can be represented by letting each obs_t^i be a binary variable observed with constant value (say 1), and setting the CPT of obs_t^i to

$$p(obs_t^i = 1|s_t^i) = Cp(obs_t^i|s^i), \tag{3.4}$$

where C is any scaling constant. This is identical to the mechanism used in hybrid hidden Markov model/artificial neural network (HMM/ANN) ASR systems [BM94], in which a neural network is trained to classify phones, and its output is converted to a scaled likelihood for use in an HMM-based recognizer. We give examples of systems using such an approach in Chapters 5 and 6.

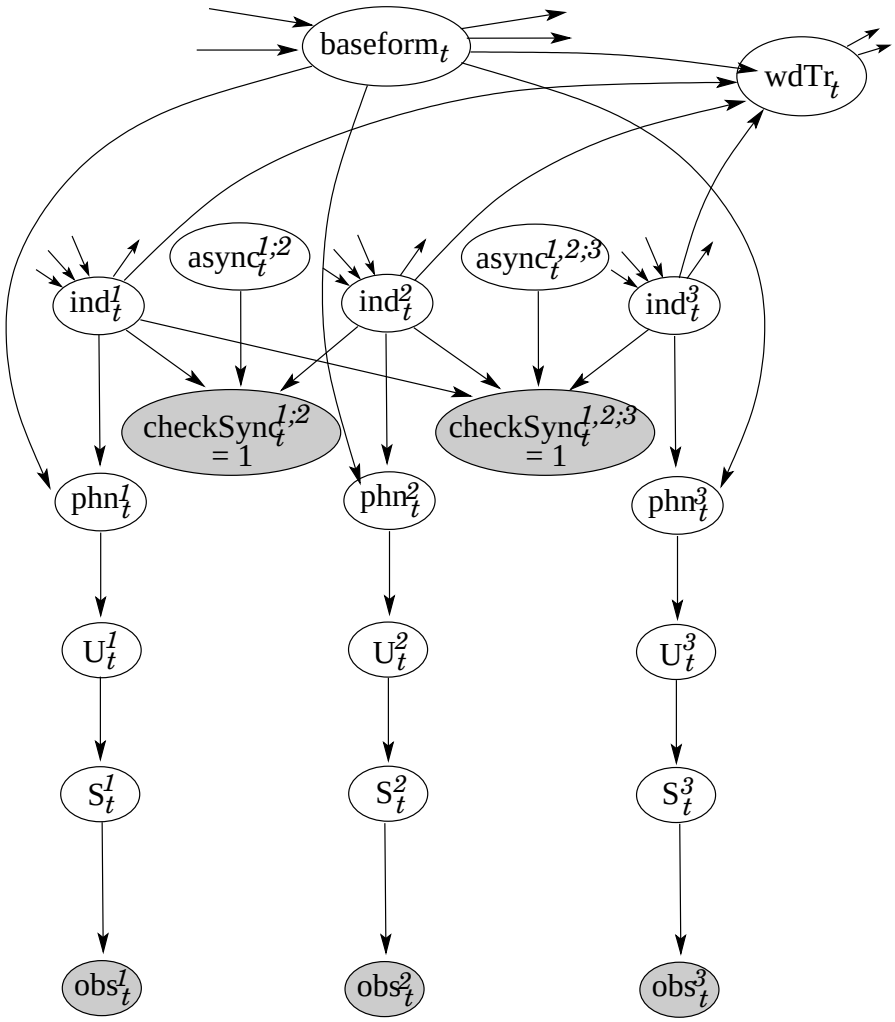

Figure 3-3: *One way of integrating the pronunciation model with acoustic observations.*

Finally, we consider the case where we wish to use a set of feature classifiers or feature-specific observations for a *different* set of features than is used in the pronunciation model. We have not placed any constraints on the choice of feature set used in our model, and in fact for various reasons we may wish to use features that are not necessarily the most acoustically salient. On the other hand, for the feature-acoustics interface, we may prefer a more acoustically salient feature set. As long as there is a mapping from the feature set of the pronunciation model to that of the acoustic model, we can construct such a system as shown in Figure 3-4. We give an example of this type of system in Chapter 5.

3.5 Relation to previous work

As mentioned previously, the idea of predicting the allowed realizations of a word by modeling the evolution of feature streams (whether articulatory or more abstract) is not new. Dynamic Bayesian networks and their use in ASR are also not new, although the particular class of DBN we propose is. To our knowledge, however, this is the first *computationally implemented* feature-based approach to pronunciation modeling suitable for use in an ASR system. We now briefly describe the relation of this approach to some of the previous work mentioned in Chapter 2.

3.5.1 Linguistics and speech science

The class of models presented in this chapter are inspired by, and share some characteristics with, previous work in linguistics and speech science. Most closely related is the articulatory phonology of Browman and Goldstein [BG92]: Our use of asynchrony is analogous to their gestural overlap, and feature substitution is a generalization of gestural reduction. Although substitutions can, in principle, include not only reductions but also increases in gesture magnitude, we will usually constrain substitutions to those that correspond to reductions (see Chapter 4.

The current approach differs from work in linguistics and speech science in its aim: Although we are motivated by an understanding of the human mechanisms of speech production and perception, our immediate goal is the applicability of the approach to the problem of automatic speech recognition. Our models, therefore, are not tuned to match human perception in terms of such measures as types of errors made and relative processing time for different utterances. We may also choose to omit certain details of speech production when they are deemed not to produce a difference in recognition performance.

Because of this difference in goals, the current work also differs from previous linguistic and speech science proposals in that it must (a) provide a *complete* representation of the lexicon, and (b) have a *computational implementation*. For example, we draw heavily on ideas from articulatory phonology (AP) [BG92]. However, to our knowledge, there has to date been no reasonably-sized lexicon represented in terms of articulatory gestures in the literature on articulatory phonology, nor is there a description of a mechanism for generating such a lexicon from existing lexica. We are also not aware of a description of the set of articulatory gestures necessary for a complete articulatory phonology, nor a computational implementation of AP allowing for the recognition of words from their surface forms. For our purposes, we must generate an explicit, complete feature set and lexicon, and a testable implementation of the model. In some cases, we must make design choices for which there is little support in the scientific literature, but which are necessary for a complete working model. A particular feature set and phone-to-feature mapping that we have developed for use in experiments are described in Chapter 4 and Appendix B. It is hoped that the model, feature sets, and phone-to-feature mappings can be refined as additional data and theoretical understandings become available.

There are also similarities between our approach and Fant's *microsegments* model [Fan73],

in which features may change values asynchronously and a new segment is defined each time a feature changes value. Our vectors of surface feature values can be viewed as microsegments. The key innovation is, again, the introduction of a framework for performing computations and taking advantage of inter-feature independencies.

3.5.2 Automatic speech recognition

In the field of automatic speech recognition, a fair amount of research has been devoted to the classification of sub-phonetic features from the acoustic signal, or to the modeling of the signal in terms of features; in other words, to the problem of feature-based *acoustic observation modeling*. This has been done both in isolation, as a problem in its own right (e.g., [WFK04]), and as part of complete ASR systems [KFS00, Eid01, MW02]. However, the problem of feature-based *pronunciation modeling* has largely been ignored. In complete ASR systems using feature-based acoustic models, the typical approach is to assume that the lexicon is represented in terms of phonemes, and that features will evolve synchronously and take on the canonical values corresponding to those phonemes.

A natural comparison is to the work of Deng and colleagues [DRS97], Richardson and Bilmes [RBD00], and Kirchhoff [Kir96]. In [DRS97] and [RBD00], multiple feature streams are "compiled" into a single HMM with a much larger state space. This results in data sparseness issues, as many states are seen very rarely in training data. These approaches, therefore, do not take advantage of the (conditional) independence properties between features. In addition, as previously mentioned, both [DRS97] and [RBD00] assume that features synchronize at the target configuration for each sub-word unit. This is quite a strong assumption, as common pronunciation phenomena often involve asynchrony across a larger span. In [Kir96], on the other hand, features are allowed to desynchronize arbitrarily within syllables, and must synchronize at syllable boundaries. This approach takes greater advantage of the independent nature of the features, but assumes that all degrees of asynchrony within a syllable are equivalent. In addition, there are many circumstances in which features do not synchronize at syllable boundaries.

3.5.3 Related computational models

Graphical model structures with multiple hidden streams have been used in various settings. Ghahramani and Jordan introduced factorial HMMs and used speech recognition as a sample application [GJ97]. Logan and Moreno [LM98] used factorial HMMs for acoustic modeling. Nock and Young [NY02] developed a general architecture for modeling multiple asynchronous state streams with coupled HMMs and applied it to the fusion of multiple acoustic observation vectors. Factorial HMMs, and related multistream HMM-type models, have received particularly widespread application in the literature on multi-band HMMs [DFA03, ZDH+03], as well as in audio-visual speech recognition [NLP+02, GSBB04]. Our approach is most similar to coupled HMMs; the main differences are the more explicit modeling of asynchrony between streams and the addition of substitutions.

3.6 Summary and discussion

This chapter has presented a general and flexible model of the evolution of multiple feature streams for use in modeling pronunciation variation for ASR. Some points bear repeating, and some bear further examination:

- The only processes generating pronunciation variants in our approach are inter-feature asynchrony and per-feature substitution. In particular, we have not included deletions or insertions of feature values. This is in keeping with articulatory phonology, the linguistic theory most closely related to our approach. It would be straightforward to incorporate deletions and insertions into the model. However, this would increase the complexity (i.e., the number of parameters) in the model, and based on our experiments thus far (see Chapter 4), there is no clear evidence that insertions or deletions are needed.

- Our approach does not assume a particular feature set, although certain feature sets may be more or less suitable in such a model. In particular, the features should obey the properties of conditional independence assumed by the DBN. For example, our model would not be appropriate for binary feature systems of the kind used by Stevens [Ste02] or Eide [Eid01]. Such feature sets are characterized by a great deal of dependence between feature values; in many cases, one feature value is given deterministically by other feature values. While it may be worthwhile to add some feature dependencies into our model, the level that would be required for this type of feature set suggests that they would be better modeled in a different way.

- We do not require that the features used in the pronunciation model be used in the acoustic observation model as well, as long as there is an information-preserving mapping between the feature sets (see the discussion of Figure 3-4). This is important in the context of previous work on feature classification, which has typically concentrated on more acoustically-motivated features which may not be the best choice for pronunciation modeling. We are therefore free to use whatever feature sets best account for the pronunciation variation seen in data.

- We do not claim that all pronunciation variation is covered by the model. We leave open the possibility that some phenomena may be related directly to the underlying phoneme string, and may not be the result of asynchrony between features or substitutions of individual feature values. For now we assume that any such variation is represented in the baseform dictionary.

- So far, we have only dealt with words one at a time, and assumed that features synchronize at word boundaries. We know that this assumption does not hold, for example, in *green beans* → [g r iy m b iy n z]. This is a simplifying assumption from a computational perspective, and one that should be re-examined in future work.

In the following chapter, we implement a specific model using an articulatory feature set and investigate its behavior on a corpus of manually transcribed data.

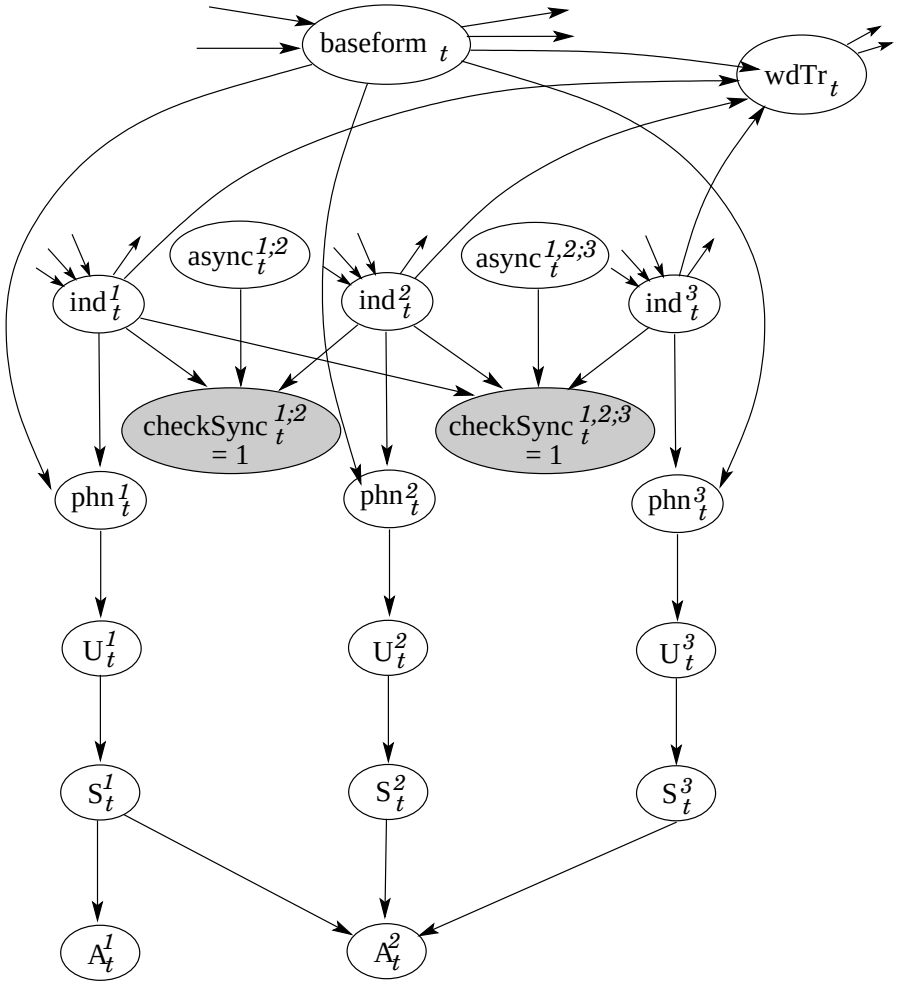

Figure 3-4: *One way of integrating the pronunciation model with acoustic observations, using different feature sets for pronunciation modeling (S^i) and acoustic observation modeling (A^i). In this example, A^1 is a function of S^1 and A^2 is a function of S^1, S^2, and S^3. As this example shows, the two feature sets are not constrained to have the same numbers of features.*

Chapter 4

Lexical Access Experiments Using Manual Transcriptions

The previous chapter introduced the main ideas of feature-based pronunciation modeling and the class of DBNs that we propose for implementing such models. In order to use such a model, design decisions need to be made regarding the baseform dictionary, feature set, and synchrony constraints. In this chapter, we present an implemented model, propose a particular feature set, and study its behavior on a lexical access task. We describe experiments performed to test the model in isolation. In order to avoid confounding pronunciation modeling successes and failures with those of the acoustic or language models, we test the model on an isolated-word recognition task in which the surface feature values are given. This gives us an opportunity to experiment with varying model settings in a controlled environment.

4.1 Feature sets

In Chapter 2 we discussed several types of sub-phonetic feature sets used in previous research. There is no standard feature set used in feature-based ASR research of which we are aware. The most common features used in research on feature-based acoustic observation modeling or acoustics-to-feature classification are based on the categories used in the International Phonetic Alphabet [Alb58] to distinguish phones. Table 4.1 shows such a feature set. *Nil* values are used when a feature does not apply; for example, **front/back** and **height** are used only for vowels (i.e., only when **manner** = *vowel*), while **place** is used only for consonants. The value *sil* is typically included for use in silence portions of the signal (a category that does not appear in the IPA). The state space of this feature set–the total number of combinations of feature values–is 7560 (although note that many combinations are disallowed, since some feature values preclude others).

Our first instinct might be to use this type of feature set in our model, so as to ensure a good match with work in acoustic observation modeling. Using this feature set with our model, it is easy to account for such effects as vowel nasalization, as in *don't* → [d ow_n n t], and devoicing, as in *from* → [f r_vl ah m], by allowing for

feature	**values**
front/back (F)	nil, front, back, mid, sil
height (H)	nil, high, mid, low, sil
manner (M)	vowel, stop release, stop, fricative, approximant, lateral, nasal, sil
nasalization (N)	non-nasal, nasal, sil
place (P)	nil, labial/labiodental, dental/alveolar, post-alveolar, velar, glottal, sil
rounding (R)	non-round, round, nil
voicing (V)	voiced, voiceless, sil

Table 4.1: *A feature set based on IPA categories.*

asynchronous onset of nasality or voicing with respect to other features. However, for many types of pronunciation variation, this feature set seems ill-suited to the task. We now re-examine a few examples that demonstrate this point.

One type of effect that does not seem to have a good explanation in terms of asynchrony and substitutions of IPA-style features is stop insertion, as in *sense* → [s eh n t s]. Part of the explanation would be that the **place** feature lags behind **voicing** and **nasalization**, resulting in a segment with the **place** of an [n] but the voicing/nasality of an [s]. However, in order to account for the **manner** of the [t], we would need to assume that either (i) part of the /n/ has had its **manner** substituted from a nasal to a stop, or (ii) part of the /s/ has had its **manner** substituted from a fricative to a stop. Alternatively, we could explicitly allow insertions in the model. In contrast, we saw in Chapter 1 that such examples can be handled using asynchrony alone.

Another type of effect 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, however, 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*, but in addition, the features **front/back** and **height** would change from *nil* to the appropriate values.

Motivated by such examples, we propose a feature set based on the vocal tract variables of Browman and Goldstein's articulatory phonology (AP) [BG92]. We have informally used this type of features in examples in Chapters 1 and 3. We formalize the feature set in Table 4.2. These features refer to the locations and degrees of constriction of the major articulators in the vocal tract, discussed in Chapter 2 and shown in Figure 4-1. The meanings of the feature values are given in Table B.1 of Appendix B, and the mapping from phones to features in Table B.2. The state space of this feature set consists of 41,472 combinations of feature values.

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 are 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,

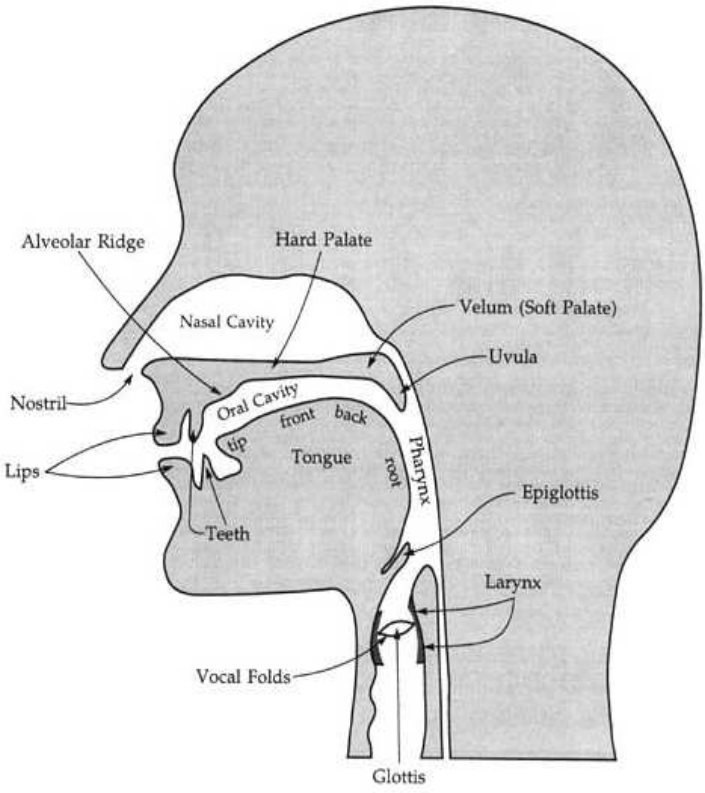

Figure 4-1: *A midsagittal section showing the major articulators of the vocal tract, reproduced from Chapter 2.*

while differentiating between as many phones as possible. In constructing phone-to-feature mappings, we have consulted the articulatory phonology literature (in particular, [BG86, BG89, BG90a, BG90b, BG92]), phonetics literature ([Lad01, Ste98]), and X-ray tracings of speech articulation [Per69].

