Speaker dynamics as a source of pronunciation variability for continuous speech recognition models

Rebecca Anne Bates

A dissertation submitted in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Washington

2004

Program Authorized to Offer Degree: Electrical Engineering

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

University of Washington

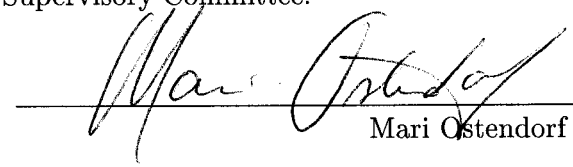
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Rebecca Anne Bates

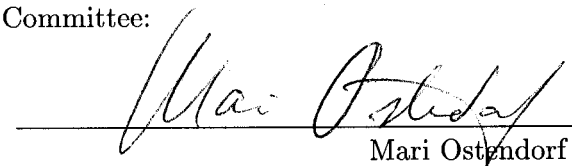
and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of Supervisory Committee:



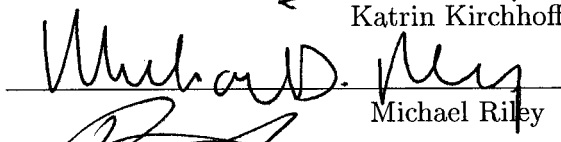
Mari Ostendorf

Reading Committee:

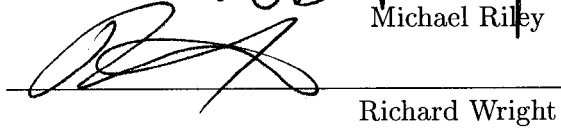


Mari Ostendorf

Katrin Kirchhoff



Michael Riley



Richard Wright

Date: 24 December 2003

In presenting this dissertation in partial fulfillment of the requirements for the Doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Bell and Howell Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature 

Date 24 December 2003

University of Washington

Abstract

Speaker dynamics as a source of pronunciation variability
for continuous speech recognition models

by Rebecca Anne Bates

Chair of Supervisory Committee:

Professor Mari Ostendorf
Electrical Engineering

A significant source of variation in spontaneous speech is due to intra-speaker pronunciation changes. Previous work has identified several factors related to pronunciation variability, such as phonetic context and speaking rate, which are useful to model in automatic speech recognition. This work examines new higher-level information sources: syntax, discourse structure and prosody, specifically the relationship between these factors and pronunciation variation as seen in reduction and hyper-articulation. The key contributions of this work include 1) analysis of high-level factors, providing new cues for improving prediction of pronunciation variation, 2) a framework for including dynamic pronunciation models in automatic speech recognition systems, and 3) an analysis of feature-based pronunciation models with suggestions for their incorporation into ASR systems.

Key findings from the analysis of high-level factors are attributes that are most useful for predicting variability, including: part-of-speech (POS) of the target word and neighboring words, location of the word in an utterance, the number of F0 slope changes within the word, word duration, and average word energy. Pronunciation prediction experiments show a reduction in phone error rate of 2.3% relative and similar reductions in perplexity over a baseline model using only phonetic context.

Incorporating higher-level information (such as hypothesis-dependent word context or

word-level F0 values) into ASR systems requires a rescoring approach. A framework for this is presented, with recognition results using various types of pronunciation models on the Switchboard task. We obtain a small but statistically significant improvement in recognition performance with a baseline static model using phonetic context but no significant gains from extending this model to incorporate POS-dependent pronunciations.

We also present a phonetic-feature-based prediction model where phones are represented by a vector of 21 symbolic features that can be on, off, unspecified or unused. Feature changes are predicted rather than phone changes, allowing for varying productions of phones, e.g., nasalized vowels. We studied feature interaction by examining different groupings of dependent features and showed that a hierarchical grouping with conditional dependencies leads to lower perplexity. We find that feature-based models are more efficient than phone-based models in the sense of requiring fewer parameters to predict variation while giving a smaller distance to the hand-labeled form and similar perplexity values.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Primary acknowledgments must go to Mari Ostendorf, my advisor, mentor, and friend, who has been a wonderful source of advice, support, and encouragement. This work, both the research and the writing, would not have been possible without her considerable input. I thank her especially for giving me a good example of mentoring. I hope to partially repay her by continuing the cycle of mentoring. I would like to thank my readers, Richard Wright, Michael Riley, and Katrin Kirchhoff, for their comments and guidance. In particular, thanks go to Richard Wright for his help with the full specification of the articulatory features used in this work and connections to the linguistic literature, and to Katrin Kirchhoff for her close reading and comments which helped make more clear the connections between linguistics and engineering.

For technical support and structure for pronunciation modeling and recognition software, models and guidance, I thank people at AT&T and JHU: Michael Riley (now at Google), Sanjeev Khudanpur, Murat Saraclar, and Zak Shafran. For prosody and discourse information tools, I must thank people at SRI and ICSI at UC Berkeley: Liz Shriberg, Andreas Stolcke, Nelson Morgan, and Eric Fossler-Lussier (now at Ohio State University). For F0 processing software, thanks to Harry Bratt, Kemal Sönmez and Andreas Stolcke and two grants awarded to SRI: NSF STIMULATE IRI-9619921 and NASA NCC 2-1256. Thanks to Stefanie Shattuck-Hufnagel for discussions about articulatory features, many examples, copies of hard-to-find papers, and macaroons.

Thanks especially for the technical and emotional support from past and current SPI and SSLI Lab people, particularly the other members of the Ostendorf Five: Dr. Randy Fish, Dr. David Palmer, Dr. Zak Shafran, and Dr. Ivan Bulyko. Thanks to Harriet Nock for a close reading of drafts, her good memory of experiments run back in 1997 and Branston pickle. Thanks to Zak Shafran, Arindam Mandal and Steve Juranich for their experimental

work upon which I built many of my recognition experiments. Thanks to Özgür Çetin, the sixth member of the O5, for perplexity clarification. Thanks to Jeremy Kahn for last minute Perl and phonetics help and for letting me bounce fine-tuning tweaks off of him. Thanks to Sarah Schwarm for close readings of papers, computer science connections, continuing the tradition of caretaking via chocolate, and feeding me Barenaked Ladies. Thanks to Scott Otterson for lab conversations that tended to the political.

Thanks to Liz Shriberg and Andreas Stolcke for introducing me to the way of awk and giving me lots of room to play with prosodic features. Additionally, thanks to Lin Chase, Rukmini Iyer, Nanette Veilleux and Angela Linse for countless blessings. Thanks to Patti Price for the helpful reminder that the point of a thesis is to finish.

I would also like to thank John Sahr, of the Electrical Engineering department, and Howard Chizeck, past chair of the EE department, for their behind-the-scenes work to help me finish this degree. Thanks to the EE department at the University of Washington, especially Frankye Jones, Tam Croswhite, Amy Feldman-Bawarshi, Ann Fuchs, Stephen Graham, Robyn Hagle, Lora Hatch, Helene Obradovich, Teri Reed, and Howard Chizeck, and to the computer support staff, especially Lee Damon, Ian Masterson, Haychoi Taing, and Sekar Thiagarajan, who made this all possible on a daily basis. Thanks to Lee Damon as well for his lunch invitations when I had the time to walk all the way off campus.

This experience has been about so much more than research on one topic. Thus, there are some people who might otherwise seem peripheral who have been very important to this process.

Thanks to the CIS department at Minnesota State University Mankato and John Frey, dean of CSET, for time, support, and their incredible patience as I finished this. Thanks to Colin Wightman for tantalizing me with interesting work but keeping it off my plate until I was done.

Thanks to the groups and people who have worked with me on teaching and learning issues, especially the participants of CS590ET, the CIS department at MSU Mankato, the freshman faculty group at MSU Mankato, the 2001-02 Huckabay Fellows, Cindy At-

man, Dean Betty Feetham, Randy Fish, Angela Linse, Eve Riskin, Nanette Veilleux, Gina Wenger, and, again, Mari Ostendorf. I am thankful for the experience of being mentored.

Thanks to the groups of people who have reminded me of the pleasures of the world outside of the lab: my teachers from the UW Dance Department, my spiritual community at St. Therese Parish, my sushi friends, Highland Figure Skating Club, the Woodshole gang, and my neighborhood support group at the Old Fifth Avenue Tavern.

The soundtrack for this dissertation was provided by Barenaked Ladies, Prince and the New Power Generation, Lyle Lovett, Magnetic Fields, k. d. lang, Penguin Cafe Orchestra, Martin Sexton, Cowboy Junkies, Badly Drawn Boy, The Mighty, Mighty Bosstones, The Shades of Praise Gospel Choir, Michael Franti and National Public Radio. Thanks to Tiffany Megargee, Teresita Heiser, Philip Clarkson, Jason Ruhl and Sarah Schwarm for CDs and introductions.