4.2 Models

4.2.1 Articulatory phonology-based models

Figure 4-2 shows the structure of the articulatory phonology-based model used in our experiments. The structure of synchrony constraints is based on linguistic considerations. First, we make the assumption that the pairs **TT-LOC**, **TT-OPEN**; **TB-LOC**, **TB-OPEN**; and **VELUM**, **GLOTTIS** are always synchronized; for this reason we use single variables for the tongue tip index ind_t^{TT}, tongue body index ind_t^{TB}, and glottis/velum index ind_t^{GV}. We base this decision on the lack of evidence of which we are aware for pronunciation variation that can be explained by asynchrony among these pairs. We impose a soft synchrony constraint on **TT** and **TB**, implemented using $async_t^{TT;TB}$. Another constraint is placed on the lips vs. tongue,

feature	values
LIP-LOC (LL)	protruded, labial, dental
LIP-OPEN (LO)	closed, critical, narrow, wide
TT-LOC (TTL)	inter-dental, alveolar, palato-alveolar, retroflex
TT-OPEN (TTO)	closed, critical, narrow, mid-narrow, mid, wide
TB-LOC (TBL)	palatal, velar, uvular, pharyngeal
TB-OPEN (TBO)	closed, critical, narrow, mid-narrow, mid, wide
VELUM (V)	closed, open
GLOTTIS (G)	closed, critical, wide

Table 4.2: *A feature set based on the vocal tract variables of articulatory phonology.*

using $async_t^{LO;TT,TB}$. Asynchrony between these features is intended to account for such effects as vowel rounding before a labial consonant. We are not using **LIP-LOC** in these experiments. This helps to reduce computational requirements, and should not have a large impact on performance since there are very few words in our vocabulary distinguished solely by **LIP-LOC**; this reduces the number of feature value combinations to 13,824. The last soft synchrony constraint is between the lips and tongue on the one hand and the glottis and velum on the other, controlled by $async^{LO,TT,TB;V,G}$. Asynchrony between these two sets of features is intended to allow for effects such as vowel nasalization, stop insertion in a nasal context, and some nasal deletions. The *checkSync* variables are therefore given as follows (refer to Eq. 3.3):

$$checkSync_t^{TT;TB}{=}1 \iff |ind_t^{TT} - ind_t^{TB}| {=} async_t^{TT;TB}$$

$$checkSync_t^{LO;TT,TB}{=}1 \iff round\Big(|ind_t^{LO} - \frac{ind_t^{TT}+ind_t^{TB}}{2}|\Big) {=} async_t^{LO;TT,TB}$$

$$checkSync_t^{LO,TT,TB;V,G}{=}1 \iff round\Big(|\frac{ind_t^{LO}+ind_t^{TT}+ind_t^{TB}}{3} - \frac{ind_t^{V}+ind_t^{G}}{2}|\Big) {=} async_t^{LO,TT,TB;V,G}$$

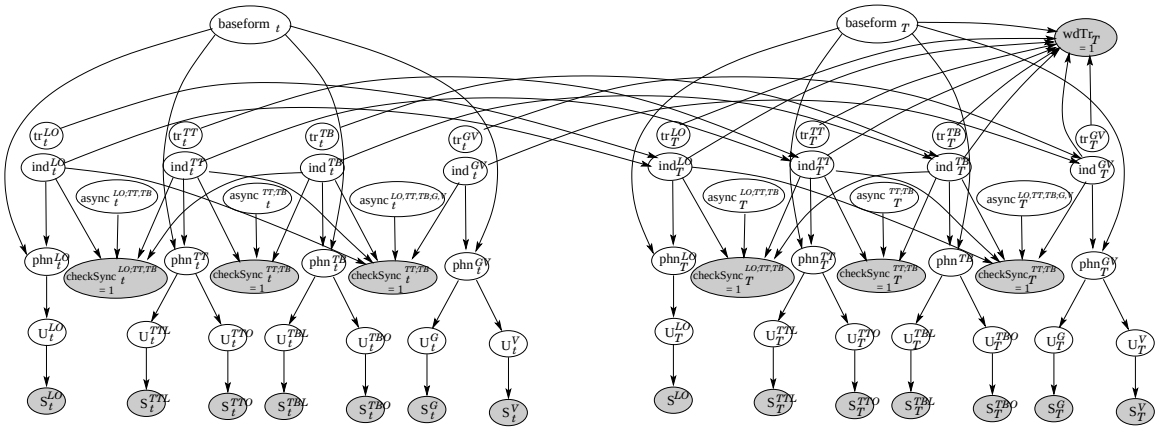

Figure 4-2: *An articulatory phonology-based model.*

4.2.2 Phone-based baselines

A phone-based model can be considered a special case of a feature-based one, where the features are constrained to be completely synchronized and no substitutions are allowed.[1] We consider two baselines, one using the same baseform dictionary as the feature-based models and one using a much larger set of baseforms, generated by applying to the baseforms a phonological rule set developed for the MIT SUMMIT recognition system [HHSL05].

4.3 Data

The data sets for these experiments are drawn from the Switchboard corpus of conversational speech [GHM92]. This corpus consists of 5-10 minute telephone conversations between randomly matched pairs of adult speakers of American English of various geographic origins within the United States. Each conversation revolves around an assigned topic, such as television shows or professional dress codes.

A small portion of this database was manually transcribed at a detailed phonetic level at the International Computer Science Institute, UC Berkeley [GHE96]. This portion consists of 72 minutes of speech, including 1741 utterances (sentences or phrases) spoken by 370 speakers drawn from 618 conversations.[2] The speakers are fairly balanced across age group and dialect region. The transcriptions were done using a version of the ARPABET phonetic alphabet [Sho80], modified to include diacritics indicating nasalization, frication (of a normally un-fricated segment), creaky voice, and several other phenomena. Appendix A describes the label set more fully. Greenberg et al. report an inter-transcriber agreement rate between 72% and 80% [Gre99], and Saraçlar reports the rate at 75.3% after mapping the labels to a smaller standard phone set [Sar00]. We acknowledge this disadvantage of using these transcriptions as ground truth but nevertheless find them useful as a source of information on the types of variation seen in real speech. The examples in Chapter 1 were drawn from these transcriptions, which we will henceforth refer to as the ICSI transcriptions.

For the experiments in this chapter, we use a 3328-word vocabulary, consisting of the 3500 most likely words in the "Switchboard I" training set [GHM92], excluding partial words, non-speech, and words for which we did not have baseform pronunciations. This is much smaller than the full Switchboard vocabulary (of roughly 20,000–30,000 words), but facilitates quick experimentation. All of our experiments have been done on the "train-ws96-i" subset of the ICSI transcriptions. We use the transcribed words in subsets 24–49 as training data; subset 20 as a held-out develop-

[1] There are two slight differences between this and a conventional phone-based model: (i) The multiple transition variables mean that we are counting the same transition probability multiple times, and (ii) when the U_t^i are not deterministic, there can be some added variability on a frame-by-frame basis. Our phone-to-feature mapping (see Table B.2 in Appendix B) is mostly deterministic. In any case, as the results will show, these details make little difference to the baseline performance.

[2] About three additional hours of speech were also phonetically transcribed and manually aligned at the syllable level, but the phonetic alignments were done by machine.

ment set; and subsets 21–22 as a final test set. The development set is used for tuning aspects of the model, whereas the test set is never looked at (neither the transcriptions nor the correct words). In addition, we manually corrected several errors in the development set transcriptions due to misalignments with the word transcriptions. For all three sets, we exclude 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).[3] 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 training set contains 2942 words, the development set contains 165, and the test set contains 236.

We prepared the data as follows. Each utterance comes with time-aligned word and phone transcriptions. For each transcribed word, we extracted the portion of the phonetic transcription corresponding to it by aligning the word and phone time stamps. The marked word boundaries sometimes fall between phone boundaries. In such cases, we considered a phone to be part of a word's transcription if at least 10ms of the phone is within the word boundaries. In addition, we collapsed the phone labels down to a simpler phone set, eliminating diacritics other than nasalization.[4] Finally, we split stops, affricates, and diphthongs into two segments each, assigning 2/3 of the original segment's duration to the first segment and the latter 1/3 to the second.

4.4 Experiments

The setup for the lexical access experiments is shown in Figure 4-3. The question being addressed is: Supposing we had knowledge of the true sequences of surface feature values $S_t^i \forall i, t$ 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 ICSI phonetic transcriptions, by assuming a deterministic mapping from surface phones to surface features values. [5] Recognition then consists of introducing these surface feature values as observations of $S = S_t^i \forall i, t$ in the DBN for each word, and computing the posterior probability of the word,

$$p(wd_j|S), 1 \leq j \leq V, \tag{4.1}$$

where V is the vocabulary size. Figure 4-3 shows the few most likely words hypothesized for *cents* transcribed as *s ah_n n t s*, along with their log probabilities, in rank order. The recognized word is the one that maximizes this posterior probability, in

[3]This means that, of the four example words considered in Chapter 1, *sense* is excluded while the remaining three are included.

[4]This was done for two reasons: First, we felt that there was somewhat less consistency in the labeling of some of the more unusual phones; and second, this allows us to use the same transcriptions in testing the baseline and proposed systems. In the future, we would like to return to this point and attempt to take better advantage of the details in the transcriptions.

[5]This mapping is similar, but not identical, to the one used for $p(U_t^i|ph_t^i)$ in the DBN; it is deterministic and contains some extra phones found in the transcriptions but not in the baseforms.

this case *cents*. In general, depending on the model, many of the words in the vocabulary may have zero posterior probability for a given S, i.e. the model considers S to not be an allowed realization of those words.

Maximum likelihood parameter learning is done using the EM algorithm, given the training set of observed word/S pairs. All DBN inference and parameter learning is done using the Graphical Models Toolkit (GMTK) [BZ02, GMT]. For these experiments, we use exact inference; this is feasible as long as the probability tables in the DBN are sparse (as we assume them to be), and it avoids the question of whether differences in performance are caused by the models themselves or by approximations in the inference.

We will mainly report two measures of performance. The *coverage* measures how well a model predicts the *allowable* realizations of a word; it is measured as the proportion of a test set for which a model gives non-zero probability to the correct word. The *accuracy* is simply the classification rate on a given test set. We say that a given surface pronunciation is *covered* by a model if the model gives non-zero probability to that pronunciation for the correct word. The coverage of a model on a given set is an upper bound on the accuracy: A word cannot be recognized correctly if the observed feature values are not an allowed realization of the word. Arbitrarily high coverage can trivially be obtained by giving some positive probability to all possible S for every word. However, this is expected to come at a cost of reduced accuracy due to the added confusability between words.

For a more detailed look, it is also informative to consider the ranks and probabilities themselves. 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 recognition scenario, confusable words may be disambiguated based on the linguistic and phonetic context (in the case of connected speech recognition). 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.

As the first two lines of Table 4.3 show, this task is not trivial: The baseforms-only model (line (1)), which has on average 1.7 pronunciations per word, has a coverage of only 49.7% on the development set and 40.7% on the test set. All of the words that are covered by the baseline model are recognized correctly; that is, the coverage and accuracy are identical. This is not surprising: The canonical pronunciations of words rarely appear in this database. A somewhat more surprising result is that expanding the baseforms with a large bank of phonological rules, giving a dictionary with up to 96 pronunciations per word (3.6 on average), increases the coverage to only 52.1%/44.5% (line (2)). The phonological rules improve both coverage and accuracy, but they do not capture many of the types of variation seen in this conversational data set.

We next ask how much we could improve performance by tailoring the dictionary to the task. If we create a dictionary combining the baseline dictionary with all pronunciations in the training set (subsets 24–29 of the transcriptions), we obtain the

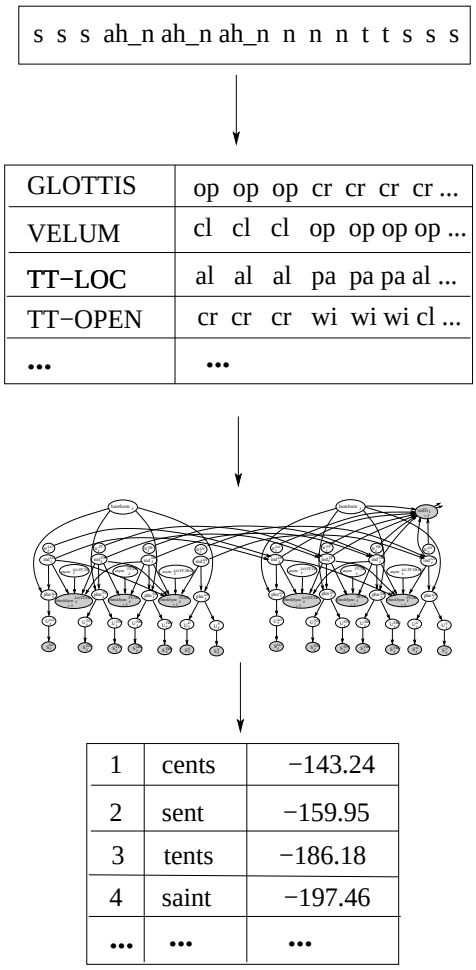

Figure 4-3: *Experimental setup.*

much-improved performance shown in line (3) of Table 4.3[6]. If we could include the pronunciations in the test set (line (4)), we would obtain almost perfect coverage[7] and accuracy of 89.7%/83.9%. This is a "cheating" experiment in that we do not in general have access to the pronunciations in the test set.

We next trained and tested an AP feature-based model, with the following hard constraints on asynchrony:

[6]The baseline and training dictionaries were combined, rather than using the training pronunciations alone, to account for test words that do not appear in the training data.

[7]On the development set, one token (namely, *number* pronounced [n ah_n m b er]) is not covered, although the phonetic pronunciation does (by definition) appear in the dictionary. This is because the duration of the transcribed [b] segment, after chunking into frames, is only one frame. Since the dictionary requires both a closure and a burst for stops, each of which must be at least one frame long, this transcription cannot be aligned with the dictionary. This could be solved by making the dictionary sensitive to duration information, including both a closure and a burst only when a stop is sufficiently long.

model	dev set		test set	
	coverage	**accuracy**	**coverage**	**accuracy**
(1) baseforms only	49.7	49.7	40.7	40.7
(2) + phonological rules	52.1	52.1	44.5	43.6
(3) all training pronunciations	72.7	64.8	66.1	53.8
(4) + test pronunciations ("cheating dictionary")	99.4	89.7	100.0	83.9
(5) AP feat-based, init 1 (knowledge-based)	83.0	73.3	75.4	60.6
(6) + EM	83.0	73.9	75.4	61.0
(7) AP feat-based, init 2 ("sparse flat")	83.0	27.9	75.4	23.7
(8) + EM	83.0	73.9	75.4	61.4
(9) async only	49.7	49.1	42.4	41.5
(10) subs only	75.8	67.3	69.9	57.2
(11) IPA feat-based	63.0	56.4	56.8	49.2
(12) + EM	62.4	57.6	55.9	50.0

Table 4.3: *Results of Switchboard ranking experiment. Coverage and accuracy are percentages.*

1. All four tongue features are completely synchronized,

$$async_t^{TT;TB} = 0 \tag{4.2}$$

2. The lips can desynchronize from the tongue by up to one index value,

$$p(async_t^{LO;TT,TB} > 1) = 0 \tag{4.3}$$

   This is intended to 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*.

3. The glottis/velum index must be within 2 of the mean index of the tongue and lips,

$$p(async_t^{LO,TT,TB;GV} > 2) = 0 \tag{4.4}$$

   This accounts for the typically longer-distance effects of nasalization, as in *trying* → [t r ay_n n].

In addition, we set many of the substitution probabilities to zero, based on the assumption that location features will not stray too far from their intended values, and that constriction degrees may be reduced from more constricted to less constricted but generally not vice versa. These synchrony and substitution constraints are based on both articulatory considerations and trial-and-error testing on the development set.

In order to get a sense of whether the model is behaving reasonably, we can look at the most likely settings for the hidden variables given a word and its surface realization S, 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 "came about". Figures 4-4 and 4-5 show spectrograms and the most likely sequences of a subset of the DBN variables for two example words from the development set, *everybody* → [eh r uw ay] and *instruments* → [ih_n s tcl ch em ih_n n s], computed using the model described above.[8] Multiple frames with identical variable values have been merged for ease of viewing.

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. Lacking access to the speaker's intentions, we cannot be sure of the correct analysis; however, this analysis seems like a reasonable one given the phonetic transcription.

Turning to the example of *instruments*, the apparent deletion of the first [n] and nasalization of both [ih]s is, as expected from the discussion in Chapter 1, 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.

In setting initial parameter values for EM training, we assumed that values closer to canonical–lower values of the *async* variables and values of S_t^i similar to U_t^i–are preferable, and set the initial parameters accordingly. We refer to this initialization as the "knowledge-based" initialization. Tables 4.4–4.6 show some of the conditional probability tables (CPTs) used for initializing EM training, and Tables 4.7–4.9 show the learned CPTs for the same variables. We note that some of the training examples necessarily received zero probability (due to the zeros in the CPTs) and therefore were not used in training. Of the 2942 training words, 688 received zero probability.

	closed	critical	narrow	wide
closed	0.95	0.04	0.01	0
critical	0	0.95	0.04	0.01
narrow	0	0	0.95	0.05
wide	0	0	0	1

Table 4.4: *Initial CPT for* **LIP-OPEN** *substitutions,* $p(S^{LO}|U^{LO})$. S^{LO} *values correspond to columns,* U^{LO} *values to rows.*

[8] In looking at the spectrograms, we might argue that these are not the best phonetic transcriptions for these examples: The [uw] in *everybody* might be more [w]-like, and the [ch] of *instruments* might be labeled as a retroflexed [t] in a finer-grained transcription. However, we still have intuitions about what constitutes a good analysis of these transcriptions, so that it is instructive to consider the analyses produced by the model.

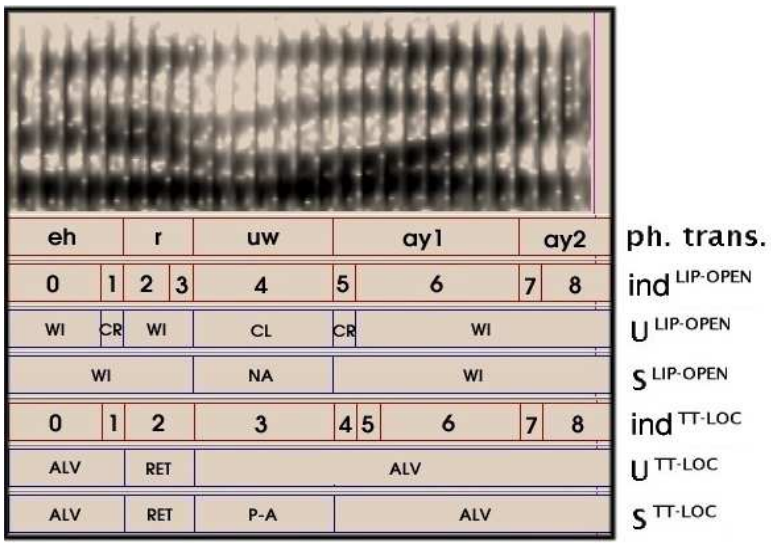

Figure 4-4: *Spectrogram, phonetic transcription, and partial alignment, including the variables corresponding to* **LIP-OPEN** *and* **TT-LOC**, *for the example* everybody → *[eh r uw ay]. Indices are relative to the underlying pronunciation /eh v r iy bcl b ah dx iy/. Adjacent frames with identical variable values have been merged for easier viewing. Abbreviations used are:* **WI** = wide*;* **NA** = narrow*;* **CR** = critical*;* **CL** = closed*;* **ALV** = alveolar*;* **P-A** = palato-alveolar*;* **RET** = retroflex.

Lines (5) and (6) of Table 4.3 show the coverage and accuracy of this model using both the initial and the trained parameters. We first note that coverage greatly increases relative to the baseline models as expected, since we are allowing vastly more pronunciations per word. As we previously noted, however, increased coverage comes with the danger of increased confusability and therefore lower accuracy. Encouragingly, the accuracy also increases relative to the baseline models. Furthermore, all of the development set words correctly recognized by the baseforms-only baseline are also correctly recognized by the feature-based model. Compared to the baseforms + rules model, however, two words are no longer correctly recognized: *twenty* → [t w eh n iy] and *favorite* → [f ey v er t]. These examples point out shortcomings of our current model and feature set; we will return to them in Section 4.5.

It is interesting to note that the accuracy does not change appreciably after training; the difference in accuracy is not significant according to McNemar's test [Die98]. (Note that coverage cannot increase as a result of training, since it is determined entirely by the locations of zeros in the CPTs.) This might make us wonder whether the magnitudes of the probabilities in the CPTs make any difference; perhaps it is the case that for this task, it is possible to capture the transcribed pronunciations simply by adding some "pronunciation noise" to each word, without increasing confusability. In other words, perhaps the only factor of importance is the locations of zeros in the CPTs (i.e. what is possible vs. impossible). To test this hypothesis, we tested the model with a different set of initial CPTs, this time having the same zero/non-zero

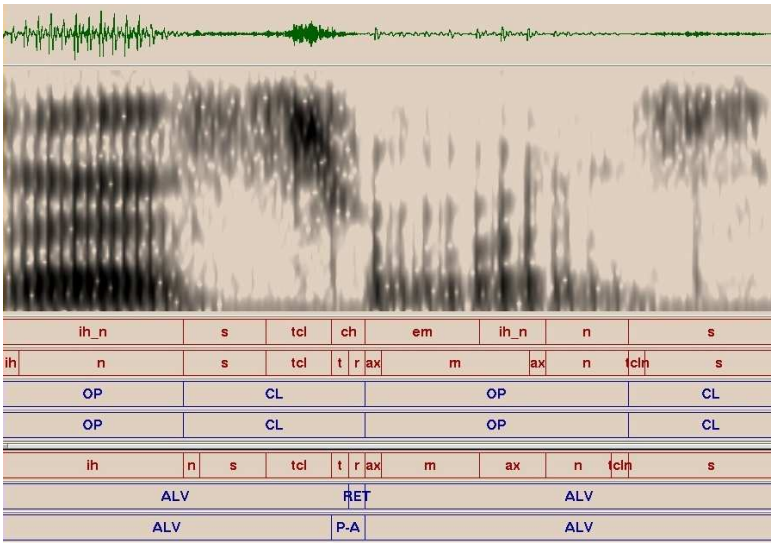

Figure 4-5: *Spectrogram, phonetic transcription, and partial alignment, including the variables corresponding to* **VELUM** *and* **TT-LOC**, *for the example* instruments → *[ih_n s tcl ch em ih_n n s].*

	interdental	alveolar	palato-alveolar	retroflex
interdental	0.95	0.05	0	0
alveolar	0.025	0.95	0.025	0
palato-alveolar	0	0.05	0.95	0
retroflex	0	0.01	0.04	0.95

Table 4.5: *Initial CPT for* **TT-LOC** *substitutions,* $p(S^{TTL}|U^{TTL})$. S^{TTL} *values correspond to columns,* U^{TTL} *values to rows.*

structure as the knowledge-based initialization, but with uniform probabilities over the non-zero values. We refer to this as the "sparse flat" initialization. In addition, to test the sensitivity of the parameter learning to initial conditions, we also re-trained the model using this new initialization. The results are shown in lines (7) and (8) of Table 4.3.