Thanks to all the people I invite to my New Year's brunch for making me feel like I could go almost anywhere and still have great friends with whom to enjoy a meal. Thanks to friends in Boston who welcome me with open arms, especially Nanette Veilleux, Ellyn Lane and Sheryl Sarokas. Thanks to Markus Josephson for knocking on my door. Thanks to Teresita Heiser for phone calls, connections to social justice, and opportunities to think about words. Thanks to Tiffany Giesler Megargee and Jennifer Ivers Holt for their presence in my life and for shouldering some of the weight of this dissertation.

Thanks to Hallie Dunham for her example of the delight of new things, for helping me learn more about learning, and for sharing her mother's time with me. Thanks to Malia Megargee for the privilege of thinking about how we raise spiritual people. Thanks to John Cooney for being the voice at the other end of the phone. Thanks to Jonathan Hardwick for help in the process of converting from a starter to a closer.

The Bates family, especially Uncle Jerry, Sarah, Joe and Zach, deserves special thanks for keeping me grounded and forcing me to think about why I did this at so many different levels. I am happy to make them proud. I dedicate this work to my extended family, teachers all of them, and to my mother, Mary Arnold Friel Bates, who first put a pencil in

my hand and set me on this path. She would have been tickled to see this day come.

This research was supported in part by the NSF, award number IIS-9618926, an Intel Ph.D. Fellowship, and by a faculty improvement grant from Minnesota State University Mankato.

"I think it's just bizarre, there are some people who want to research things just because they're curious about it."

–U.S. Congressman Patrick Toomey, 15th District, R-Penn., July 2003

Chapter 1

INTRODUCTION

When people talk, there is a lot of variability in how they speak. When looking at the speech of many people, the speakers could be female or male, young or old, healthy or sick, native or non-native speakers of the language, from different geographic regions, and with varying levels of education. When the speech is recorded, it could first be transmitted by a cell-phone, it could be recorded on a close-talking microphone or it could be recorded on a generic computer-mounted microphone. The range of possibilities make it difficult for machines to recognize speech. These are all examples of inter-speaker variability. Another aspect of variability that also confounds recognition systems is at the intra-speaker level: pronunciation variation. The same speaker could say the same word differently depending on contextual information, such as the state of the speaker's conversation. This variability and its high-level context is the focus of this dissertation. The goal is to determine relevant measures of conversation dynamics, predict the variation in pronunciation, and use the results for computer speech recognition.

1.1 Problem Background

Speaker-independent continuous speech recognition systems for high quality recordings of read speech are currently at a state where word recognition results are quite good, reliably greater than 90% accuracy on large (50,000+) vocabularies. Unfortunately, the results are not as good when the speech being recognized is spontaneous. For example, the 2002 Switchboard evaluation of spontaneous, conversational speech over telephone lines had word error rates on the order of 20% and error rates on Meeting Data (multiple speakers wearing headsets) were on the order of 35% [24]. During the 1999 DARPA Broadcast News

benchmark tests, spontaneous speech portions had word error rates (14-16%) nearly double those of the baseline condition of news announcer recordings (8-9%) [72].[1] This difference is due to many factors, but many attribute the problem primarily to the variability in how different words are pronounced. In order to improve recognition results on spontaneous speech, recognition systems need to be robust to the many different possibilities of pronunciation changes. Unfortunately, enumerating the possibilities by expanding the dictionary can lead to increased errors due to confusability, e.g., two possible additional pronunciations for the word "and" are the same as for the words "an" and "end". This work addresses the problem of pronunciation modeling in an existing speech recognition system by allowing the pronunciation model to change dynamically and by explicitly representing the different levels of information that can influence variations.

The study of speech physiology and production began with attempts to find the invariant aspects of speech for use in speech recognition. Early discoveries made it clear that the talker's message will still be conveyed even with a large amount of variability. While much of speech recognition involves statistical modeling to characterize typical realizations of sounds, it is more and more important to deal with known sources of variability in our models as applications of speech recognition become more widespread. Variation is addressed in part by adaptation, including adapting an acoustic model to a particular speaker or telephone channel, adapting a language model to fit the words said earlier in a conversation, or normalizing the cepstral vectors representing speech for different vocal tract lengths. All of these adaptation methods have yielded improvements for continuous recognition of spontaneous speech, but there is still a great deal of room for improvement, especially with new recognition evaluations moving to tasks with more conversational variety such as meeting transcription [64] rather than conversations between two individuals or computer-directed speech.