The coverage is again trivially the same as before. The accuracies, however, are quite poor when using the initial model, indicating that the magnitudes of the probabilities are indeed important. After training, the performance is the same as or better than when using the knowledge-based initialization, indicating that we need not be as careful with the initialization.

The coverage and accuracy do not give the full picture, however. In an end-to-end recognizer, the model's scores would be combined with the language and observation model scores. Therefore, it is important that, if the correct word is not top-ranked, its rank is as high as possible, and that the correct word scores as well as possible relative to competing words. Figure 4-6 shows the empirical cumulative distribution

async degree	0	1	2	3+
TT;TB	1	0	0	0
LO;TT,TB	0.67	0.33	0	0
LO,TT,TB;G,V	0.6	0.3	0.1	0

Table 4.6: *Initial CPTs for the asynchrony variables.*

	closed	critical	narrow	wide
closed	0.999	8.2×10^{-4}	2.7×10^{-4}	0
critical	0	0.77	0	2.3×10^{-1}
narrow	0	0	0.98	1.9×10^{-2}
wide	0	0	0	1

Table 4.7: *Learned CPT for* **LIP-OPEN** *substitutions,* $p(S^{LO}|U^{LO})$. S^{LO} *values correspond to columns,* U^{LO} *values to rows.*

functions of the correct word's rank for the test set, using the frame-based model with both initializations, before and after training. Figure 4-7 shows the cumulative distributions of the *score margin*–the difference in log probability between the correct word and its highest-scoring competitor–in the same conditions. The score margin is positive when a word is correctly recognized and negative otherwise. Since the correct word's score should be as far as possible from its competitors, we would like this curve to be as flat as possible. These plots show that, although the accuracy does not change after training when using the knowledge-based initialization, the ranks and score margins do improve. The difference in both the rank distributions and score margin distributions is statistically significant on the test set (according to a paired t-test [Die98]). On the development set, however, only the score margin differences are significant.

Next, we ask what the separate effects of asynchrony and substitutions are. Lines (9) and (10) of Table 4.3 show the results of using only substitutions (setting the asynchrony probabilities to zero) and only asynchrony (setting the off-diagonal values in the substitution CPTs to zero). In both cases, the results correspond to the models after EM training using the knowledge-based initialization, except for the additional zero probabilities. The asynchrony-only results are identical to the phone baseline, while the substitution-only performance is much better. This indicates that virtually all non-baseform productions in this set include some substitutions, or else that more asynchrony is needed. Anecdotally, looking at examples in the development set, we believe the former to be the case: Most examples contain some small amount of substitution, such as /ah/ $\rightarrow$ [ax]. However, asynchrony is certainly needed, as evidenced by the improvement from the substitution-only case to the asynchrony + substitution case; the improvement in accuracy is significant according to McNemar's test ($p = .003/.008$ for the dev/test set). Looking more closely at the performance on the development set, many of the tokens on which the synchronous models failed

	interdental	alveolar	palato-alveolar	retroflex
interdental	0.98	2.1×10^{-2}	0	0
alveolar	9.7×10^{-4}	0.99	1.1×10^{-2}	0
palato-alveolar	0	1.5×10^{-2}	0.98	0
retroflex	0	1.1×10^{-2}	4.0×10^{-3}	0.99

Table 4.8: *Learned CPT for* **TT-LOC** *substitutions,* $p(S^{TTL}|U^{TTL})$. S^{TTL} *values correspond to columns,* U^{TTL} *values to rows.*

async degree	0	1	2	3+
TT;TB	1	0	0	0
LO;TT,TB	0.996	4.0×10^{-3}	0	0
LO,TT,TB;G,V	0.985	1.5×10^{-2}	5.3×10^{-5}	0

Table 4.9: *Learned CPTs for the asynchrony variables.*

but the asynchronous models succeeded were in fact the kinds of pronunciations that we expect to arise from feature asynchrony, such as nasals replaced by nasalization on a preceding vowel.

Finally, we may wonder how a model using IPA-style features would fare on this task. We implemented an IPA-based model with synchrony constraints chosen so as to mirror those of the AP-based model to the extent possible. For example, **voicing** and **nasality** share an index variable, analogously to **GLOTTIS** and **VELUM** in the AP-based model, and the **front/back**–**height** soft synchrony constraint is analogous to the one on **TT**–**TB**. The remaining synchrony constraints are on the pairs **height**–**place**, **place**–**mann**, **mann**–**round**, and **round**–**voi/nas**. We imposed the following hard constraints on asynchrony, also chosen to match as much as possible those of the AP-based model:

$$\begin{aligned} p(async_t^{F;H} > 0) &= 0 \\ p(async_t^{H;P} > 1) &= 0 \\ p(async_t^{P;M} > 1) &= 0 \\ p(async_t^{M;R} > 1) &= 0 \\ p(async_t^{R;V,N} > 2) &= 0 \end{aligned}$$

Lines (11) and (12) of Table 4.3 show the performance of this model in terms of coverage and accuracy, before and after training. Both measures are intermediate to those of the baseline and AP-based models. This might be expected considering our argument that IPA features can capture some but not all pronunciation phenomena in which we are interested. In addition, the IPA model committed eight errors on words correctly recognized by the baseforms-only model. However, we note that this model was not tuned as carefully as the AP-based one, and requires further experimentation

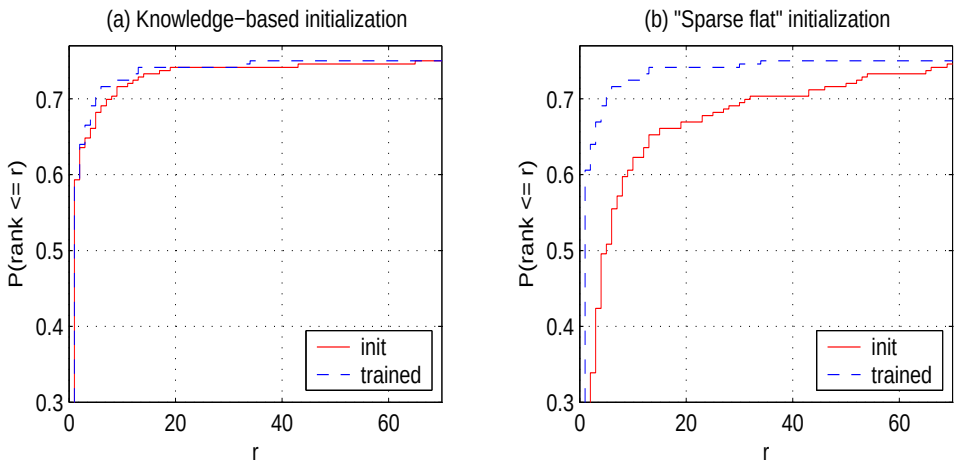

Figure 4-6: *Empirical cumulative distribution functions of the correct word's rank, before and after training.*

before firm conclusions can be drawn.

4.5 Discussion

We look to problematic examples to guide us toward future improvements. We have noted that the feature-based model failed to recognize the examples *favorite* → [f ey v er t] and *twenty* → [t w eh n iy]. In the first example, our current model does not allow the transformation /ax r ih/ → [er]. In order to properly model this, we may need to either consider retroflexion as a separate feature capable of desynchronizing from others, or else allow for context-dependent substitutions of tongue feature values.

In the second example, this is because although the model allows stops to be nasalized, it still expects them to have a burst; however, when a stop is nasalized, there is no longer enough pressure built up behind the constriction to cause a burst when the constriction is released. This may indicate a general problem with our modeling of stops. One possible modification would be to consider a stop to be a closure only, with the burst occurring as a function of surrounding context. We note that this is an example for which allowing deletions might seem to be a natural solution. However, since the underlying phenomenon is not one of deletion but rather an acoustic consequence of asynchrony, we prefer to improve our modeling of stops to account for this.

We can gain insight not only from misrecognized words, but also from words whose ranks are inappropriate. For example, in Figure 4-5 we analyzed the realization of *instruments* as [ih_n s tcl ch em ih_n n s]. While this example was correctly recognized, we may be surprised that the second-ranked word for this transcription was *investment*. Figure 4-8 shows part of the model's analysis of this "realization" of *investment*. The most disturbing aspect is the realization of the underlying critical

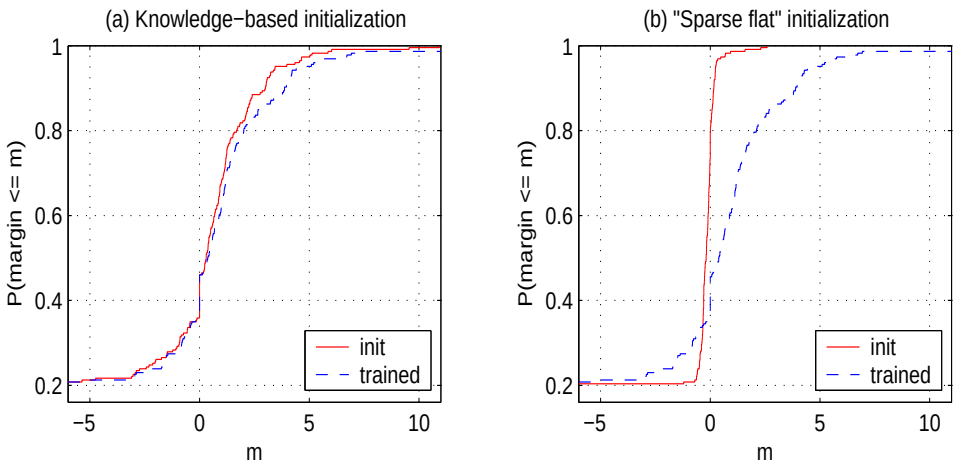

Figure 4-7: *Empirical cumulative distribution functions of the score margin, before and after training.*

lip closure for /v/ as a wide lip opening. We cannot disallow such a substitution in principle, as it is needed to explain such pronunciations as *probably* → [p r aa l iy]. However, this is again a case of context-dependence: While a labial consonant can greatly reduce in an unstressed syllable, it is much less likely to do so in a stressed syllable as in *investment*. As long as the model does not account for the context-dependency of substitutions, such anomalous pronunciations are assigned unnaturally high probability.

These types of examples indicate that adding context-dependency to our model of substitutions is an important next step. We hypothesize that, in the constrained setting of lexical access experiments using manual transcriptions, recognition performance is not significantly impaired by the "over-permissiveness" of the substitution modeling because most of the more aberrant pronunciations are not seen in the data. In end-to-end recognition experiments, however, we are not given the surface form but must instead rely on noisy observation models, and therefore expect that context-independent substitution modeling is inadequate. For this reason, when we turn to end-to-end recognition in Chapters 5 and 6, we allow asynchrony but, for the most part, disallow substitutions.

There are more general issues to be considered as well. There is a large space of synchrony structures and parameter initializations to be explored. Linguistic considerations have biased us toward the ones we have chosen. However, our linguistic knowledge is incomplete, so that it may be helpful to learn the synchrony constraints from data.

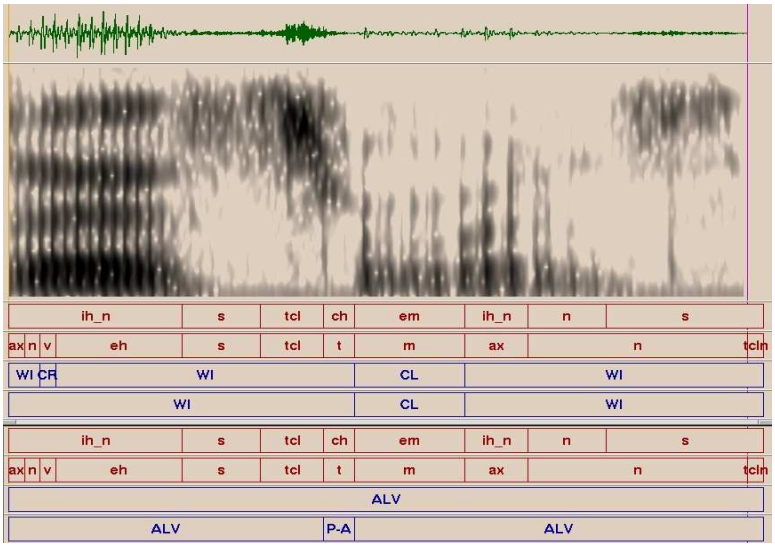

Figure 4-8: *Spectrogram, phonetic transcription, and partial alignment, including the variables corresponding to* **LIP-OPEN** *and* **TT-LOC**, *for* investment → *[ih_n s tcl ch em ih_n n s]. Indices are relative to the underlying pronunciation* /ih n s tcl t r ax m ax n tcl t s/.

4.6 Summary

In this chapter we have investigated the performance of a particular feature-based pronunciation model. We have developed a feature set and phone-to-feature mappings based on the vocal tract variables of articulatory phonology, and have argued that this is a more natural fit to a feature-based pronunciation model than the IPA-style features that prevail in acoustic observation modeling work. As noted in Chapter 3, the use of one feature set in the pronunciation model does not preclude the use of a different one in the observation model. In Chapter 5, we will describe such a system. In fact, since these experiments were based on *phonetic* transcriptions, they suggest that we may see a benefit from a feature-based pronunciation model even when using conventional phone-based observation models. We believe, however, that this would be a handicap, as much detail is lost in the phone-based representation.

The main results from this chapter's experiments are that

1. A phone-based model, even when augmented with a large set of phonological rules, fails to allow for most of the variation seen in a set of manual phonetic transcriptions of conversational speech.

2. The proposed feature-based model has greatly increased coverage, while also recognizing words with higher accuracy. This is not simply due to the addition of "pronunciation noise" in combination with a constrained task: When we change the non-zero probabilities in the model to uniform values, accuracy drops far below the baseline level.

3. When learning the parameters of the feature-based model using EM, the same coverage and accuracy are obtained for both a careful knowledge-based parameter initialization and one in which only the structure of zero/non-zero CPT elements is determined in a knowledge-based way, while the non-zero probabilities are uniform.

4. For this data set, a model with asynchrony alone does not outperform the baseline; one with substitutions alone performs much better; and the combination of both significantly outperforms the substitution-only case. We hypothesize that this is because there are very frequent minor deviations from baseform phones in the transcriptions, such as [ah] $\rightarrow$ [ax].

Chapter 5

Integration with the speech signal: Acoustic speech recognition experiments

As described in Chapter 3, there is a number of options for combining feature-based pronunciation models with observation models:

1. Analogously to HMM-based recognition, using Gaussian mixture models over acoustic observations, conditioned on the current vector of surface feature values.

2. Using feature-specific observations obs_t^i, and factoring

$$p(obs_t^1, \ldots, obs_t^N | s_t^1, \ldots, s_t^N) = \prod_{i=1}^{N} p(obs_t^i | s_t^i) \tag{5.1}$$

3. Using feature classifiers, with probabilistic outputs used as soft evidence in the DBN.

In the latter two cases, there is also the option of using a feature set for observation modeling that differs from the lexical features, as long as there is a mapping from the lexical features to the acoustic ones.

Comparison of these approaches is outside the scope of this thesis. However, we have built several systems to demonstrate the use of feature-based pronunciation models using different observation modeling strategies. In this chapter, we describe two systems applied to different tasks, one using Gaussian mixture-based observation models and one using distinctive features classified at various (non-uniformly spaced) points in the signal. For all experiments, the Graphical Models Toolkit [BZ02, GMT] was used for DBN training and testing.

set	# utterances	# one-word utterances	# non-silence words	length (hours)
train (A–C)	8352	4506	16360	4.25
dev (D)	3063	1707	5990	1.58
test (E)	3204	1730	6294	1.65

Table 5.1: *Sizes of sets used for SVitchboard experiments.*

5.1 Small-vocabulary conversational speech recognition with Gaussian mixture observation models[1]

We first describe an experiment using a simple Gaussian mixture-based end-to-end recognizer for a small-vocabulary conversational domain. The goal of this experiment is simply to determine how an end-to-end system using a feature-based pronunciation model would compare with a conventional HMM-based recognizer using the same type of observation model and acoustic observation vectors. Although we believe that feature-based systems will benefit from feature-specific acoustic measurements, it is reasonable to ask how they would fare without them.

5.1.1 Data

The data set for these experiments is a portion of the beta version of SVitchboard, a **S**mall **V**ocabulary subset of Sw**itchboard** [KBB05].[2] SVitchboard consists of several "tasks", corresponding to different subsets of Switchboard with different vocabulary sizes, ranging from 10 to 500 words (plus a silence word). Starting with a vocabulary of the five most frequent words in Switchboard, new words are added to the vocabulary one at a time, such that each new word maximizes the size of the resulting data set containing only the current vocabulary. We use the 100-word task from the SVitchboard beta version, as a compromise between computation and variety of words. Most of the words are common function words (*I, and, the, you*, and so on), but some are longer content words (*exactly, interesting, wonderful*). In addition, more than half of the utterances consist of only one word.

Each SVitchboard task is subdivided into five sets, A–E, corresponding to disjoint sets of speakers. We use sets A–C for training, D as a held-out development set, and E for final testing. Table 5.1 gives some descriptive statistics for these sets.

[1]This section describes work done in collaboration with Jeff Bilmes. We gratefully acknowledge the assistance of Chris Bartels and Simon King.

[2][KBB05] describes the released version of SVitchboard, rather than the beta version. The experiments in this section were done before the official release of SVitchboard, so the data sets are slightly different from the published description.

5.1.2 Model

To minimize computation, we used a simplified version of the model presented in Section 4.2.1, shown in Figure 5-1. We allowed asynchrony but no substitutions, and assumed that the following sets of features are always completely synchronous:

1. **LIP-LOC, LIP-OPEN**
2. **TT-LOC, TT-OPEN, TB-LOC, TB-OPEN**
3. **GLOTTIS, VELUM**

Effectively, then, we have three features consisting of these combinations of features, which we will refer to as **L, T**, and **G**, respectively. Since we allow no substitutions or asynchrony within these subsets, the number of possible combinations of feature values is also greatly reduced: six values for **L**, nineteen for **T** and three for **G**, resulting in a total of 342 possible states. Table B.3 of Appendix B lists the allowed values for each feature. We used the hard synchrony constraints, similarly to Section 4.4, that **L** and **T** must be within one state of each other and **G** must be within one state of either **L** or **T**. As in Chapter 4, we assumed that all features synchronize at the end of each word. Given these constraints and our vocabulary, 130 out of the possible 342 states are actually used in the model.

The observation model, $p(obs_t|u_t^L, u_t^T, u_t^G)$, is implemented as a mixture of Gaussians. We imposed a three-frame minimum duration for each segment in each feature stream.