Intra-speaker variability in general is not very well addressed by adaptation. One source of variability is pronunciation, which can include such phenomena as reduction or deletions of sounds (saying "dunno" instead of "don't know"), insertions of additional sounds

[1] While the announcer was not necessarily reading, the content was outlined beforehand rather than composed spontaneously.

when a word is stressed (saying "way-ell" instead of "well"), and substitution of sounds which can happen because of the effects of surrounding words or simply because of a mistake in pronunciation. Indeed, in a phonetically hand-labeled corpus of only four hours of speech (see [31]), there were over 80 different "pronunciations" of the word "and", including many different vowel substitutions (e.g., /ax/ or /eh/ for /ae/), deletions of the final /d/, and pronunciations that were a single nasal consonant, whether /n/, /m/, or a nasal flap. McAllaster *et al.*, performed experiments using simulated data that suggest pronunciation modeling can give large wins (i.e., word error rates on the order of 5-10% when all pronunciations in the test set match the canonical dictionary pronunciations) [60]. Further, experiments done to examine the difference between spontaneous, read and acted speech show that speaking style is a major factor in recognition accuracy. Results are significantly better when the test data is read speech, even given a mismatch of spontaneous speech for training data and read speech for test data [84, 111].

In this work, pronunciation modeling will be viewed as a locally adaptive problem which is different from the typical view of adaptation. Instead of adapting to a particular speaker or a noisy environment, modeling will take into account the information and conversational context of the speech. The amount of information a speaker is trying to convey will have an impact on the quality of the pronunciation. When individual words are conveying less information, they may be reduced (or hypo-articulated) while words conveying more information may be hyper-articulated [54]. While a particular context (i.e., spontaneous speech) can be statically modeled by using a dictionary that matches the context, dynamic modeling will be used to change pronunciation possibilities over the course of a conversation. We develop modeling frameworks that take advantage of high-level conversational information such as prosody, syntax and discourse.

1.2 Main Contributions

This work contributes to the field of speech science in three aspects: an exploration of the high-level contextual factors that affect speech production, a recognition framework for locally-adapted pronunciation models, and a model of pronunciation based on phonetic

features. The specific contributions in these areas are described in this section.

The main contribution of this work is a better understanding of pronunciation variability and the factors associated with it. Previous work on pronunciation modeling looked at phonetic context, trigram context and speaking rate. This work assesses the usefulness of different types of structure, including syntax and discourse, as well as prosodic cues, for predicting variation. To this end, we evaluate automatic use of high-level information variables for prediction of pronunciation variants by decision trees. We introduce a two-stage approach to pronunciation prediction where intermediate variables characterize variation and can then be used to predict surface-form realizations of speech. This approach aims to more efficiently characterize the relationship between pronunciation variability and higher level information such as that seen in both hypothesized words and prosodic cues. Results on intermediated variable prediction show that text-based word variables such as part-of-speech categories are strong indicators of pronunciation variation and that acoustic correlates of prosody such as fundamental frequency (F0) and duration are also useful for predicting actual speech production given canonical dictionary pronunciations. The two stage approach to surface-form prediction shows great promise in oracle experiments, though gains in the fully automatic case are not statistically significant.

The primary applications of pronunciation models are in automatic recognition systems, and the incorporation of phone-based pronunciation modeling with recognition systems is evaluated here. We show that the use of pronunciation models can have a positive effect on recognition results when word-level pronunciation weights are trained on a large set of training data. Incorporating part-of-speech (POS) information in static pronunciation models appears to have no beneficial effect on recognition results, most likely because POS is not included in the language model and so the POS dependence is only modeled in the sense of a knowledge-based mixture. Dynamic pronunciation modeling can address this short-coming, as evidenced by improvements on some speakers, but requires further work on word-based representations. In the dynamic pronunciation modeling framework presented here, we see improvements as more higher-level information is incorporated. However, results with dynamic models still do not meet the static model baselines, in part because pronunciation weights for generated phone strings are only trained on 3.5 hours of data and

also because of implementation differences which lead to no word-level constraints in the dynamic model.

This work also includes an evaluation of the use of distinctive articulatory features (first introduced discussed by Jakobson, Fant and Halle in 1952 [41] and furthered by Chomsky and Halle in 1968 [8]) versus phones for pronunciation modeling, where articulatory features are symbolic linguistic characteristics that distinguish phones. Although phone-based systems are considered state-of-the-art, various types of articulatory features have been used in recognition systems in the past, in particular to address issues of environmental variability. In this work, twenty-one symbolic features (that can be on, off, unspecified or unused) are used to describe a phone. Using articulatory features rather than the phone set (on the order of 45 phones) reduces the number of parameters needed for modeling pronunciation changes. This is further motivated by the hypothesis that the 80 different "variations" of "and" could be characterized more efficiently than with a phone-based representation since there is only a small difference between /ae/ and /ax/ in the articulatory feature space. Articulatory features also allow a more detailed description of the sounds that may be produced in spoken language. For example, a phone that is reduced may not actually be a distinctly different phone but, instead, be a slightly different combination of articulatory features. We first use the articulatory features in a distance measure designed to quantify pronunciation variation, which is useful in phone string alignment and as an intermediate variable for predicting pronunciation variation. In pronunciation modeling experiments, this work shows that variation can be predicted using articulatory features more efficiently than in a phone-based model and that predicted feature vectors are closer to hand-labeled versions than predicted phones.

1.3 Dissertation Overview

This work begins with two background chapters. Chapter 2 examines traditional speech recognition systems as well as previous work incorporating pronunciation modeling in recognition. The context and motivation for improved pronunciation prediction is developed here. Acoustic modeling techniques are presented to show how pronunciation modeling is incor-

porated into ASR systems. Previous work in the field is described, showing the connections between lexicon development and different styles of pronunciation modeling.

Chapter 3 presents linguistic foundations of this research, beginning with a description of the types of high level information used in this work. The connections between the linguistic context of a word (represented by discourse, syntax and prosody) and pronunciation variation are presented, as well as a detailed description of the factors that are used in modeling variation. We present the motivation for the articulatory features used in lieu of phones in parts of this work, and the features and the values used to represent them are explained in this chapter.

Next, the corpus used and available results are described in Chapter 4. This work uses the Switchboard corpus because it is a set of spontaneous conversations, resulting in a great deal of pronunciation variability [30]. Additionally, this corpus has been used in the speech recognition community for 10 years, resulting in a large amount of supporting data such as part-of-speech labels, disfluency markers and prosodic data. Most importantly, a 4-hour portion of this data is phonetically hand-labeled giving us the "true" surface-form pronunciation rather than just the dictionary, or baseform, pronunciation [31, 32].

Chapter 5 describes an analysis of the connections between high-level linguistic information and pronunciation changes. In particular, we show connections between high-level information variables such as a window of part-of-speech categories and pronunciation variation. The connections are established automatically via the use of decision trees to find the attributes most useful for pronunciation prediction. This chapter also presents a two-stage model of pronunciation prediction. The first stage could be either (or both) of two intermediate predictors: a word-level pronunciation distance measure and a phone-level indicator of transformation (e.g., reduction vs. remaining the same). The second stage is the prediction of surface-form pronunciation given the dictionary baseform, the conversational attributes and possibly either of the intermediate predictors. The experimental results in this work are based on the hand-labeled portion of the Switchboard corpus.

Chapter 6 presents the combination of our phone-based pronunciation prediction with a recognition system. The pronunciation models used in this portion of the work use the best sets of attributes found in Chapter 5 and are trained on two different sets of data: the hand-

labeled portion of the corpus and an 80-hour portion of the acoustic model training data that has been re-aligned using the pronunciation model built with the hand-labeled data. We present results using the pronunciation models based on different sizes of training data as well as different sets of attributes (phone-level information only, and this combined with text-based information). Additionally, we present a framework for incorporating prosodic information in a dynamic modeling context.

Chapter 7 describes the work performed to examine the use of articulatory features versus phones for the prediction of pronunciation changes. Rather than prediction of phone transformations, decision trees are built for each articulatory feature using the best sets of attributes found in Chapter 5. We show that features efficiently model pronunciation variation in the sense of using fewer parameters while improving prediction results. This chapter also sets up possible frameworks for including the articulatory features used here in a speech recognition system.

This dissertation concludes in Chapter 8 with a summary of the main contributions and implications of this work, as well as suggestions for future directions for this work in the areas of analysis, prediction and recognition, particularly related to feature-based systems.

Chapter 2

SPEECH RECOGNITION

In this chapter, after a brief description of a basic recognition system, current approaches to pronunciation modeling will be described. This will present the context for the incorporation of higher level information and the connection between speech technology and linguistics that will be discussed in the next chapter.

2.1 Basic Recognition System

The major components of a speech recognition system are signal processing and pattern recognition as shown in Figure 2.1. Pattern recognition involves finding the highest scoring word string, W, given cepstral or other observation vectors $x(n)$, the output of signal processing. Most automatic speech recognition (ASR) systems are statistical and use both an acoustic model and a language model. This section will briefly describe signal processing then the basic models used in speech pattern recognition.