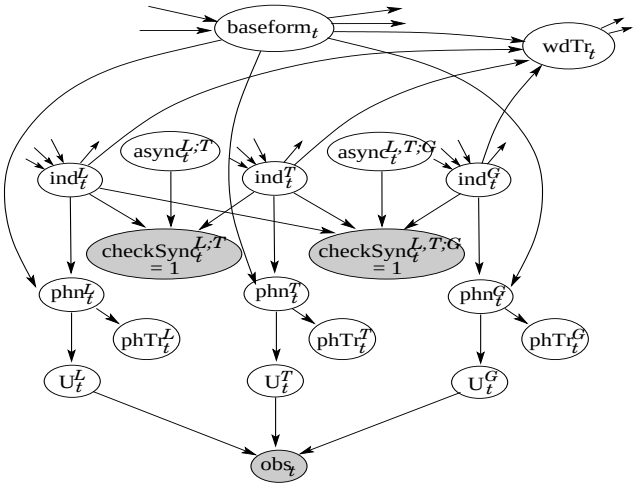

Figure 5-1: *DBN used for experiments on the SVitchboard database.*

We compare this model to a baseline three-state context-independent monophone HMM-based recognizer. The acoustic observations for both recognizers consisted of the first 13 Mel-frequency cepstral coefficients and their first and second differences, resulting in a 39-dimensional observation vector. This is a commonly used observation vector in ASR systems [RJ93].

system	# mix.	# Gauss.	S	I	D	WER	SER
phone, ph. 5, dev	126	9803	42.5	15.7	13.3	71.6	65.0
phone, ph. 6, dev	”	14311	41.8	15.9	12.0	69.7	63.9
phone, ph. 7, dev	”	20693	43.5	16.0	11.7	71.2	65.0
phone, ph. 6 + converge, test	”	**14311**	**41.2**	**16.6**	**11.3**	**69.1**	**63.9**
feat, ph. 6, dev	130	7092	40.5	9.0	22.3	71.8	67.4
feat, ph. 7, dev	”	9819	39.8	8.6	22.5	70.9	66.9
feat, ph. 8, dev	”	13527	39.5	8.6	23.2	71.3	66.9
feat, ph. 7 + converge, test	”	**9819**	**38.8**	**7.0**	**23.3**	**69.1**	**65.9**

Table 5.2: *SVitchboard experiment results. S = word substitution rate; I = word insertion rate; D = word deletion rate; WER = word error rate; SER = sentence error rate. All error rates are percentages.*

5.1.3 Experiments

For both the baseline and feature-based models, we used diagonal Gaussian mixtures as the observation model. Each Gaussian mixture was initialized with a single Gaussian with random near-zero mean and equal variances for all 39 dimensions. EM training was done in several phases. In the first few EM iterations of each phase, mixture components with high weights are each split into two components (with identical covariances and slightly offset means) and, optionally, ones with low weights are removed. EM iterations then continue until convergence, defined as a $< 2\%$ relative difference in the log likelihood between successive iterations. The recognizer is tested on the development set after each phase, and training is stopped after the first phase in which the word error rate does not decrease. The number of Gaussians for the baseline and feature-based systems may therefore differ, but each is chosen to optimize that system's performance for a fair comparison (up to the limits of this training procedure). Once the number of Gaussians has been chosen, additional EM iterations are done on the combined training + development sets (i.e., sets A–D), with no splitting or vanishing of Gaussians, until convergence with a $< 0.2\%$ relative log likelihood difference.

Table 5.2 shows the performance of the baseline and feature-based systems, in the latter phases of training on the development set and on the final test set, in terms of both word and sentence error rates. The difference in sentence error rates on the test set is significant according to McNemar's test ($p = 0.008$). The language model is an unsmoothed bigram trained on the utterances in sets A–C.[3] As is common in ASR systems, an insertion penalty was imposed for each hypothesized word to account for the recognizers' preference for deletions; the size of the penalty was manually tuned on the development set.

[3]In preliminary experiments not reported here, a smoothed bigram was found to produce higher word error rates.

5.1.4 Discussion

The results of this experiment show that a very simple feature-based model has comparable performance to a context-independent monophone HMM-based system. A more extensive study is needed to determine whether additional asynchrony or substitutions would improve this performance further. The constraint that features synchronize at word boundaries is also an artificial one that should be relaxed in future work. To be competitive with state-of-the-art systems with highly context-dependent observation models, a version of context-dependent modeling will be needed for feature-based models as well. Analogously with triphone-type systems, the context may be the neighboring feature values or any combination thereof. An interesting direction for further work would be to study the relative benefits of using different combinations of feature contexts, as well as to automatically learn the most informative contexts. Finally, the use of context-dependent observation distributions would greatly increase the number of distributions. This could be addressed using the commonly applied strategy of state clustering; the most useful contexts for clustering would also be an interesting direction for future work.

5.2 Landmark-based speech recognition[4]

Stevens [Ste02] has proposed a method for speech recognition based on the acoustic analysis of important points in the signal, referred to as *landmarks*. Stevens' landmarks correspond to points of closure and release for consonants, vowel steady states, and energy dips at syllable boundaries. The first step of landmark-based recognition is to locate the landmarks using various acoustic cues, which may differ for different types of landmarks. At each detected landmark, a bank of classifiers hypothesize values for a number of binary *distinctive features*. Distinctive features are acoustically and articulatorily motivated features intended to be sufficient for word classification. Different features are classified at different landmarks, depending on the type of landmark. The detected feature values are then matched against the lexicon, which is also represented in terms of distinctive features.

Juneja and Espy-Wilson [JEW04] have implemented phonetic and small-vocabulary recognizers based on these principles. They use support vector machine (SVM) classifiers to detect both landmarks and features. Figure 5-2 shows an example of detected landmarks. At each type of landmark, different features are classified; for example, there are separate vowel-related and consonant-related feature classifiers for the corresponding landmarks. In addition, when a certain feature applies to multiple types of landmarks, different SVMs are trained for the same feature at different landmarks. In general, there is a hierarchical structure determining which classifiers are used in each context, and each SVM is trained only on data from the corresponding context. This

[4]This section describes work done at the 2004 Summer Workshop of the Johns Hopkins University Center for Language and Speech Processing, as part of the project "Landmark-Based Speech Recognition" led by Mark Hasegawa-Johnson. Jeff Bilmes provided assistance with GMTK for this project. Parts of this section have appeared in the workshop project final report [ea04] and in [ea05]. All material appearing here, however, has been written by the author unless otherwise indicated.

results in a situation where the feature classifier outputs are not all "interpretable" in every frame; we return to this point later.

The SVM outputs are converted to posterior probabilities using a learned histogram method [JEW04], and the final phone or word hypotheses are determined using a dynamic programming algorithm over the graph of possible landmark/feature values. In this work, a deterministic mapping is assumed between a phonetic lexicon and detected feature values.

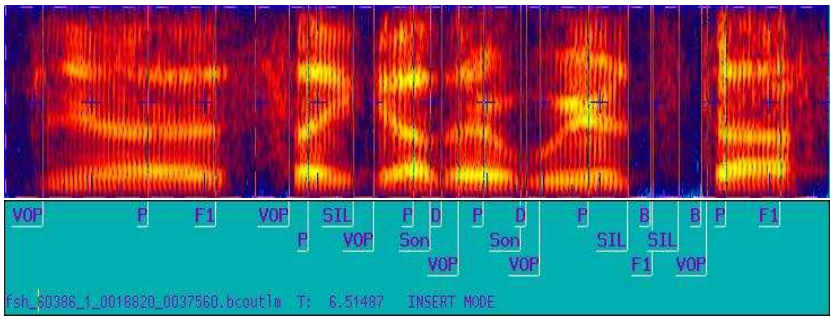

Figure 5-2: *Example of detected landmarks, reproduced from [ea04]. VOP = voice onset point; P = syllabic peak; F1 = fricative onset; SIL = silence onset; Son = sonorant onset; D = syllabic dip; B = stop burst.*

.

One issue with this approach is that the surface feature values and landmark locations may not neatly correspond to phonetic segments. Different features may evolve at different rates and may not reach their target values, resulting in segments of speech that do not correspond to any phone in the English phonetic inventory and in which boundaries between segments are not clearly defined. In order to relax this assumption, we have developed a recognizer based on combining SVM classifiers of landmarks and distinctive features with an articulatory feature-based pronunciation model allowing for asynchrony and (limited) substitution of feature values. In particular, we use the AP-based pronunciation model described in Chapter 4.

This work was carried out in the context of a larger project using an isolated-word landmark-based recognizer to rescore word lattices produced by an HMM-based baseline recognizer [ea05]. The landmark and distinctive feature classification is described in detail in [ea04]. Here we describe only the integration of the pronunciation model with the feature classifiers and give an example analysis of a difficult phrase. The goal is to show that it is possible to combine a feature-based pronunciation model of the type we have proposed with an observation modeling approach that may at first seem incompatible.

5.2.1 From words to landmarks and distinctive features

Combining an AP-based pronunciation model with distinctive feature classifier outputs involves two tasks: (1) conversion between articulatory features (AFs) and dis-

tinctive features (DFs), and (2) incorporation of likelihoods computed from SVM outputs. Our solutions for both of these are depicted in Figure 5-3.

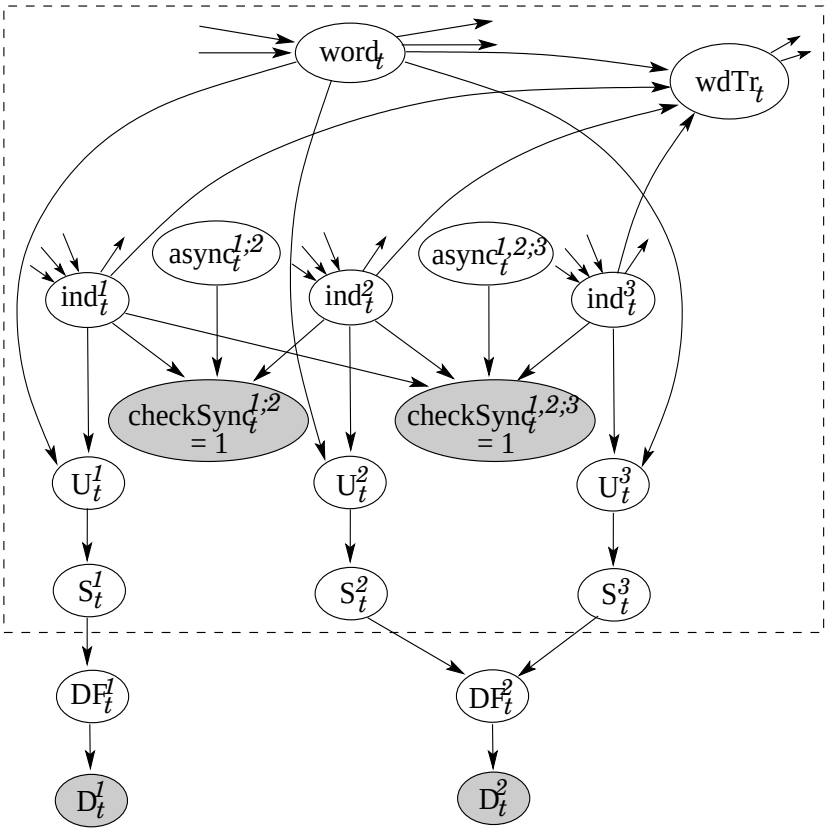

Figure 5-3: *Example of a DBN combining a feature-based pronunciation model with landmark-based classifiers of a* different *feature set. The actual model used in experiments is not shown, as it uses more than 70 additional variables to represent the context-specific distinctive features. However, for completeness, the mapping from articulatory to distinctive features is given in appendix Tables B.4–B.5.*

For the first task, we simply used a deterministic mapping from articulatory to distinctive features, implemented by adding to the DBN a variable corresponding to each DF and its associated dependencies; e.g., *sonorant* = 1 whenever the glottis is in the voiced state and either the lip and tongue openings are narrow or wider (a vowel, glide, or liquid) or there is a complete lip/tongue closure along with an open velum (a nasal consonant). The AF-to-DF mapping can be complicated, but it need only be specified once for a given set of AFs and a given set of DFs. In this way, pronunciations and acoustics can be modeled using completely different feature sets, as long as there is a deterministic mapping between the pronunciation model's feature set and the one used to model the acoustics. In the case of our feature sets, the mapping is almost deterministic; the main exceptions include the silence DF (for which there is no analogue in terms of articulatory features) and, possibly, the lateral DF (since the horizontal dimension of the tongue is not represented in the AF set).

In order to incorporate the outputs of the SVMs, we used the Bayesian network construct of *virtual* (or *soft*) *evidence* [Bil04, Pea88]. This is used when a variable is not observed, i.e. there is no *hard evidence* about it, but we have some information about it that causes us to favor some values over others; this is exactly what the SVM outputs tell us about the values of the DFs. This is done by adding, for each DF, a "dummy" variable D_{DF}, whose value is always 1 and whose distribution is constructed so that $P(D_{DF} = 1|DF = f)$ is proportional to the likelihood for $DF = f$. This "hybrid DBN/SVM" is the final DBN used for recognition.

The hierarchical organization of the SVMs gives rise to an interesting problem, however. Since each SVM is trained only in a certain context (e.g. separate *Labial* classifiers are trained for stop-vowel, vowel-stop, fricative-vowel, and vowel-fricative boundaries), only those SVM outputs relevant to the hypothesis being considered in a given frame are used. For example, the output of the "dental fricative" classifier is only meaningful in frames that correspond to fricatives. Which SVMs will be used in a given frame can be determined by the values of certain variables in the DBN. For example, if the current frame corresponds to a closure and the previous frame corresponds to a vowel (both of which can be determined by examining the values of **LIP-OPEN, TT-OPEN**, and **TB-OPEN**), the vowel-stop SVMs will be used. This is implemented using the mechanism of switching dependencies (see, e.g., [BZR+02, Bil00, GH96]); e.g., **LIP-OPEN, TT-OPEN**, and **TB-OPEN** are switching parents of the *Labial* SVM soft evidence "dummy" variables. Appendix B, Tables B.4–B.5, show the complete articulatory-to-distinctive feature mapping, along with the context (i.e. the values of the switching parents) in which each SVM is licensed. Such mapping tables were used to automatically generate a DBN structure for a given set of articulatory and distinctive features.

The problem with this mechanism is that different hypotheses that are being compared during decoding may have different numbers of relevant SVMs, and therefore different numbers of probabilities being multiplied to form the overall probability of each hypothesis. For example, the first and last frames of a fricative with an adjacent vowel will license both the isolated *Strident* classifier and the *Strident* classifier specific to a vowel-fricative or fricative-vowel landmark. For this reason, hypotheses that license fewer SVMs will be preferred; e.g. we can imagine a situation where a hypothesis containing one long fricative will be preferred over one containing two fricatives with a short intervening vowel.

Our solution to this, for the time being, has been to score words in two passes: The manner SVMs *silence, sonorant, continuant, and stop*, which are interpretable in all frames, are used to obtain a manner segmentation, using either the Juneja/Espy-Wilson alignment system referred to above, or else the DBN itself with only the manner DF variables; the full DBN is then used along with the remaining SVM outputs to compute a score conditioned on the manner segmentation, using each SVM only in the context in which it is licensed. This issue, however, merits further study.

The DBN parameters (i.e., the entries in the various conditional probability tables) were estimated via EM, using as training data either a subset of the Switchboard ICSI transcriptions, as in Chapter 4, or the SVM outputs themselves. We used three

training conditions: (a) a 1233-word subset of the phonetic transcriptions, consisting of all words in the training set of Chapter 4 except for the ones to which the model assigns zero probability; (b) the SVM outputs computed on this same 1233-word set; and (c) the SVM outputs for the entire training set of ICSI transcriptions, consisting of 2942 words. While these sets are small, the DBN has only several hundred trainable parameters.

For all experiments, all of the AFs besides **LIP-LOC** were used. **LIP-LOC** was excluded in order to limit the required computation, and because there is only one pair of phones ([aa] and [ao]) that are distinguished only by their **LIP-LOC** values. We imposed the same synchronization constraints as in Chapter 4, reproduced here:

1. All four tongue features are completely synchronized,

$$async_t^{TT;TB} = 0 \tag{5.2}$$

2. The lips can desynchronize from the tongue by up to one index value,

$$p(async_t^{LO;TT,TB} > 1) = 0 \tag{5.3}$$

3. The glottis/velum index must be within 2 of the mean index of the tongue and lips,

$$p(async_t^{LO,TT,TB;GV} > 2) = 0 \tag{5.4}$$

These constraints result in 3 free synchronization parameters to be learned:

$$P(async_t^{LO;TT,TB} = 1) \tag{5.5}$$
$$P(async_t^{LO,TT,TB;GV} = 1) \tag{5.6}$$
$$P(async_t^{LO,TT,TB;GV} = 2) \tag{5.7}$$

the remaining asynchrony probabilities either are set to zero or can be computed from these three probabilities. This may seem like a very small amount of variation; however, this limited degree of asynchrony accounts for the majority of phenomena we are aware of. The types of asynchrony phenomena that are not allowed under these constraints are extreme spreading, as can sometimes happen with nasality (e.g., *problem* → [p r aa_n m]) or rounding (e.g., of the [s] in *strawberry*). When the DBN was trained on the phonetic transcriptions, the learned asynchrony probabilities were found to be: $P(async_t^{LO;TT,TB} = 1) = 1.05 \times 10^{-3}$; $P(async_t^{LO,TT,TB;GV} = 1) = 7.20 \times 10^{-4}$; and $P(async_t^{LO,TT,TB;GV} = 2) = 3.00 \times 10^{-27}$.

For most experiments, the only feature whose surface value was allowed to differ from the underlying value was **LIP-OPEN**. This constraint was again intended to reduce computational requirements. **LIP-OPEN** was chosen because of the high frequency of (anecdotally observed) reductions such as *probably* → [p r aw l iy]. We allowed **LIP-OPEN** to reduce from **CL** to **CR** or **NA**, and from **CR** to **NA** or **WI**; all other values were assumed to remain canonical. The learned reduction probabilities, when training from either phonetic transcriptions or SVM outputs, are shown in Table 5.3.

	S=**CL**	S=**CR**	S=**NA**	S=**WI**
U=**CL**	9.996×10^{-1}	2.555×10^{-11}	4.098×10^{-4}	0
U=**CR**	0	7.933×10^{-1}	1.619×10^{-35}	2.067×10^{-1}
U=**NA**	0	0	1	0
U=**WI**	0	0	0	1

	S=**CL**	S=**CR**	S=**NA**	S=**WI**
U=**CL**	8.350×10^{-1}	1.102×10^{-2}	1.540×10^{-1}	0
U=**CR**	0	3.014×10^{-1}	3.030×10^{-1}	3.955×10^{-1}
U=**NA**	0	0	1	0
U=**WI**	0	0	0	1

Table 5.3: *Learned reduction probabilities for the* **LIP-OPEN** *feature,* $P(S^{LO} = s|U^{LO} = u)$*, trained from either the ICSI transcriptions (top) or actual SVM feature classifier outputs (bottom).*

A final time-saving measure for these experiments was the use of relatively low frame rates: All experiments used either 20ms or 15ms frames. Since the SVMs were applied every 5ms, their outputs were downsampled to match the frame rate of the DBN.

As a way of qualitatively examining the model's behavior, we can compute a Viterbi "forced alignment" for a given waveform, i.e. the most probable values of all of the DBN variables given the word identities and the SVM outputs. Figure 5-4 shows an alignment for the phrase "I don't know". [5] In this example, both the /d/ and the /n t n/ sequence have been produced essentially as glides. In addition, the final /ow/ has been nasalized, which is accounted for by asynchrony between **VEL** and the remaining AFs. The fact that we can obtain reasonable alignments for such reduced pronunciations is an encouraging sign. Furthermore, this example suggests one reason for choosing an acoustically-motivated representation for observation modeling and an articulatory one for pronunciation modeling: There is arguably little hope of determining the locations and opening degrees of the various articulators in this extremely reduced phrase; and there is also arguably little hope of predicting the observed distinctive feature values based on transformations from an underlying distinctive feature-based lexical representation.

5.2.2 Discussion

We have only scratched the surface of the issues that need to be explored in a system using an articulatory pronunciation model with a landmark-based observation model. We are continuing to examine the proper way to account for the distinctive feature hierarchy. Additional issues to be investigated are:

[5] This figure was generated using an xwaves-based display tool developed at the JHU '04 workshop by Emily Coogan.

The weighting of the soft evidence relative to other probabilities in the DBN. This is analogous to the weighting of Gaussian mixture likelihoods and transition probabilities in a conventional HMM.

Iterative training of the DBN and SVMs. As currently implemented, there is a mismatch between the DBN and SVMs, which are trained on phonetic transcriptions that do not contain the "non-phonetic" feature value combinations that are allowed in the DBN. Given the initial set of SVMs trained on phonetic transcriptions, the DBN could be used to re-transcribe the training data in terms of feature values, and to use this re-transcribed data to retrain the SVMs. This process can be iterated, akin to Viterbi training in conventional systems.

Finally, there are some more general questions that this research brings up. For example, how does one choose the feature set and ranges of feature values? This project has taken the position that articulatory features (rather than, e.g., distinctive features) are natural units for modeling pronunciation variation, whereas distinctive features are more natural for modeling the acoustic signal. It may be argued that the particular choice of feature values is an arbitrary one, making for an inelegant model. However, phonetic units are arguably more arbitrary, as they have little justification from a linguistic point of view and are a poor fit to highly reduced speech of the type we have discussed. There is room, nevertheless, for research into the most appropriate linguistic feature space for speech recognition, as this issue has not been widely studied to date.

5.3 Summary

In this chapter we have presented two possible scenarios for the integration of acoustic observations with a feature-based pronunciation model. We have presented a simple mixture Gaussian-based system using features rather than phones, which performs similarly to a baseline monophone system. We have also shown that we can combine a feature-based pronunciation model with acoustically-motivated landmark-based classifiers, and we have seen an example that shows the benefits of using the two different representations in different parts of the recognizer.

Much work is required to make such systems competitive with state-of-the-art recognizers. In Gaussian mixture-based systems, context dependency, state clustering, and the word synchronization constraint will need to be dealt with. For landmark-based recognition, there are more fundamental issues such as the problem of different numbers of feature classifiers for different hypotheses. However, a similar system in which all SVMs are trained on all frames would not have the same problem.

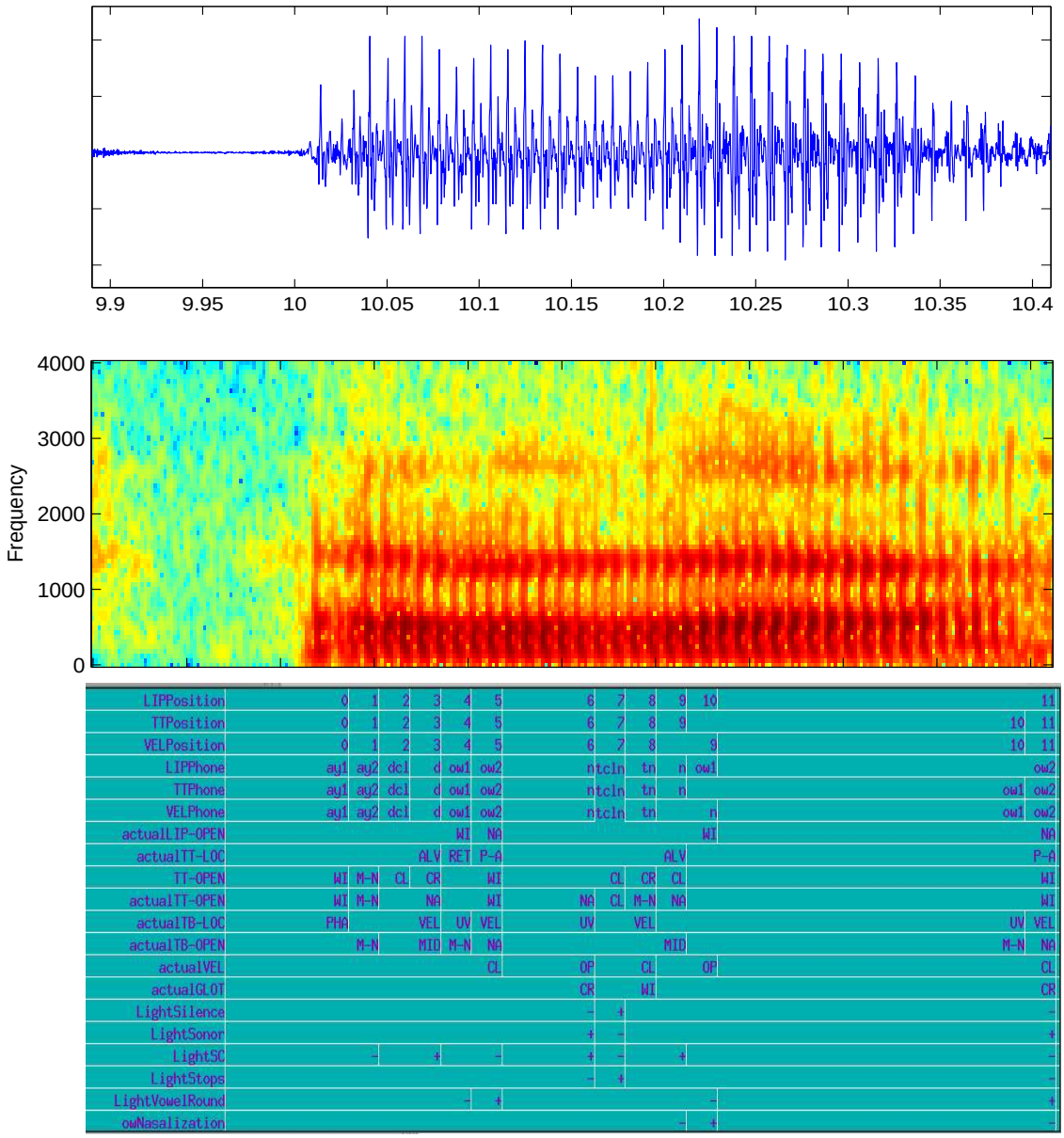

Figure 5-4: *Waveform, spectrogram, and some of the variables in an alignment of the phrase "I don't know". The notation in this figure differs slightly from the previously used variable names: The <feature>Position variables correspond to the ind variables in Figure 5-3;* <feature>*Phone correspond to the ph variables;* <feature> *is the underlying feature value U; actual*<feature> *is the surface value S; and Light*<DF> *is the value of the distinctive feature DF ("Light" in front of a DF name simply refers to the fact that the SVMLight package was used to train the SVM, as opposed to the other packages used in this project). While the underlying value for the tongue tip opening (***TT-OPEN***) is "closed" (***CL***) during the underlying /dcl/ and /n t n/, the surface value (***actualTT-OPEN***) is "narrow" (***NA***). The effect of asynchrony can be seen, e.g., during the initial portion of the /ow/: This segment is nasalized, which is hypothesized to be the result of asynchrony between the velum and remaining features.*

Chapter 6

Lipreading with feature-based models

Until now we have been treating speech as an acoustic signal. However, speech is communicated through both the acoustic and visual modalities, and similar models can be applied to both acoustic and visual speech recognition. In fact, in the visual modality, there is a closer connection between the signal and the articulatory features we have discussed, because some of the features are directly measurable. In this chapter, we apply the ideas of feature-based modeling to visual speech recognition, or lipreading. Lipreading can be used in combination with acoustic speech recognition, for example to improve performance in noise [NLP+02]. We can also imagine scenarios where lipreading may be useful by itself, for example if there is sufficient acoustic noise, if the acoustic signal has been corrupted or lost, or if the acoustic environment is highly variable and there is insufficient matched acoustic training data.

The sub-word unit of choice in most lipreading systems (or the lipreading components of audio-visual recognition systems) is the *viseme*, the visual analogue of the phoneme, defined as a set of visually indistinguishable phonemes (or, sometimes, phones). For example, in a lipreading lexicon, the words *pan, ban, man, pad, bad, mad, pat, bat*, and *mat* may all have the same pronunciation, e.g., [labial-closure low-front-vowel alveolar-consonant]. Different viseme sets have been defined for different systems, and they are often defined automatically through clustering [Haz]. In this chapter, we consider whether the single stream of states in visual recognition may also benefit from factoring into multiple semi-independent feature streams. In the following sections we discuss two types of lipreading tasks: medium-vocabulary word *ranking*, which we can imagine using to enhance an acoustic speech recognizer in difficult acoustic conditions; and small-vocabulary *isolated phrase recognition*, which could be used as a stand-alone system, for example in a noisy car or kiosk environment.

6.1 Articulatory features for lipreading

Our first task is to define a feature set. We make the simplifying assumption that the only visible features are those associated with the lips. In general, the tongue and teeth may also be visible at times, but their frequent occlusion by the lips presents a complication that we prefer not to introduce for the time being. We take as a starting point the articulatory phonology-based feature set defined in Chapter 4. The only two relevant features are **LIP-LOC** and **LIP-OPEN**, whose values are listed again in Table 6.1 for convenience.

feature	**values**
LIP-LOC (LL)	protruded, labial, dental
LIP-OPEN (LO)	closed, critical, narrow, wide

Table 6.1: *The lip-related subset of the AP-based feature set.*

Until now we have always constrained these features to be synchronous. Figure 6-1 shows a situation in which they can be asynchronous. In this example, a labial closure is followed by a rounded (i.e., protruded) vowel, and the early onset of rounding affects the shape of the mouth during the closure.

Figure 6-1: *Example of lip opening/rounding asynchrony. Compare the shape of the mouth during the latter part of the [m] lip closure (i.e. the second image) in the words* milk *(top) and* morning *(bottom) from the sentence "Greg buys fresh milk each weekday morning."*

One drawback of these features is that they do not allow for independent control of labio-dental articulation and rounding. It is possible to produce a labio-dental ([f] or [v]) while protruding the lips, and in fact this happens in rounded contexts, as shown in Figure 6-2. This can be explained as asynchrony between the labio-dental and rounded articulations, but is not allowed as long as *labio-dental* and *protruded* are values of the same feature. For this reason, we have modified the feature set to separate these into two binary features, **LAB-DENT (LD)** and **LIP-ROUND (LR)**.

Finally, we slightly modify the definition of **LIP-OPEN**, collapsing *critical* and *narrow* into a single value and adding a *medium* opening value. The former is motivated by the practical concern that the available image resolution is often too low to distinguish between critical and narrow lip constrictions; the latter allows us to distinguish among a larger number of configurations (which, in the acoustic experiments, were distinguished using other features). Our final feature set for lipreading experiments is shown in Table 6.2.

Figure 6-2: *Example of rounding/labio-dental asynchrony. Compare the shape of the mouth during the [f] in the words* breakfast *(top) and* four *(bottom) from the sentence "He had four extra eggs for breakfast." (For example, frame 10 in the top and frame 4 in the bottom.)*

feature	**values**
LAB-DENT (LD)	yes, no
LIP-ROUND (LR)	yes, no
LIP-OPEN (LO)	closed, narrow, medium, wide

Table 6.2: *Feature set used in lipreading experiments.*

6.2 Experiment 1: Medium-vocabulary isolated word ranking[1]

In this section we experiment with a medium-vocabulary, isolated-word lipreading system of the type that might be used in combination with an acoustic speech recognizer. For anything but a small, carefully-constructed vocabulary, recognizing words from lip images alone is impossible. However, the word scores provided by a lipreading system may still be useful in combination with an acoustic recognizer, even if the lipreading system alone is very poor. We therefore do not measure performance by error rate but by the rank given to the correct word. The task is, therefore, given a sequence of mouth images corresponding to a spoken word, to rank all of the words in the lexicon according to their posterior probabilities given the image sequence.

6.2.1 Model

We use a model architecture combining a feature-based DBN with the outputs of SVM feature classifiers converted to soft evidence, as described in Section 3.4. For each feature, an SVM is trained to classify among the feature's values. The output of each SVM is converted to a posterior using a sigmoidal mapping. so that, for each lip image x and each feature F, we have $P(F_t = f|X_t = x)$. The posterior is converted to a scaled likelihood by dividing by the feature priors: $P(X_t = x|F_t = f) \propto P(F_t = f|X_t = x)/P(F_t = f)$. This scaled likelihood is then used as soft evidence in the DBN shown in Figure 6.2.1. $P(obs_t^F = 1|S_t^F = f)$ is proportional to the likelihood

[1]This section describes joint work with Kate Saenko and Trevor Darrell [SLGD05]. The visual processing and SVM training are described in [SLGD05]; here we focus on the modeling of the hidden feature dynamics.

$P(X_t = x|S_t^F = f)$ computed from the SVM outputs. In these experiments, we concentrate only on the effects of asynchrony between the features; in other words, we allow asynchrony but no feature changes (i.e. $async_t^j$ can vary, but $S_t^F = U_t^F$). For this reason, the S_t^F variables have been dropped in Figure 6.2.1.

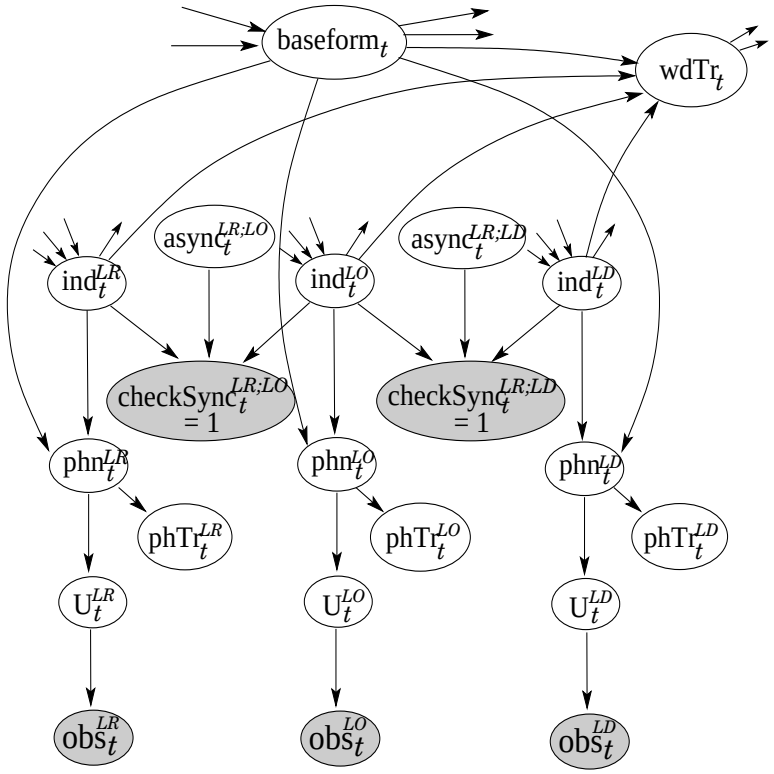

Figure 6-3: *One frame of a DBN used for lipreading.*

6.2.2 Data

For these experiments, we used 21 utterances taken from a single speaker in AV-TIMIT [Haz], a corpus of audio-visual recordings of subjects reading phonetically balanced sentences. Of these, 10 utterances were used for training and 11 for testing. To simulate the isolated-word task, utterances were split into words, resulting in a 70-word test set. Each visual frame was also manually transcribed with 3 AF values. The vocabulary contains 1793 words, and up to three baseforms are allowed per words.

6.2.3 Experiments

We have conducted experiments to investigate several questions that arise in using the proposed feature-based system. First, we would like to compare the effects of using feature-based versus viseme-based *classifiers*, as well as of using a feature-based

viseme	LO	LR	LD
1	closed	any	no
2	any	any	yes
3	narrow	rounded	no
4	medium	unrounded	any
5	medium	rounded	any
6	wide	any	any

Table 6.3: *The mapping from visemes to articulatory features.*

versus viseme-based *pronunciation model*. A viseme-based pronunciation model is a special case of our DBN in which the features are constrained to be completely synchronous (i.e. $async_t^j$ is identically 0). Using viseme classifiers with a viseme-based pronunciation model is similar to the conventional viseme-based HMM that is used in most visual speech recognition systems, with the exception that the likelihoods are converted from classifier outputs. In order to use a feature-based pronunciation model with viseme classifiers, we use a (many-to-one) mapping from surface features (S_t^F) to visemes. Also, since we do not have ground truth articulatory feature labels, we investigate how sensitive the system is to the quality of the training labels and classifiers using manual transcriptions of the data.[2] In order to facilitate quick experimentation, these experiments focus on an isolated-word recognition task and use only a small data set, with manual settings for the (small number of) DBN parameters.

The mapping between the six visemes and the feature values they correspond to is shown in Table 6.3. Although there are more than six possible combinations of feature values, only these six are represented in the manually labeled training data.

The details of SVM training and performance can be found in [SLGD05]. Here we concentrate on the word ranking experiments. For each spoken word in the test set, we compute the Viterbi path for each word in the vocabulary and rank the words based on the relative probabilities of their Viterbi paths. Our goal is to obtain as high a rank as possible for the correct word. Performance is evaluated both by the mean rank of the correct word over the test set and by examining the entire distribution of the correct word ranks.

In the models with asynchrony, **LIP-ROUND** and **LIP-OPEN** were allowed to desynchronize by up to one index, as were **LIP-OPEN** and **LAB-DENT**. Table 6.4 summarizes the mean rank of the correct word in a number of experimental conditions, and Figure 6.2.3 shows the empirical cumulative distribution functions (CDFs) of the correct word ranks in several of these conditions. In the CDF plots, the closer the distribution is to the top left corner, the better the performance. We consider the baseline system to be the viseme-based HMM, i.e. the synchronous pronunciation model using the viseme SVM.

Table 6.4 also gives the significance (p-value) of the mean rank differences between each model and the baseline, as given by a one-tailed paired t-test [Die98]. The

[2]Thanks to Kate Saenko for these transcriptions.

Classifier unit	Mean rank, sync model	Mean rank, async model
Viseme	281.6	262.7 (.1)
Feature, forced train	216.9 (.03)	209.6 (.02)
Feature, manual train	165.4 (.0005)	149.4 (.0001)
Feature, oracle	113.0 (2×10^{-5})	109.7 (3×10^{-5})

Table 6.4: *Mean rank of the correct word in several conditions. The significance of the difference between each system and the baseline, according to a one-tailed paired t-test [Die98], is given in parentheses.*

models using multiple feature-dependent classifiers always significantly outperform the ones using viseme classifiers. For each type of classifier and training condition, the asynchronous pronunciation model outperforms the synchronous one, although these differences are not significant on this test set. It may seem counterintuitive that asynchrony should make a difference when viseme classifiers are used; however, it is possible for certain apparently visemic changes to be caused by feature asynchrony; e.g., a [k] followed by an [uw] may look like an [ao] because of **LIP-OPEN**/**LIP-ROUND** asynchrony.

Next, the forced train vs. manual train comparison suggests that we could expect a sizable improvement in performance if we had more accurate training labels. While it may not be feasible to manually transcribe a large training set, we may be able to improve the accuracy of the training labels using an iterative training procedure, in which we alternate training the model and using it to re-transcribe the training set.

To show how well the system could be expected to perform if we had ideal classifiers, we replaced the SVM soft evidence with likelihoods derived from our manual transcriptions. In this "oracle" test, we simulated soft evidence by assigning a very high likelihood ($\approx$0.95) to feature values matching the transcriptions and very low likelihood to the incorrect feature values.

6.3 Experiment 2: Small-vocabulary phrase recognition[3]

We now describe experiments with a small-vocabulary, isolated-phrase visual speech recognition system. In these experiments the phrase vocabulary has been designed to be, in principle at least, visually distinguishable given our feature set. With these constraints, it is feasible to have a stand-alone lipreading system whose performance can be judged by error rate rather than by word ranks. The system described here is part of a larger project on end-to-end visual detection and recognition of spoken phrases [SLS+05]; here we will concentrate on the task of recognizing a phrase given

[3]This section describes joint work with Kate Saenko, Michael Siracusa, Kevin Wilson, and Trevor Darrell, described in [SLS+05].

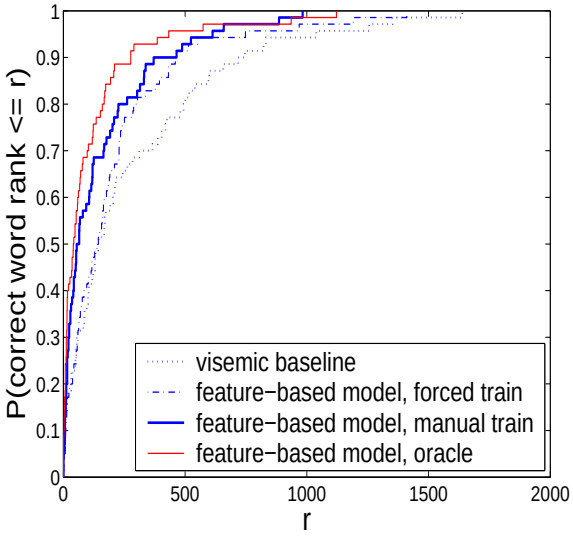

Figure 6-4: *CDF of the correct word's rank, using the visemic baseline and the proposed feature-based model. The rank* r *ranges from 1 (highest) to the vocabulary size (1793).*

a sequence of mouth images.

Here we experiment with a model that differs from the others we have considered in two main ways:

1. Rather than using either per-feature classifiers (as in Sections 5.2 and 6.2) or observation distributions conditioned on all feature values (as in Section 5.1), here we use per-feature observation vectors. These observations happen to be the outputs of SVM feature classifiers, but they are used as observations and modeled with Gaussian distributions.

2. It is a *whole-phrase* model, meaning that there are no explicit underlying phones or feature values, only several streams of phrase-specific states. Whole-word models typically outperform sub-word-based models on small-vocabulary tasks (e.g., [ZBR+01]). However, we will compare the whole-phrase model to sub-word models, including the one of Section 6.2.

6.3.1 Model

Figure 6-5 shows the DBN used in our experiments. The model essentially consists of three parallel HMMs per phrase, one per articulatory feature, where the joint evolution of the HMM states is constrained by synchrony requirements. For comparison, Figure 6-6 shows a conventional single-stream viseme HMM, which we use as a baseline in our experiments.

For each feature stream, the observations obs^F consist of the continuous SVM outputs for that feature (or for visemes, in the case of the baseline model). There

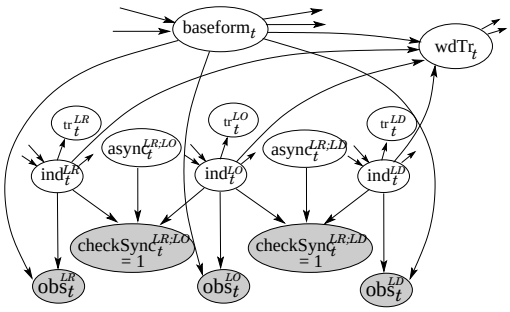

Figure 6-5: *DBN for feature-based lipreading. ind_t^F is an index into the state sequence of feature F, where F is one of $\{LO,LR,LD\}$.*

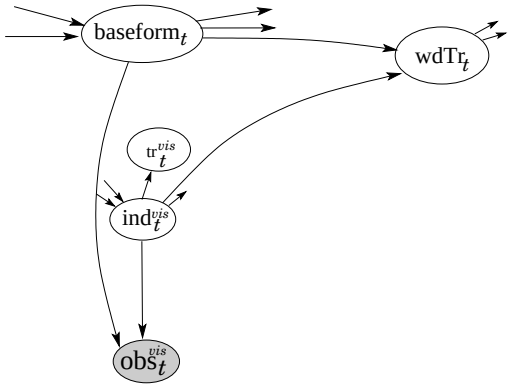

Figure 6-6: *DBN corresponding to a single-stream viseme HMM-based model.*

is a separate DBN for each phrase in the vocabulary, with ind^F ranging from 1 to the maximum number of states. Recognition is done by finding, for each DBN, the Viterbi path, or the most likely settings of all variables given the observations, and choosing the phrase whose DBN has the highest Viterbi score.