Signal processing transforms speech into observation vectors which are then used in training and recognition. Cepstral analysis, the most commonly used procedure, is the

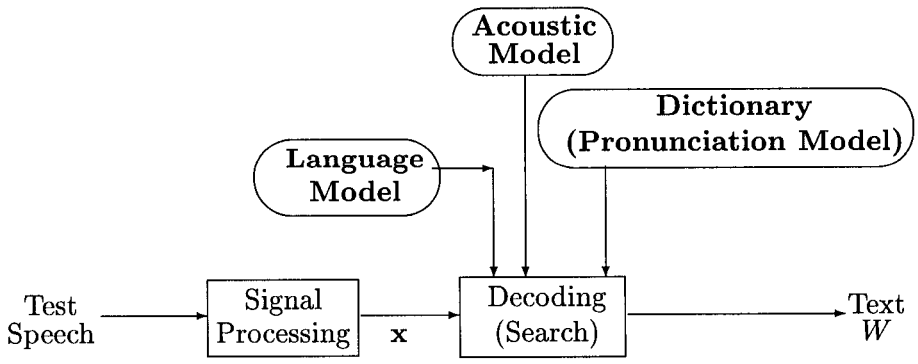


Figure 2.1: Basic speech recognition system with optional pronunciation model.

processing used to prepare speech for recognition. Assuming a linear model of speech production, cepstral analysis provides a method for separating the excitation source from the effect of the vocal tract. The vocal tract effects are the necessary aspect of speech for word recognition. One possibility for implementing cepstral analysis involves taking the Fourier transform of the observations, the log of the transformed observations, then the inverse Fourier transform, and finally the result is windowed to retain only the low order cepstra. Information about the temporal evolution of the cepstra is incorporated by including the derivatives (and often the second derivatives) of the cepstral coefficients.

In speech recognition, the time-varying patterns of a word are represented using **acoustic models** of speech. The models are estimated using a large amount of training data. It is important to have enough training data in order to model variability and to create robust models. For acoustic models in general, probabilistic models are used to represent $p(\mathbf{x}|W)$ where $\mathbf{x}$ is the sequence of acoustic observations representing the training speech and W is the associated word string. The most common acoustic unit used for modeling is a phone, i.e., words are composed of a sequence of phones. A phone is related to the linguistic notion of a phoneme. Phonemes are the smallest units of a given language that if exchanged, will change the meaning of a word. The words "pin" and "bin" are differentiated by the difference between the minimal pair, phonemes /p/ and /b/. The acoustic realization of a phoneme is a phone. Differences in phones will not necessarily result in different meanings but do result in acoustically different sounds. For example, the /t/ in "butter" can be made to sound like the /t/ in "cat" or it can be flapped, creating a sound more like a /d/. The meaning of the word does not change. Further discussion of the difference can be found in [10]. In this work, we will use phonemes to describe canonical sounds but will use phones when discussing work in acoustic modeling or labeled speech data. Triphones, or phones tagged with the phones immediately following and preceding them, are typically modeled in order to reduce the effects of coarticulation, where the acoustic realization of one phone depends on its neighbors. The expansion of the model set to triphones introduces a data sparsity problem that has generally been dealt with through clustering methods. Decision tree based triphone clustering with phonetic questions is an example of the successful inclusion of linguistic information into ASR systems. (See [68] for more on this topic.) The most

popular type of model is the Hidden Markov model (HMM) [3, 75], which is what we will use in this work. HMMs and decision trees are described in more detail in Section 2.2.

Language models represent the probability of word strings $p(W)$. For small vocabulary tasks, a finite state network can be used which maps out specific alternatives for the vocabulary. A typical language model used in large vocabulary tasks is an n-gram model where the probability of a word is found given the $n-1$ previous words. Training of the language model is done using the text of the training data to compute relative word frequencies that are smoothed in some way to avoid zero probability cases. The tri-gram language model probability for a word string W can be represented as

$$p(W) = p(w_1)p(w_2|w_1)\prod_{n=3}^{N} p(w_n|w_{n-1}, w_{n-2}).$$

Decoding involves finding the maximum likelihood word string given an observation sequence,

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p(W|\mathbf{x}) \tag{2.1}$$

or, using the acoustic and language models,

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, p(\mathbf{x}|W)p(W). \tag{2.2}$$

Various efficient search methods, such as the Viterbi algorithm, can be used in this task to search through a network of words (or a tree for a first pass of recognition), which themselves are networks of phonemes, to find the most probable sequence. One way to reduce the search space is to use a multi-pass system [85], where the first pass narrows the space using simple acoustic models and subsequent passes incorporate more detailed models (higher order and/or more parameters).