6.3.2 Data

We evaluated this recognizer on the task of isolated short phrase recognition. In particular, we used a set of 20 commands that could be used to control an in-car stereo system (see Table 6.5). The data consists of video of two speakers saying these twenty stereo control commands three times each. To test the hypothesis that coarticulation increases in fast, sloppy speech, and that a DBN that allows articulator asynchrony will better account for co-articulation in faster test conditions, the three repetitions were done at different speaking rates. During the first repetition, the speaker clearly enunciated the phrases (slow condition), then spoke successively faster during the second and third repetitions (medium and fast conditions.)

SVM classifiers for both articulatory features and visemes were trained on a separate data set, consisting of three speakers reading sentences from the TIMIT

1	"begin scanning"	11	"shuffle play"
2	"browse country"	12	"station one"
3	"browse jazz"	13	"station two"
4	"browse pop"	14	"station four"
5	"CD player off"	15	"station five"
6	"CD player on"	16	"stop scanning"
7	"mute the volume"	17	"turn down the volume"
8	"next station"	18	"turn off the radio"
9	"normal play"	19	"turn on the radio"
10	"pause play"	20	"turn up the volume"

Table 6.5: *Stereo control commands.*

train/test	dictionary-based			whole-phrase		
	viseme+GM	feature+SE	feature+GM	viseme+GM	feature+GM	async feature+GM
slow-med/fast	10	7	13	16	23 (p=0.118)	25 (p=0.049)
slow-fast/med	13	13	21	19	29 (p=0.030)	30 (p=0.019)
med-fast/slow	14	21	18	27	25	24
average	12.3	13.7	17.3	20.7	25.7	26.3
average %	**30.8**	**34.2**	**43.3**	**51.6**	**64.1**	**65.8**

Table 6.6: *Number of phrases, out of 40, recognized correctly by various models. The first column lists the speed conditions used to train and test the model. The next three columns show results for dictionary-based models. The last three columns show results for whole-phrase models.*

database [GLF+93]. Each frame in this set was manually annotated with values of the three features.

6.3.3 Experiments

The SVMs in these experiments use the one-vs.-all multi-class formulation, so that there are six classifiers for the three features: four for **LO**, one for **LR**, and one for **LD**. To evaluate the viseme-based baseline, we use a six-viseme SVM classifier trained on the same data, with feature labels converted to viseme labels. The mapping between visemes and features is the same as in Table 6.3. Again, only these six feature combinations are represented in the manually labeled training data. For additional information regarding the training and performance of these SVMs, as well as the data preprocessing, see [SLS+05].

6.3.4 Phrase recognition

In this section, we evaluate various phrase recognizers on the stereo control command task. The experimental setup is as follows. Recall that the two speakers spoke

each stereo control command at slow, medium and fast speeds. We repeat each experiment three times, training the system on two speed conditions and testing it on the remaining condition, and average the accuracies over the three trials.

The rightmost three columns of Table 6.6 (labeled *whole-phrase*) compare the feature-based model of Figure 6-5 to the viseme-based HMM model of Figure 6-6, referred to as *viseme+GM*. We evaluate two versions of the feature-based model: one with strict synchrony enforced between the feature streams, i.e. $async_t^{A;B} = 0 \forall A, B$ (*feature+GM*), and one with some asynchrony allowed between the streams (*async feature+GM.*) In the model with asynchrony, **LR** and **LO** are allowed to de-synchronize by up to one index value (one state), as are **LO** and **LD**. The two asynchrony probabilities, $p(async_t^{LR;LO} = 1)$ and $p(async_t^{LO;LD} = 1)$, are learned from the training data. All three of the above systems use whole-phrase units and Gaussian models (GMs) of observations (in this case, single Gaussians with tied diagonal covariance matrices.) On average, the models using feature classifiers outperform the ones using viseme classifiers, and the asynchronous feature-based systems slightly outperform the synchronous ones.

Looking at each of the train and test conditions in more detail, we see that, when the training set includes the slow condition, the asynchronous feature model outperforms the synchronous one, which, in turn, outperforms the viseme baseline. (The McNemar significance levels p [Die98] for these differences are shown in the table.) However, when the models are trained on faster speech and tested on slow speech, the baseline slightly outperforms the feature-based models.

For comparison, the leftmost three columns of the table correspond to three *dictionary-based* models with the structure of Figure 6.2.1, using either soft evidence-based observation models (labeled SE) or Gaussians. In particular, the dictionary-based *feature+SE* model is the model of Section 6.2. The whole-phrase models outperform the *feature+SE* model by a large margin on this task. To evaluate the relative importance of the differences between the models, we modify the *feature+SE* baseline to use single Gaussians with diagonal covariance over SVM outputs; this is the *feature+GM* dictionary-based model. We can see that, while this improves performance over using soft evidence, it is still preferable to use whole-phrase models for this task. Finally, we evaluate a dictionary-based version of the viseme HMM baseline, *viseme+GM*. As was the case for whole-phrase models, the dictionary-based model using feature classifiers outperforms the one using viseme classifiers.

6.4 Summary

In both experiments presented in this chapter, we have found that lipreading models using feature classifiers outperform viseme-based ones. We have also found that, in a real-world command recognition task, whole-phrase models outperform dictionary-based ones, and observation models consisting of Gaussians over SVM outputs outperform soft evidence-based ones.

For the purpose of fast evaluation, this work has used a limited feature set consisting only of features pertaining to the lips, ignoring other visible features such as

tongue positions. Incorporating these additional features will require some care, as they are often occluded, but may improve performance enough to make a system such as the one of Section 6.3 practical to use.

As mentioned at the beginning of this chapter, lipreading is often intended not for stand-alone applications but for combination with acoustic speech recognition. We believe a promising area for further work is the application of feature-based models to audio-visual recognition, where they can be used to address the well-known problem of audio-visual asynchrony [NLP+02]. We return to this idea when we consider future work in greater depth in Chapter 7.

Chapter 7

Discussion and conclusions

In this thesis we have argued for a new class of models of pronunciation variation for use in automatic speech recognition. The main distinguishing aspects of the proposed class of models are

- A representation of words using multiple streams of linguistic features rather than a single stream of phones. Many common pronunciation effects can be described in terms of small amounts of asynchrony and changes in individual feature values. This is not a new idea; however, the use of this idea in a pronunciation model is new.
- The direct modeling of the *time course* of sub-word units. Previous pronunciation models have always assumed that it is their job to produce the list of sub-word units, but not their time alignment, leaving the modeling of sub-word unit durations up to a separate duration model. We have argued that this is an artificial separation, and that sub-word units and time alignments should be modeled jointly.
- A unified probabilistic description in terms of a graphical model. This allows us to (i) naturally represent the factorization of the state space defined by feature value combinations, leading to a parsimonious model; (ii) perform recognition in a single pass, as opposed to the two-pass approaches used by some previous efforts at using multiple feature streams in recognition; and (iii) explore a potentially huge set of related models relatively quickly, using a unified computational framework.

In this chapter we describe several possible directions for future work, including refinements and extensions to the basic model and applications to new tasks, and close with our main conclusions.

7.1 Model refinements

In Chapter 3, we presented a very basic class of models incorporating the ideas of feature asynchrony and feature substitutions, and in Chapter 4 we instantiated a

particular model using a feature set based on articulatory phonology. In these models, we have assumed that the distributions for both substitutions and asynchrony are context-independent: The $async_t^{A;B}$ variables have no parents in the graph, and each surface feature value S_t^i depends only on the corresponding underlying target U_t^i. It is clear from linguistic considerations that this is a very large assumption, and it would be useful to study the use of additional dependencies, such as between surface feature values and neighboring underlying or surface values.

Studies of pronunciation variation in the ASR literature suggest some useful contextual 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 ends (codas) of syllables are far more likely to be realized non-canonically [Gre99]. Ostendorf et al. [OBB+96] have shown that pronunciation variation depends on a "hidden mode" or speaking style, which may vary during the course of an utterance. Finally, Bell et al. [BJFL+03] have shown that pronunciation variants depend on contextual factors such as nearby disfluencies (e.g., hesitations), word predictability, and utterance position. These studies are based on phone-based measurements of pronunciation variation, so that we cannot extract specific implications for feature substitution or asynchrony modeling from them. However, it is clear that there are some effects of many contextual factors on pronunciation variation.

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 [BG92]. The existence of attested phenomena such as *football* → [f uh b ao l] [GHE96] but not, as far as we know, of *haptic* → [h ae p ih k] implies that tongue-lip asynchrony may also be asymmetric.

Browman and Goldstein use evidence from linguistic and articulatory data to devise specific ordering constraints among their vocal tract variables [BG92]. 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 the automatic learning of certain aspects of the model, such as the asynchrony structure. In addition, since this model is, after all, intended not only to model pronunciation but to be part of a system that performs a specific task—namely, automatic recognition of speech—it may be the case that the optimal structure of the model for ASR purposes may differ from the "correct" linguistic interpretation.

Finally, we have implemented the model within the standard approach to speech recognition using a generative statistical model trained under the maximum likelihood criterion. This allows us to make immediate comparisons with existing baselines. However, recent work suggests that discriminative approaches may be preferable in many contexts [McD00, DB03, LMP01, Bil99, BZR+02, GMAP05]. It should in principle be straightforward to adapt the proposed model to a discriminative framework.

7.2 Additional applications

7.2.1 A new account of asynchrony in audio-visual speech recognition

One promising application for feature-based models is audio-visual speech recognition (AVSR). The modeling of audio-visual asynchrony has long been used in AVSR systems to account for such effects as anticipatory lip rounding. The prevailing approach to audio-visual asynchrony modeling is to use two-stream HMMs, in which one state stream corresponds to the visual "state" while the other corresponds to the audio "state" [NLP+02]. This implicitly assumes that there are two separate hidden processes, one generating the acoustics and another generating the visual signal. This is demonstrated in Figure 7-1, which shows a Viterbi alignment produced by this type of AVSR system. The hidden acoustic state is the current phoneme, while the visual state is the viseme. The word in this example is "housewives", and asynchrony can be seen in both the initial [h] and the medial [s]. The interpretation that this recognizer gives to this example is that the visual "process" is "ahead" of the acoustic one at the beginning of the word, giving the visual impression of an [ae] but the acoustic impression of an [h]. Similarly, in the middle of the word, the visual "process" is already producing a [w] while the acoustic one is still in the [s] state.

However, since both signals are produced by the same speech articulators, this is a rather artificial analysis. A more satisfying explanation can be given through feature-based considerations, as shown in the alignment of Figure 7-2 produced by a feature-based model. The system used here is similar to the one used in Section 5.1, with the addition of a visual observation variable depending on the visible articulators S_t^L and S_t^T. This alignment is more informative of the underlying processes: At the beginning of the word, the lips and tongue are in position for the [ae] while the voicing state is still [hh]; and in the middle of the word, the lips proceed ahead of the other articulators, giving rise to the rounded [s].

This analysis not only provides a more intuitive explanation, but, we believe, may also help to improve recognition performance. The phoneme-viseme approach does not take into account the fact that when the [s] is rounded, the acoustics are also affected, as evidenced in the spectrogram by the lower cutoff frequency of the medial [s] than of the final [z]. Similarly, the spectrum of the initial [hh] also shows formant values appropriate for [ae]. While phoneme-viseme models can take some of this variation into account using context-dependent observation models, this requires more training data and, for typical training set sizes, cannot account for longer-distance effects such as the rounding of [s] in *strawberry*.

7.2.2 Application to speech analysis

More speculatively, we can imagine the type of model we have proposed having applications in speech analysis, for both scientific exploration and more immediate applications.

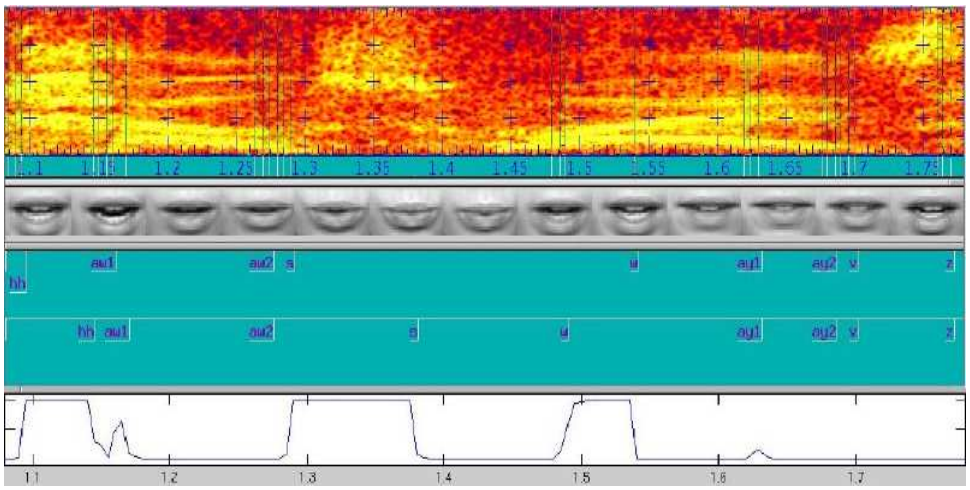

Figure 7-1: *A Viterbi alignment and posteriors of the async variables for an instance of the word* housewives*, using a phoneme-viseme system. From top to bottom, the panels are: spectrogram; lip images; values of* ph_t^{ph} *and* ph_t^{vis} *in the best path; and posterior probability of* $async_t^{ph;vis}$*.*

One possible use of the model would be to make automatic articulatory transcriptions of large amounts of recorded speech, with some bootstrapping from existing articulatory measurement data. The resulting larger transcribed set can be used to study articulation phenomena on a larger scale than is possible with existing corpora. In addition, the newly transcribed data could be recorded in a less invasive environment than is typical for articulatory measurement settings. Of course, we could only hope to measure phenomena allowed by the model. However, within the scope of effects included in the model, we could learn new aspects those phenomena, for example the relative timing of articulatory gestures or the probabilities of various reductions. Rather than building models of pronunciation phenomena from the ground up, by making measurements and devising a model that explains them, we may be able to use a bit of a priori knowledge to analyze much more data at once. Finally, the results of such analysis could be used to refine the model itself for ASR or other purposes.

As a more immediate application, we can imagine building a pronunciation analysis tool, for example for correction of non-native pronunciations. This would involve using a more detailed feature set in the model, including features that exist in both the speaker's native tongue and the target language. In an articulatory measurement study, [BG92] show that speakers of different languages use different relative timings in their articulatory movements. If successful, we can imagine such a system producing advice, say for Italian learner of English, such as "When you make a [t] or a [d], try moving the tip of your tongue farther back. You should not be able to feel your teeth with your tongue when you do this." This could provide a more time effective and comfortable language learning environment.

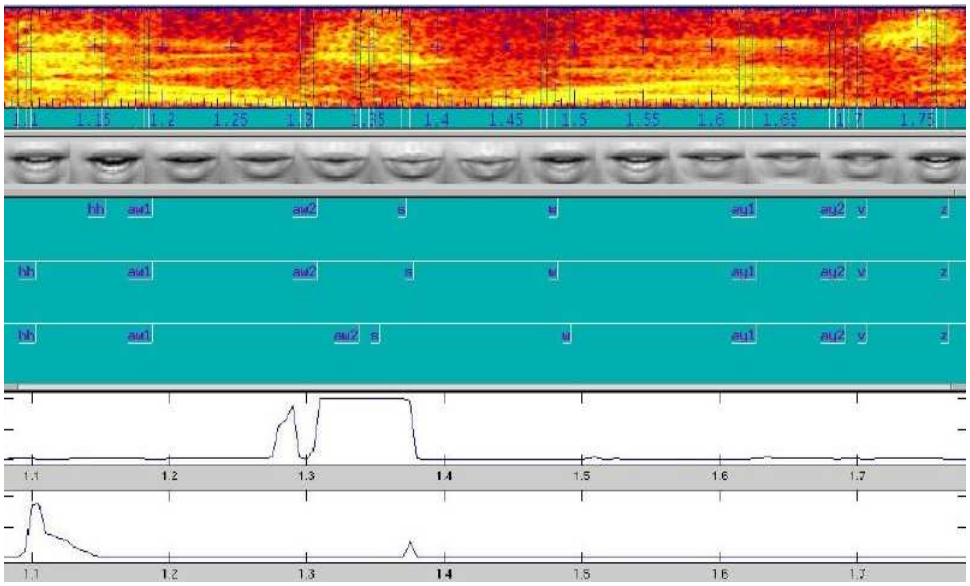

Figure 7-2: *A Viterbi alignment and posteriors of the async variables for an instance of the word* housewives*, using a feature-based recognizer with the "LTG" feature set. From top to bottom, the panels are: spectrogram; lip images; values of* ph_t^G*,* ph_t^T*, and* ph_t^L *in the best path; and posterior probabilities of* $async_t^{L;T}$ *and* $async_t^{LT;G}$*.*

7.3 Conclusions

The main contributions of this thesis are:

- Introduction of a flexible class of feature-based models for pronunciation variation, with a unified computational implementation allowing for explicit modeling of independencies in feature streams.

- Investigation of this model, along with a feature set based on articulatory phonology, in a lexical access task using manual transcriptions of conversational speech. In these experiments, we have shown that the proposed model outperforms a phone-based one in terms of coverage of observed pronunciations and ability to retrieve the correct word.

- Demonstration of the model's use in complete recognition systems for (i) landmark-based ASR and (ii) lipreading applications.

Perhaps the most important of these is the unified computational implementation. While there have been previous efforts at using feature streams for similar purposes, they have typically been forced to make rigid assumptions, have not been able to take advantage of the parsimony of the factored state space, or have used multi-pass approaches for recognition. We have, of course, made some rather strong assumptions in our initial experiments with the proposed class of models. However, it is straightforward to refine or extend the model with additional variables and dependencies as

the need arises. The graphical model implementation means that, to a large extent, we need not write new algorithms for new model varieties, greatly facilitating experimentation. We have not attempted to provide a new state of the art in automatic speech recognition, and it is likely that model refinements will be needed before significant ASR improvements on standard, well-studied tasks can be achieved. It is our hope that the framework we have proposed will facilitate such future investigation.

Appendix A

Phonetic alphabet

label	description	example	transcription
[iy]	high front tense	sweet	[s w iy t]
[ih]	high front lax	Bill	[b ih l]
[ey]	middle front tense	ate	[ey t]
[eh]	middle front lax	head	[h eh d]
[ae]	low front lax	after	[ae f t er]
[er]	high central lax r-colored (stressed)	bird	[b er d]
[axr]	high central lax r-colored (unstressed)	creature	[k r iy ch axr]
[uh]	middle central lax (stressed)	butter	[b uh dx axr]
[ax]	middle central lax (unstressed)	about	[ax b aw t]
[ay]	low central tense diphthong	kite	[k ay t]
[aw]	low central lax	flower	[f l aw er]
[aa]	low back lax	hot	[h aa t]
[ow]	middle back tense rounded	goat	[g ow t]
[oy]	middle back tense rounded diphthong	toy	[t oy]
[ao]	middle back lax rounded	bought	[b ao t]
[uw]	high back tense rounded	smooth	[s m uw dh]
[uh]	high back lax rounded	wood	[w uh d]

Table A.1: *The vowels of the ARPABET phonetic alphabet, with descriptions and examples. Based on http://www.billnet.org/phon/arpabet.php*

label	description	example	transcription
[p]	voiceless bilabial stop	put	[p uh t]
[t]	voiceless alveolar stop	top	[t aa p]
[k]	voiceless velar stop	crazy	[k r ey z iy]
[b]	voiced bilabial stop	buy	[b ay]
[d]	voiced alveolar stop	dull	[d uh l]
[g]	voiced velar stop	bug	[b uh g]
[m]	voiced bilabial nasal	mouth	[m aw th]
[n]	voised alveolar nasal	night	[n ay t]
[ng]	voiced velar nasal	sing	[s ih ng]
[f]	voiceless labiao dental fricative	find	[f ay n d]
[v]	voiced labio dental fricative	vine	[v ay n]
[th]	voiceless dental fricative	cloth	[k l aa th]
[dh]	voiced dental fricative	clothe	[k l ow dh]
[s]	voiceless alveolar fricative	see	[s iy]
[z]	voiced alveolar fricative	zoo	[z uw]
[sh]	voiceless palato-alveolar fricative	cash	[k ae sh]
[zh]	voiced palato-alveolar fricative	leisure	[l iy zh axr]
[ch]	voiceless palato-alveolar affricate	chicken	[ch ih k ih n]
[jh]	voiced palato-a;veolar affricate	judge	[jh uh jh]
[l]	voiced alveolar lateral	liquid	[l ih k w ih d]
[w]	voiced bilabial approximant	water	[w ah dx axr]
[r]	voiced alveolar approximate	round	[r aw n d]
[y]	voiced velar approximant	year	[y iy r]
[h]	voiceless glottal fricative	happy	[h ae p iy]
[q]	voiceless glottal stop	kitten	[k ih q n]
[dx]	voiceless tap (allophone of /t/)	latter	[l ae dx er]

Table A.2: *The consonants of the ARPABET phonetic alphabet, with descriptions and examples. Based on http://www.billnet.org/phon/arpabet.php.*

Appendix B

Feature sets and phone-to-feature mappings

Feature name	Description	# values	value = meaning
LIP-LOC	position (roughly, horizontal displacement) of the lips	3	PRO = protruded (rounded) LAB = labial (default/neutral position) DEN = dental (labio-dental position)
LIP-OPEN	degree of opening of the lips	4	CL = closed CR = critical (labial/labio-dental fricative) NA = narrow (e.g., [w], [uw]) WI = wide (all other sounds)
TT-LOC	location of the tongue tip	4	DEN = inter-dental (e.g., [th], [dh]) ALV = alveolar (e.g., [t], [n]) P-A = palato-alveolar (e.g., [sh]) RET = retroflex (e.g., [r])
TT-OPEN	degree of opening of the tongue tip	6	CL = closed (stop consonant) CR = critical (fricative, e.g. [s]) NA = narrow (e.g. [r] or alveolar glide) M-N = medium-narrow MID = medium WI = wide
TB-LOC	location of the tongue body	4	PAL = palatal (e.g. [sh], [y]) VEL = velar (e.g., [k], [ng]) UVU = uvular (default/neutral position) PHA = pharyngeal (e.g. [aa])
TB-OPEN	degree of opening of the tongue body	6	CL = closed (stop consonant) CR = critical (e.g. fricated [g] in "legal") NA = narrow (e.g. [y]) M-N = medium-narrow MID = medium WI = wide
VEL	state of the velum	2	CL = closed (non-nasal) OP = open (nasal)
GLOT	state of the glottis	3	CL = closed (glottal stop) CR = critical (voiced) OP = open (voiceless)

Table B.1: *Definition of the articulatory phonology-based feature set.*

phone	LIP-LOC	LIP-OPEN	TT-LOC	TT-OPEN	TB-LOC	TB-OPEN	VEL	GLOT
aa	LAB	W	ALV	W	PHA	M-N	CL(.9),OP(.1)	CR
ae	LAB	W	ALV	W	VEL	W	CL(.9),OP(.1)	CR
ah	LAB	W	ALV	M	UVU	M	CL(.9),OP(.1)	CR
ao	PRO	W	ALV	W	PHA	M-N	CL(.9),OP(.1)	CR
aw1	LAB	W	ALV	W	VEL	W	CL(.9),OP(.1)	CR
aw2	PRO	N	P-A	W	UVU	M-N	CL(.9),OP(.1)	CR
ax	LAB	W	ALV	M	UVU	M	CL(.9),OP(.1)	CR
axr	LAB	W	RET	CR(.1),N(.8), M-N(.1)	VEL(.1),UVU(.8), PHA(.1)	CL(.1),CR(.2), M-N(.1),M(.1),W(.5)	CL(.9),OP(.1)	CR
ay1	LAB	W	ALV	W	PHA	M-N	CL(.9),OP(.1)	CR
ay2	LAB	W	ALV	M-N	PAL	M-N	CL(.9),OP(.1)	CR
b	LAB	CR	ALV	M	UVU	W	CL	CR
bcl	LAB	CL	ALV	M	UVU	W	CL	CR
ch	LAB	W	P-A	CR	PAL	M-N	CL	W
d	LAB	W	ALV	CR	VEL	M	CL	CR
dcl	LAB	W	ALV	CL	VEL	M	CL	CR
dh	LAB	W	DEN	CR	UVU	M	CL	CR
dx	LAB	W	ALV	N	VEL	M	CL	CR
eh	LAB	W	ALV	M	PAL	M	CL(.9),OP(.1)	CR
el	LAB	W	ALV	CL	UVU	N	CL(.9),OP(.1)	CR
em	LAB	CL	ALV	M	UVU	M	OP	CR
en	LAB	W	ALV	CL	UVU	M	OP	CR
er	LAB	W	RET	CR(.1),N(.8), M-N(.1)	VEL(.1),UVU(.8), PHA(.1)	CL(.1),CR(.2), M-N(.1),M(.1),W(.5)	CL(.9),OP(.1)	CR
ey1	LAB	W	ALV	M	PAL	M	CL(.9),OP(.1)	CR
ey2	LAB	W	ALV	M-N	PAL	M-N	CL(.9),OP(.1)	CR
f	DEN	CR	ALV	M	VEL	M	CL	W
g	LAB	W	P-A	W	VEL	CR	CL	CR
gcl	LAB	W	P-A	W	VEL	CL	CL	CR
hh	LAB	W	ALV	M	UVU	M	CL	W
ih	LAB	W	ALV	M-N	PAL	M-N	CL(.9),OP(.1)	CR
iy	LAB	W	ALV	M-N	PAL	N	CL(.9),OP(.1)	CR
jh	LAB	W	P-A	CR	PAL	M	CL	CR
k	LAB	W	P-A	W	VEL	CR	CL	W
kcl	LAB	W	P-A	W	VEL	CL	CL	W
l	LAB	W	ALV	CL	UVU	N	CL(.9),OP(.1)	CR
m	LAB	CL	ALV	M	UVU	M	OP	CR
n	LAB	W	ALV	CL	UVU	M	OP	CR
ng	LAB	W	P-A	W	VEL	CL	OP	CR
ow1	PRO	W	P-A	W	UVU	M-N	CL(.9),OP(.1)	CR
ow2	PRO	N	P-A	W	VEL	N	CL(.9),OP(.1)	CR
oy1	PRO	W	ALV	W	UVU	M-N	CL(.9),OP(.1)	CR
oy2	LAB	W	ALV	M-N	PAL	M-N	CL(.9),OP(.1)	CR
p	LAB	CR	ALV	M	UVU	W	CL	W
pcl	LAB	CL	ALV	M	UVU	W	CL	W
r	LAB	W	RET	CR(.1),N(.8), M-N(.1)	VEL(.1),UVU(.8), PHA(.1)	CL(.1),CR(.2), M-N(.1),M(.1),W(.5)	CL(.9),OP(.1)	CR
s	LAB	W	ALV	CR	UVU	M	CL	W
sh	LAB	W	P-A	CR	PAL	M-N	CL	W
t	LAB	W	ALV	CR	VEL	M	CL	W
tcl	LAB	W	ALV	CL	VEL	M	CL	W
th	LAB	W	DEN	CR	UVU	M	CL	W
uh	PRO	W	P-A	W	UVU	M-N	CL(.9),OP(.1)	CR
uw	PRO	N	P-A	W	VEL	N	CL(.9),OP(.1)	CR
v	DEN	CR	ALV	M	VEL	M	CL	CR
w	PRO	N	P-A	W	UVU	N	CL(.9),OP(.1)	CR
y	LAB	W	ALV	M-N	PAL	N	CL(.9),OP(.1)	CR
z	LAB	W	ALV	CR	UVU	M	CL	CR
zh	LAB	W	P-A	CR	PAL	M	CL	CR
epi	PRO	CL	DEN	CL	PAL	N	CL	CL
sil	DEN	CL	DEN	CL	PAL	CL	CL	CL
dn	LAB	W	ALV	CR	VEL	M	CL(.9),OP(.1)	CR
dcln	LAB	W	ALV	CL	VEL	M	CL(.9),OP(.1)	CR
tn	LAB	W	ALV	CR	VEL	M	CL(.9),OP(.1)	W
tcln	LAB	W	ALV	CL	VEL	M	CL(.9),OP(.1)	W

Table B.2: *Mapping from phones to underlying (target) articulatory feature values. Entries of the form "$x(p_1), y(p_2), \ldots$" indicate that the feature's value is x with probability p_1, y with probability p_2, and so on. Diphthongs have been split into two phones each (e.g. [ay1] and [ay2]), corresponding to the starting and ending articulatory configurations of the diphthong. [dcln], [dn], [tcln], and [tn] refer to post-nasal stops; they are included to account for effects such as* finding $\rightarrow$ *[f ay n ih ng].*

feature subset	values
L	0=P-N, 1=P-W, 2=L-CL, 3=L-CR, 4=L-W, 5=D-CR
T	0=D-CR-U-M, 1=A-CL-U-N, 2=A-CL-U-M, 3=A-CR-U-M, 4=A-N-U-M, 5=A-MN-PA-N, 6=A-MN-PA-MN, 7=A-M-PA-M, 8=A-M-U-M, 9=A-W-V-M, 10=P-CR-PA-MN, 11=P-M-U-MN, 12=P-W-V-CL, 13=P-W-V-CR, 14=P-W-V-N, 15=P-W-U-N, 16=P-W-U-MN, 17=P-W-PH-MN, 18=R-N-U-M
G	0=C-CR, 1=C-O, 2=O-CR

Table B.3: *Definition of feature values used in SVitchboard experiments, in terms of abbreviated feature value labels from Table B.1 in the following orders: L = LL-LO; T = TTL-TTO-TBL-TBO; G = V-G. For example, L=0 corresponds to protruded lips with a narrow opening (as for a [w]); T=D-CR-U-M corresponds to an interdental tongue tip with a critical opening, and tongue body in the uvular location with a medium opening (as for a [th]).*

SE?	variable	context	definition
0	clo	nil	!actualVEL(0) && (actualLIP-OPEN(0)=0 || (actualTT-OPEN(0)=0 && !(TTPhone(0)=L)) || (actualTB-OPEN(0)=0))
0	hh	nil	actualLIP-OPEN(0)>2 && actualTT-OPEN(0)>2 && actualTB-OPEN(0)>2 && actualGLOT(0)=2
0	voi	nil	actualGLOT(0)=1
0	stri	Fr(0)	actualTT-OPEN(0)=1 && (actualTT-LOC(0)=1 && actualTT-LOC(0)=2)
1	Silence	nil	clo(0) || TBPhone(0)=SIL
1	Sonor	!Silence(0)	hh(0) || (actualGLOT(0)=1 && ((actualLIP-OPEN(0)>1 && (actualTT-OPEN(0)>1 || TTPhone(0)=L || TTPhone(0)=EL) && actualTB-OPEN(0)>1) || (actualVEL(0)=1 && (actualLIP-OPEN(0)=0 || actualTT-OPEN(0)=0 || actualTB-OPEN(0)=0))))
1	SC	Sonor(0)	((actualLIP-OPEN(0)<3 && !(LIPPhone(0)=EM) && !(LIPPhone(0)=UW) && !(LIPPhone(0)=OW2)) || (actualTT-OPEN(0)<3 && !(TTPhone(0)=EN) && !(TTPhone(0)=ER) && !(TTPhone(0)=AXR) && !(TTPhone(0)=EL)) || (actualTB-OPEN(0)<3 && !(TBPhone(0)=AXR) && !(TBPhone(0)=EL) && !(TBPhone(0)=IY) && !(TBPhone(0)=UW) && !(TBPhone(0)=OW2))) || hh(0)
0	syl	Sonor(0)	!SC(0)
0	NC	SC(0)	actualVEL(0)=1 && (actualLIP-OPEN(0)=0 || actualTT-OPEN(0)=0 || actualTB-OPEN(0)=0)
0	LG	SC(0)	!NC(0) || hh(0)
1	Stops	!Sonor(0)	clo(0) && !clo(1)
0	Fr	!Silence(0) && !Sonor(0)	!Stops(0) || hh(0)
0	StriFr	Fr(0)	TT-LOC(0)=1
0	actualAA_AY1_AO	syl(0)	TT-LOC(0)=1 && TT-OPEN(0)=5 && TB-LOC(0)=3 && TB-OPEN(0)=3
0	actualAE_AW1	syl(0)	TT-LOC(0)=1 && TT-OPEN(0)=5 && TB-LOC(0)=1 && TB-OPEN(0)=5
0	actualAH_AX	syl(0)	TT-LOC(0)=1 && TT-OPEN(0)=4 && TB-LOC(0)=2 && TB-OPEN(0)=4
0	actualAW2_OW1_UH	syl(0)	TT-LOC(0)=2 && TT-OPEN(0)=5 && TB-LOC(0)=2 && TB-OPEN(0)=3
0	actualAXR	syl(0)	TT-LOC(0)=3
0	actualAY2_IH_EY2_OY2	syl(0)	TT-LOC(0)=1 && TT-OPEN(0)=3 && TB-LOC(0)=0 && TB-OPEN(0)=3
0	actualEH_EY1	syl(0)	TT-LOC(0)=1 && TT-OPEN(0)=4 && TB-LOC(0)=0 && TB-OPEN(0)=4
0	actualIY	syl(0)	TT-LOC(0)=1 && TT-OPEN(0)=3 && TB-LOC(0)=0 && TB-OPEN(0)=2
0	actualOW2_UW	syl(0)	TT-LOC(0)=2 && TT-OPEN(0)=5 && TB-LOC(0)=1 && TB-OPEN(0)=2
0	actualOY1	syl(0)	TT-LOC(0)=1 && TT-OPEN(0)=5 && TB-LOC(0)=2 && TB-OPEN(0)=3
0	actualDX	SC(0)	TT-LOC(0)=TT_ALV && TT-OPEN(0)=2
0	FrVBoundary	nil	Fr(0) && syl(1)
0	VFrBoundary	nil	Fr(0) && syl(-1)
0	StriFrVBoundary	FrVBoundary(0)	stri(0)
0	VStriFrBoundary	VFrBoundary(0)	stri(0)
0	SCVBoundary	nil	SC(0) && syl(1)
0	VSCBoundary	nil	SC(0) && syl(-1)
0	NCVBoundary	nil	NC(0) && syl(1)
0	VNCBoundary	nil	NC(0) && syl(-1)
0	LGVBoundary	nil	LG(0) && syl(1)
0	VLGBoundary	nil	LG(0) && syl(-1)
0	VStBoundary	nil	clo(0) && syl(-1)
1	AspirationPreVocalic	FrVBoundary(0)	hh(0)
1	StopVoicingPrevocalic	Stops(0)	voi(0)
1	StopVelarPrevocalic	Stops(0)	actualTB-LOC(0)=TB_VEL && actualTB-OPEN(0)=0
1	StopAlveolarPrevocalic	Stops(0)	actualTT-LOC(0)=TT_ALV && actualTT-OPEN(0)=0
1	StopLabialPrevocalic	Stops(0)	actualLIP-OPEN(0)=0
1	FricVoicingPrevocalic	FrVBoundary(0)	voi(0)
1	FricStridentPrevocalic	FrVBoundary(0)	stri(0)
1	FricAnteriorPrevocalic	StriFrVBoundary(0)	actualTT-LOC(0)=TT_ALV

Table B.4: *Mapping from articulatory features to distinctive features. Each row represents a variable. The first column indicates whether or not we have soft evidence for the variable (in the form of likelihoods computed from SVM discriminant values). The second column gives the name of the variable. The third column describes the context in which the variable is relevant, expressed as a regular expression over time-indexed variables. Finally, the fourth column contains a regular expression giving the value of the variable in terms of other previously-defined variables. For example, the variable "VStBoundary" is one for which we do not have soft evidence, it is relevant in all contexts (indicated by "nil" in the context column), and its value is 1 when "clo" is 1 in the current frame and "syl" is 1 in the previous frame; and "FricLabialPostvocalic" is a variable for which we do have a classifier, it is relevant in frames corresponding to vowel-fricative boundaries, and its value is 1 if "actualLIP-OPEN" is 1 (critical) in the current frame. The variables for which we do not have SVMs are simply "helper" variables, used to more concisely define regular expressions for other variables.*

SE?	variable	context	definition
1	VowelHigh	syl(0)	actualIY(0) || actualOW2_UW(0)
1	LateralPrevocalic	LGVBoundary(0)	actualTT-LOC(0)=1 && actualTT-OPEN(0)=0 && actualTB-LOC(0)=2 && actualTB-OPEN(0)=2
1	RhoticPrevocalic	LGVBoundary(0)	actualTT-LOC(0)=TT_RET
1	RoundPrevocalic	LGVBoundary(0)	actualLIP-OPEN(0)=2
1	YPrevocalic	LGVBoundary(0)	actualTB-LOC(0)=TB_PAL && actualTB-OPEN(0)=2
1	LateralPostvocalic	VLGBoundary(0)	actualTT-LOC(0)=1 && actualTT-OPEN(0)=0 && actualTB-LOC(0)=2 && actualTB-OPEN(0)=2
1	RhoticPostvocalic	VLGBoundary(0)	actualTT-LOC(0)=TT_RET
1	RoundPostvocalic	VLGBoundary(0)	actualLIP-OPEN(0)=2
1	YPostvocalic	VLGBoundary(0)	actualTB-LOC(0)=TB_PAL && actualTB-OPEN(0)=2
1	StridentIsolated	Fr(0)	stri(0)
1	FricLabialPostvocalic	VFrBoundary(0)	actualLIP-OPEN(0)=1
1	FricLabialPrevocalic	FrVBoundary(0)	actualLIP-OPEN(0)=1
1	Rhotic	LG(0)	actualTT-LOC(0)=TT_RET
1	Lateral	LG(0)	actualTT-LOC(0)=1 && actualTT-OPEN(0)=0 && actualTB-LOC(0)=2 && actualTB-OPEN(0)=2
1	Round	LG(0)	actualLIP-OPEN(0)=2
1	Body	LG(0)	actualTB-LOC(0)=TB_PAL && actualTB-OPEN(0)=2
1	VowelNasal	syl(0)	actualVEL(0)=1 && (actualLIP-OPEN(0)=0 || actualTT-OPEN(0)=0 || actualTB-OPEN(0)=0)
1	VowelRhotic	syl(0)	actualTT-LOC(0)=TT_RET
1	VowelLateral	syl(0)	actualTT-LOC(0)=1 && actualTT-OPEN(0)=0 && actualTB-LOC(0)=2 && actualTB-OPEN(0)=2
1	VowelRound	syl(0)	actualLIP-OPEN(0)=2
1	VowelBody	syl(0)	actualTB-LOC(0)=TB_PAL
1	VowelTenseHigh	syl(0)	actualIY(0) || actualEH_EY1(0) || actualOW2_UW(0) || actualAW2_OW1_UH(0)
1	VowelTenseLow	syl(0)	actualAA_AY1_AO(0) || actualOY1(0) || actualAE_AW1(0)
1	GlideAspiration	LG(0)	!voi(0)
1	StopVoicingPostvocalic	VStBoundary(0)	voi(0)
1	FricPalatal	Fr(0)	actualTB-OPEN(0)=1 && actualTB-LOC(0)=TB_PAL
1	FricDental	Fr(0)	actualTT-OPEN(0)=1 && actualTT-LOC(0)=TT_DEN
1	FlapPrevocalic	SCVBoundary(0)	actualDX(0)
1	FlapPostvocalic	VSCBoundary(0)	actualDX(0)
1	FlapFrame	SC(0)	actualDX(0)
1	aaNasalization	syl(0) && actualAA_AY1_AO(0)	actualVEL(0)=1
1	aeNasalization	syl(0) && actualAE_AW1(0)	actualVEL(0)=1
1	axNasalization	syl(0) && actualAH_AX(0)	actualVEL(0)=1
1	ehNasalization	syl(0) && actualEH_EY1(0)	actualVEL(0)=1
1	ihNasalization	syl(0) && actualAY2_IH_EY2_OY2(0)	actualVEL(0)=1
1	iyNasalization	syl(0) && actualIY(0)	actualVEL(0)=1
1	owNasalization	syl(0) && actualAW2_OW1_UH(0) || actualOW2_UW(0)	actualVEL(0)=1
1	oyNasalization	syl(0) && actualOY1(0) || actualAY2_IH_EY2_OY2(0)	actualVEL(0)=1
1	uwNasalization	syl(0) && actualOW2_UW(0)	actualVEL(0)=1
1	NasalPrevocalic	SCVBoundary(0)	NC(0)
1	NasalPostvocalic	VSCBoundary(0)	NC(0)
1	NasalLabialPrevocalic	NCVBoundary(0)	actualLIP-OPEN(0)=0
1	NasalAlveolarPrevocalic	NCVBoundary(0)	actualTT-OPEN(0)=0 && actualTT-LOC(0)=TT_ALV
1	NasalVelarPrevocalic	NCVBoundary(0)	actualTB-OPEN(0)=0 && actualTB-LOC(0)=TB_VEL
1	NasalLabialPostvocalic	VNCBoundary(0)	actualLIP-OPEN(0)=0
1	NasalAlveolarPostvocalic	VNCBoundary(0)	actualTT-OPEN(0)=0 && actualTT-LOC(0)=TT_ALV
1	NasalVelarPostvocalic	VNCBoundary(0)	actualTB-OPEN(0)=0 && actualTB-LOC(0)=TB_VEL
1	FricVoicingPostvocalic	VFrBoundary(0)	voi(0)
1	FricStridentPostvocalic	VFrBoundary(0)	stri(0)
1	FricAnteriorPostvocalic	VStriFrBoundary(0)	actualTT-LOC(0)=TT_ALV
1	StopVelarPostvocalic	VStBoundary(0)	actualTB-LOC(0)=TB_VEL && actualTB-OPEN(0)=0
1	StopAlveolarPostvocalic	VStBoundary(0)	actualTT-LOC(0)=TT_ALV && actualTT-OPEN(0)=0
1	StopLabialPostvocalic	VStBoundary(0)	actualLIP-OPEN(0)=0

Table B.5: *Mapping from articulatory features to distinctive features, continued.*

Bibliography

[AHD00] *The American Heritage Dictionary of the English Language*. Houghton Mifflin, Boston, fourth edition, 2000.