Traditionally, word pronunciations are specified by a static **dictionary,** which defines the sequence of subword units associated with the acoustic models that comprise each word in the vocabulary. When pronunciation probabilities are included, we simply change the definition of the probability of the acoustics given the words, to include the dependence of the acoustics on the pronunciations and the pronunciations on the words. In this case, the

$p(\mathbf{x}|W)$ can be modified to become:

$$p(\mathbf{x}|W) = \sum_{\phi} p(\phi|W)p(\mathbf{x}|\phi), \tag{2.3}$$

where ϕ is the phone string corresponding to the actual pronunciation and $p(\phi|W)$ is the **pronunciation model**. In this case, the equation used becomes:

$$\hat{W} = \max_{W} \sum_{\phi} p(\phi|W)p(\mathbf{x}|\phi)p(W) \approx \max_{W,\phi} p(\phi|W)p(\mathbf{x}|\phi)p(W). \tag{2.4}$$

The probability of a phone string can depend on either the entire word string W, taking in information connected with the utterance, or it may simply depend on the individual word, w_i. The different possible ways of designing and incorporating $p(\phi|W)$ will be described in the next section.

2.2 Models: HMMs and Decision Trees

This section describes the most commonly used acoustic model, the hidden Markov model (HMM), and decision trees as probabilistic models used in combination with HMMs for acoustic model clustering and as predictive models for pronunciation. All of these basic tools will be used in this thesis.

2.2.1 Hidden Markov Models

In order to model the connections inherent in the acoustics of speech, the modeling framework needs to allow for time-varying conditional dependencies associated with speech production. However, modeling long-term dependencies would require a great deal of training data, and the time to search through model possibilities is prohibitive. Modeling these dependencies is addressed by using hidden Markov models. Given that the vocal tract is in a particular position, or state, for producing a sound, we model both the probability of the acoustic output and the probability that the vocal tract will change from the current state to the next state. The probability that the state $S_t = i$ will change to state $S_{t+1} = j$ at the next time step is

$$\mathrm{a}_{ij} = P(S_{t+1} = j|S_t = i)$$

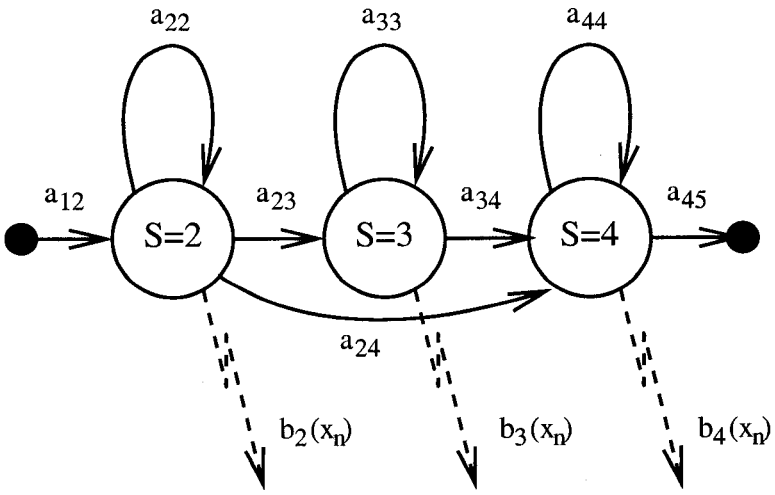


Figure 2.2: A state diagram of a 3-state hidden Markov model.

where $\sum_j a_{i,j} = 1 \forall i$. The state is not observed directly, hence the "hidden" portion of HMMs. The acoustic output or observation for a given state at a particular time, x_t, is described by the distribution

$$b_j(x_t) = p(X_t = x_t | S_t = j).$$

The modeling assumptions made are reasonable given a sufficiently large state space.

Phones change during their physical production. Sounds like /p/ and /b/ are created by building pressure behind the lips and releasing the air in a sudden burst. Vowels are more likely to stay the same throughout production, but their boundaries may be affected by the phones on either side. Acoustic models need to describe this temporal variation within the phones, so the model topology consists of states connected to each other in a left-to-right transition network. Figure 2.2 shows a sample model topology for a single phone model. This is a three state model with the transition probabilities, a_{ij}, and associated observation probability distributions, $b_j(x_t)$, labeled. (States 1 and 5 are used for joining phone models.) A typical distribution assumption is that $b_j(x)$ is a Gaussian mixture with diagonal covariances. Full covariance Gaussian distributions are sometimes also used.