[Alb58] R. W. Albright. The international phonetic alphabet: Its background and development. *International journal of American linguistics*, 24(1, part 3), 1958.

[Bat04] R. Bates. *Speaker dynamics as a source of pronunciation variability for continuous speech recognition models*. PhD dissertation, University of Washington, Seattle, WA, 2004.

[BD96] H. Bourlard and S. Dupont. A new ASR approach based on independent processing and recombination of partial frequency bands. In *Proc. ICSLP*, Philadelphia, October 1996.

[BG86] C. P. Browman and L. Goldstein. Towards an articulatory phonology. *Phonology Yearbook*, 3:219–252, 1986.

[BG89] C. P. Browman and L. Goldstein. Articulatory gestures as phonological units. *Phonology*, 6:201–251, 1989.

[BG90a] C. P. Browman and L. Goldstein. Gestural specification using dynamically-defined articulatory structures. *Journal of Phonetics*, 18:299–320, 1990.

[BG90b] C. P. Browman and L. Goldstein. Tiers in articulatory phonology, with some implications for casual speech. In T. Kingston and M. E. Beckman, editors, *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, pages 341–376. Cambridge University Press, Cambridge, UK, 1990.

[BG92] C. P. Browman and L. Goldstein. Articulatory phonology: An overview. *Phonetica*, 49:155–180, 1992.

[BGK+84] C. P. Browman, L. Goldstein, J. A. S. Kelso, P. Rubin, and E. Saltzman. Articulatory synthesis from underlying dynamics. *JASA*, 75:S22–S23, 1984.

[Bil99] J. Bilmes. *Natural Statistical Models for Automatic Speech Recognition*. PhD dissertation, U. C. Berkeley, Berkeley, CA, 1999.

[Bil00] J. Bilmes. Dynamic Bayesian multinets. In *16th Conference on UAI*, Stanford, CA, July 2000.

[Bil03] J. Bilmes. Graphical models and automatic speech recognition. In M. Johnson, S. P. Khudanpur, M. Ostendorf, and R. Rosenfeld, editors, *Mathematical Foundations of Speech and Language Processing*, The IMA Volumes in Mathematics and its Applications. Springer-Verlag, 2003.

[Bil04] J. A. Bilmes. On soft evidence in Bayesian networks. Technical Report UWEETR-2004-0016, 2004, University of Washington Dept. of Electrical Engineering, 2004.

[BJFL+03] A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, M. Gregory, and D. Gildea. Effects of disfluencies, predictability, and utterance position on word form variation in english conversation. *JASA*, 113(2):1001–1024, 2003.

[BJM83] L. R. Bahl, F. Jelinek, and R. L. Mercer. A maximum likelihood approach to continuous speech recognition. *Proc. IEEE Trans. PAMI*, 5(2):179–190, 1983.

[Bla96] S. Blackburn. *Articulatory methods for speech production and recognition*. PhD dissertation, Cambridge University Engineering Department, Cambridge, UK, 1996.

[BM94] H. Bourlard and N. Morgan. *Connectionist Speech Recognition - A Hybrid Approach*. Kluwer Academic Publishers, 1994.

[BPSW70] L. E. Baum, T. Peterie, G. Souled, and N. Weiss. A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains. *Ann. Math. Statist.*, 41(1):164–171, 1970.

[BY01] C. Blackburn and S. Young. Enhanced speech recognition using an articulatory production model trained on x-ray data. *Computer Speech and Language*, 15(3), 2001.

[BZ02] J. Bilmes and G. Zweig. The Graphical Models Toolkit: An open source software system for speech and time-series processing. In *Proc. ICASSP*, Orlando, FL, May 2002.

[BZR+02] J. Bilmes, G. Zweig, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne. Structurally discriminative graphical models for automatic speech recognition. In *Proc. ICASSP*, Orlando, FL, May 2002.

[CH68] N. Chomsky and M. Halle. *The Sound Pattern of English*. Harper & Row, New York, NY, 1968.

[DB03] V. Doumpiotis and W. Byrne. Lattice segmentation and minimum bayes risk discriminative training for large vocabulary continuous speech recognition. In *Proc. Eurospeech*, Geneva, Switzerland, September 2003.

[DFA03] K. Daoudi, D. Fohr, and C. Antoine. Dynamic Bayesian networks for multi-band automatic speech recognition. *Computer Speech and Language*, 17(2–3):263–285, 2003.

[Die98] T. G. Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, 10(7):1895–1924, 1998.

[DK89] T. Dean and K. Kanazawa. A model for reasoning about persistence and causation. *Computational Intelligence*, 5:142–150, 1989.

[DLR77] 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):1–38, 1977.

[DRS97] L. Deng, G. Ramsay, and D. Sun. Production models as a structural basis for automatic speech recognition. *Speech Communication*, 33:93–111, 1997.

[ea04] M. Hasegawa-Johnson et al. Landmark-based speech recognition: Final report. Technical report, JHU CLSP, 2004.

[ea05] M. Hasegawa-Johnson et al. Landmark-based speech recognition. In *ICASSP*, 2005.

[EF94] K. Erler and G. H. Freeman. An articulatory-feature-based hmm for automated speech recognition. In *Acoustical Society of America, 127th Meeting*, June 1994.

[Eid01] E. Eide. Distinctive features for use in an automatic speech recognition system. In *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

[Fan73] G. Fant. *Speech Sounds and Features*. Current studies in linguistics. MIT Press, Cambridge, MA, 1973.

[FK01] J. Frankel and S. King. Asr - articulatory speech recognition. In *Proc. Eurospeech*, Aalborg, Denmark, September 2001.

[FL99] J. E. Fosler-Lussier. *Dynamic Pronunciation Models for Automatic Speech Recognition*. PhD dissertation, U. C. Berkeley, Berkeley, CA, 1999.

[FW97] M. Finke and A. Waibel. Speaking mode dependent pronunciation modeling in large vcabulary conversational speech recognition. In *Proc. Eurosepeech*, Rhodes, Greece, September 1997.

[GD04] D. Grossman and P. Domingos. Learning bayesian network classifiers by maximizing conditional likelihood. In *Proc. ICML*, Banff, Canada, July 2004.

[GGS97] R. Greiner, A. Grove, and D. Schuurmans. Learning bayesian nets that perform well. In *Proc. UAI*, Providence, RI, August 1997.

[GH96] D. Geiger and D. Heckerman. Beyond Bayesian networks: Similarity networks and Bayesian multinets. *Artificial Intelligence*, 82:45–74, 1996.

[GHE96] S. Greenberg, J. Hollenback, and D. Ellis. Insights into spoken language gleaned from phonetic transcription of the switchboard corpus. In *Proc. ICSLP*, Philadelphia, October 1996.

[GHM92] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *Proc. ICASSP*, San Francisco, March 1992.

[GJ97] Z. Gharamani and M. I. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–273, 1997.

[Gla03] J. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 17:137–152, 2003.

[GLF+93] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, D. S. Pallett, and N. L. Dahlgren. Darpa timit acoustic-phonetic continuous speech corpus cd-rom, nist, 1993.

[GMAP05] A. Gunawardana, M. Mahajan, A. Acero, and J. C. Platt. Hidden conditional random fields for phone classification. In *Proc. Interspeech*, Lisbon, Portugal, September 2005.

[GMT] The Graphical Models Toolkit. http://ssli.ee.washington.edu/ bilmes/gmtk/.

[Gol76] J. A. Goldsmith. *Autosegmental Phonology*. PhD dissertation, MIT Department of Linguistics, Cambridge, MA, 1976.

[Gol90] J. A. Goldsmith. *Autosegmental and metrical phonology*. Basil Blackwell, Oxford, UK, 1990.

[GPN02] G. Gravier, G. Potamianos, and C. Neti. Asynchrony modeling for audio-visual speech recognition. In *Proc. HLT*, San Diego, CA, March 2002.

[Gre99] Steven Greenberg. Speaking in shorthand – a syllable-centric perspective for understanding pronunciation variation. *Speech Communication*, 29:159–176, 1999.

[GSBB04] J. N. Gowdy, A. Subramanya, C. Bartels, and J. Bilmes. DBN-based multi-stream models for audio-visual speech recognition. In *Proc. ICASSP*, Montreal, Canada, May 2004.

[Haz] T. J. Hazen. Visual model structures and synchrony constraints for audio-visual speech recognition. *IEEE Trans. SAP*. to appear.

[HD96] C. Huang and A. Darwiche. Inference in belief networks: A procedural guide. *J. Approximate Reasoning*, 15(3):225–263, 1996.

[Hef50] R.-M. S. Heffner. *General Phonetics*. Foundations of Modern Linguistics. The University of Wisconsin Press, Madison, WI, 1950.

[HHSL05] Timothy J. Hazen, I. Lee Hetherington, Han Shu, and K. Livescu. Pronunciation modeling using a finite-state transducer representation. *Speech Communication*, 46(2):189–203, June 2005.

[HP00] H. G. Hirsch and D. Pearce. The Aurora experimental framework for the performance evaluation of speech recognition systems under noisy conditions. In *Proc. International Speech Communication Association Workshop on Automatic Speech Recognition*, Paris, September 2000.

[Huc94] M. A. Huckvale. Word recognition from tiered phonological models. In *IOA Conference in Speech and Hearing*, Windermere, November 1994.

[HZ84] D. Huttenlocher and V. Zue. A model of lexical access from partial phonetic information. In *Proc. ICASSP*, San Diego, CA, 1984.

[Jel98] F. Jelinek. *Statistical Methods for Speech Recognition*. MIT Press, Cambridge, MA, 1998.

[JEW04] A. Juneja and C. Espy-Wilson. Event-based system. In *Sound2Sense*, 2004.

[JFH52] R. Jakobson, C. Gunnar M. Fant, and M. Halle. Preliminaries to speech analysis: The distinctive features and their correlates. Technical Report 13, MIT Acoustics Laboratory, 1952.

[JGJS99] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[JM97] K. Johnson and J. W. Mullennix, editors. *Talker Variability in Speech Processing*. Academic Press, San Diego, CA, 1997.

[Joh02] K. Johnson. Abstract, Speech Communication Group seminar, October 2002.

[Jor98] M. I. Jordan. *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1998.

[JWJ+01] D. Jurafsky, W. Ward, Z. Jianping, K. Herold, Y. Xiuyang, and Z. Sen. What kind of pronunciation variation is hard for triphones to model? In *Proc. ICASSP*, Salt Lake City, UT, May 2001.

[Kai85] E. M. Kaisse. *Connected Speech*. Academic Press, Orlando, FL, 1985.

[KBB05] S. King, J. Bilmes, and C. Bartels. Svitchboard: Small-vocabulary tasks from switchboard. In *Proc. Interspeech*, Lisbon, Portugal, September 2005.

[KFS00] K. Kirchhoff, G. A. Fink, and G. Sagerer. Combining acoustic and articulatory feature information for robust speech recognition. *Speech Communication*, 37:303–319, 2000.

[Kir96] K. Kirchhoff. Syllable-level desynchronisation of phonetic features for speech recognition. In *Proc. ICSLP*, Philadelphia, PA, October 1996.

[Kla77] D. Klatt. Review of the arpa speech understanding program. *JASA*, 62:1345–1366, 1977.

[Lad01] P. Ladefoged. *A Course in Phonetics*. Harcourt, Brace, Jovanovich, 2001.

[Lau96] S. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, UK, 1996.

[LGB03] K. Livescu, J. Glass, and J. Bilmes. Hidden feature models for automatic speech recognition. In *Proc. Eurospeech*, 2003.

[LM98] B. T. Logan and P. J. Moreno. Factorial hmms for acoustic modeling. In *Proc. ICASSP*, Seattle, WA, May 1998.

[LMP01] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, Williams College, MA, July 2001.

[LR02] A. Lahiri and H. Reetz. Underspecified recognition. *Laboratory Phonology VII*, 2002.

[M-W] Merriam-Webster Online. http://www.m-w.com/.

[McD00] E. McDermott. Discriminative training for large vocabulary telephone-based name recognition. In *Proc. ICASSP*, Istanbul, Turkey, June 2000.

[McK99] D. McKay. An introduction to Monte Carlo methods. In M. I. Jordan, editor, *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

[MGSN98] D. McAllaster, L. Gillick, F. Scattone, and M. Newman. Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. In *Proc. ICSLP*, Sydney, Australia, November–December 1998.

[Mur02] K. Murphy. *Dynamic Bayesian Networks: Representation, Inference and Learning*. PhD dissertation, U. C. Berkeley, Berkeley, CA, 2002.

[MW02] F. Metze and A. Waibel. A flexible stream architecture for ASR using articulatory features. In *Proc. ICSLP*, Denver, CO, September 2002.

[NLP+02] A. V. Nefian, L. Liang, X. Pi, L. Xiaoxiang, C. Mao, and K. Murphy. A coupled HMM for audio-visual speech recognition. In *Proc. ICASSP*, Orlando, FL, May 2002.

[NY02] H. J. Nock and S. J. Young. Modelling asynchrony in automatic speech recognition using loosely-coupled HMMs. *Cognitive Science*, 26(3):283–301, may–jun 2002.

[OBB+96] M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In *Proc. ICSLP*, 1996. (supplementary paper).

[oL04] The Ohio State University Department of Linguistics. *Language Files*. Ohio State University Press, Columbus, OH, seventh edition, 2004.

[Ost99] M. Ostendorf. Moving beyond the 'beads-on-a-string' model of speech. In *Proc. IEEE ASRU Workshop*, 1999.

[Ost00] M. Ostendorf. Incorporating linguistic theories of phonological variation into speech recognition models. *Phil. Trans. Royal Society*, 358(1769):1325–1338, 2000.

[OZW+75] B. Oshika, V. Zue, R. Weeks, H. Neu, and J. Aurbach. The role of phonological rules in speech understanding research. *IEEE Trans. ASSP*, 23(1):104–112, 1975.

[Pea88] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.

[Per69] J. S. Perkell. *Physiology of speech production: results and implications of a quantitative cineradiographic study*. MIT Press, Cambridge, MA, 1969.

[PK86] J. S. Perkell and D. H. Klatt, editors. *Invariance and Variability in Speech Processes*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1986.

[RBD00] M. Richardson, J. Bilmes, and C. Diorio. Hidden-articulator markov models for speech recognition. In *Proc. International Training and Research Workshop on Automatic Speech Recognition*, Paris, September 2000.

[RBF+99] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraçlar, Charles Wooters, and George Zavaliagkos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29(2–4):209–224, November 1999.

[Ree98] H. Reetz. *Automatic speecch recognition with features*. PhD dissertation, University of the Saarland, Saarbrücken, Germany, 1998.

[RHD87] *Random House Unabridged Dictionary*. Random House, Inc., New York, second edition, 1987.

[RJ93] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Englewood Cliffs, NJ, 1993.

[RL96] M. D. Riley and Andrej Ljolje. Automatic generation of detailed pronunciation lexicons. In Chin-Hui Lee, Frank K. Soong, and Kuldip K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, pages 285–302. Kluwer Academic Publishers, Boston, 1996.

[RSCJ04] B. Roark, M. Saraçlar, M. Collins, and M. Johnson. Discriminative language modeling with conditional random fields and the perceptron algorithm. In *Proc. 42nd meeting of the ACL*, Barcelona, July 2004.

[RSS94] R. C. Rose, J. Schroeter, and M. M. Sondhi. An investigation of the potential role of speech production models in automatic speech recognition. In *Proc. ICSLP*, Yokohama, Japan, 1994.

[Sar00] M. Saraçlar. *Pronunciation Modeling for Conversational Speech Recognition*. PhD dissertation, Johns Hopkins University, Baltimore, MD, 2000.

[SC99] H. Strik and C. Cucchiarini. Modeling pronunciation variation for asr: A survey of the literature. *Speech Communication*, 29(2–4):225–246, 1999.

[Sch73] S. A. Schane. *Generative Phonology*. Foundations of Modern Linguistics. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1973.

[SH02] H. Shu and I. L. Hetherington. EM training of finite-state transducers and its application to pronunciation modeling. In *Proc. ICSLP*, Denver, CO, September 2002.

[Sho80] J. E. Shoup. Phonological aspects of speech recognition. In *Trends in Speech Recognition*, pages 125–138. Prentice Hall, Englewood Cliffs, NJ, 1980.

[SK04] M. Saraçlar and S. Khudanpur. Pronunciation change in conversational speech and its implications for automatic speech recognition. *Computer Speech and Language*, 18(4):375–395, October 2004.

[SLGD05] K. Saenko, K. Livescu, J. Glass, and T. Darrell. Multi-stream articulatory feature model for visual speech recognition. In *ICASSP*, 2005.

[SLS+05] K. Saenko, K. Livescu, M. Siracusa, K. Wilson, J. Glass, and T. Darrell. Visual speech recognition with loosely synchronized feature streams. In *ICCV*, Beijing, China, October 2005.

[SM89] E. Saltzman and K. G. Munhall. A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, 1:333–382, 1989.

[SMDB04] T. A. Stephenson, M. Magimai-Doss, and H. Bourlard. Speech recognition with auxiliary information. *IEEE Trans. SAP*, 12(3):189–203, 2004.

[SMSW03] S. Stuker, F. Metze, T. Schultz, and A. Waibel. Integrating multilingual articulatory features into speech recognition. In *Proc. Eurospeech*, Geneva, Switzerland, September 2003.

[Smy98] P. Smyth. Belief networks, hidden markov models, and markov random fields: a unifying view. *Pattern Recognition Letters*, 1998.

[Ste98] K. N. Stevens. *Acoustic Phonetics*. MIT Press, Cambridge, MA, 1998.

[Ste02] K. N. Stevens. Toward a model for lexical access based on acoustic landmarks and distinctive features. *JASA*, 111:1872–1891, 2002.

[SW89] J. Simpson and E. Weiner, editors. *The Oxford English Dictionary*. Clarendon Press, second edition, March 1989.

[SW96] T. Sloboda and A. Waibel. Dictionary learning for spontaneous speech recognition. In *Proc. ICSLP*, Philadelphia, October 1996.

[Tan05] M. Tang. *Large vocabulary continuous speech recognition using linguistic features and constraints*. PhD dissertation, MIT Department of Electrical Engineering and Computer Science, Cambridge, MA, 2005.

[WFK04] M. Wester, J. Frankel, and S. King. Asynchronous articulatory feature recognition using dynamic bayesian networks. In *Proc. IEICI Beyond HMM Workshop*, Kyoto, Japan, December 2004.

[WR00] A. Wrench and K. Richmond. Continuous speech recognition using articulatory data. In *Proc. ICSLP*, 2000.

[WTHSS96] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass. Effect of speaking style on LVCSR performance. In *Proc. ICSLP*, Philadelphia, PA, October 1996.

[WWK+96] M. Weintraub, S. Wegmann, Y.-H. Kao, S. Khudanpur, C. Galles, E. Fosler, and M. Saraçlar. Automatic learning of word pronunciation from data. In *Proc. ICSLP*, Philadelphia, PA, October 1996.

[YB03] I. W. Yoo and B. Blankenship. Duration of epenthetic [t] in polysyllabic american english words. *Journal of the International Phonetic Association*, 33(2):153–164, 2003.

[ZBR+01] G. Zweig, J. Bilmes, T. Richardson, K. Filali, K. Livescu, P. Xu, K. Jackson, Y. Brandman, E. Sandness, E. Holtz, J. Torres, and B. Byrne. Structurally discriminative graphical models for automatic speech recognition: final report. Technical report, Johns Hopkins University Center for Language and Speech Processing, 2001.

[ZDH+03] Y. Zhang, Q. Diao, S. Huang, W. Hu, C. Bartels, and J. Bilmes. DBN based multi-stream models for speech. In *Proc. ICASSP*, Hong Kong, April 2003.

[Zwe98] G. Zweig. *Speech recognition using dynamic Bayesian networks*. PhD dissertation, U. C. Berkeley, Berkeley, CA, 1998.