In order to model the coarticulation affects of the vocal tract, we expand the set of models from phones into triphones based on the previous and following phones. This creates models

that better specify the acoustics. For example, the word "mat", initially described by the string of three models /m/ /ae/ /t/, now becomes /#-m-ae/ /m-ae-t/ /ae-t-#/. Triphones describe contextual differences but still only represent the center phone. On the other hand, there may not be enough instances of a particular triphone to train the associated model parameters. Or, the final states of the triphone /m-ae-t/ (describing the center phone /ae/) may be very similar to the final states of the triphone /n-ae-t/. Splitting up the data makes the models for those states less robust. This raises the question of how best to cluster the states of triphones for parameter sharing so that we gain the benefits of knowing the phone context but do not suffer from sparse training data. One way is to do agglomerative clustering based on acoustic similarity [118, 40]. This works well but does not allow modeling of triphones unseen in training. Decision tree based clustering is a method that allows for unseen triphones to be modeled and will be described next.

2.2.2 Decision Trees

Decision trees are commonly used for classification and regression problems [6]. Decision trees are designed to take a collection of prediction variables or "attributes" as input, and return an output variable. When the output variable is discrete, decision trees can be used to either predict in which class a token belongs or to give a conditional probability distribution of the different classes given the input variables. When the output variable is continuous, the trees predict a number or a conditional probability density given the inputs. Decision trees make their decisions by asking questions based on available attributes in the training set. See Figure 2.3 for an example tree.

In decision tree design, at each node in the tree, a question is asked and the training data is divided and sent down branches. If the question does not partition the data into homogeneous sets, more questions can be asked until a stopping criteria is met. The stopping criteria could be no further improvement in the overall classification as measured by error rate or change in entropy, or when the data in the node is homogeneous so there are no remaining questions to ask about the data. The trees are grown in a greedy fashion and optimize decisions at the current node only. Thus, tree growing does not necessarily give

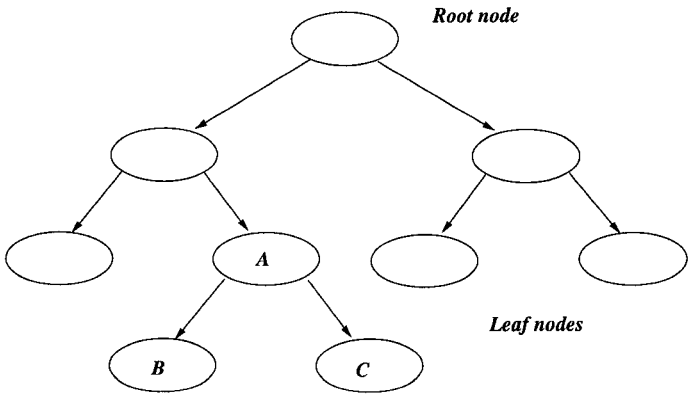


Figure 2.3: An example of a decision tree. The top node is the root of the tree and every terminal node is a leaf. Branching nodes are referred to as the parents of the resulting nodes (the children). Here, node A is the parent of nodes B and C, which are, in turn, siblings. At each node, questions are asked and the input travels down the appropriate branch. When a leaf node is reached, the result is a probability distribution, a category decision, or a number in the case of a regression tree.

a globally optimal partitioning in terms of the objective function. The objective of the growth can vary depending on the goals. Maximizing the likelihood of the training data is one example of an optimization criteria.

Decision trees are used in two different ways in this work: acoustic model clustering and pronunciation modeling. In acoustic modeling, triphone states are clustered together to reduce the effects of data sparsity. Decision trees are used to cluster observation densities associated with states of triphone models of a particular center phone by asking questions about the phone's neighbors, which are the prediction variables [115, 117]. Such questions might include phone class (nasal, stop, vowel, etc.) or about production of the neighbors (tongue placement, voicing, use of lips, etc.). In training, the output variable is a cepstral vector and its derivatives. The objective is to maximize the likelihood of the data, which is equivalent to minimizing the conditional entropy of the observation distributions. After design, the leaf node is associated with a Gaussian distribution, which may later be re-estimated as a Gaussian mixture. Shafran has explored the use of of questions about information other than triphone context (i.e., syllable, stress and prosody information) for acoustic model clustering [87]. The information used by Shafran is similar to that used for

pronunciation modeling in this work and we will take advantage of the existing clustering software for incorporating other linguistic information.

For pronunciation modeling, while we may ask some of the same questions, the objective of the trees is different. Given the expected phone from the canonical dictionary, the pronunciation model predicts a phone label that should result in a better match with the acoustic model. The output in this case is a phone label rather than an acoustic cluster. A discrete probability mass function is generated at the leaf nodes, so either the distribution or the discrete label can be used in recognition. When the output is a probability distribution, it can be incorporated into the probabilistic framework of the recognition search. Minimizing the posterior distribution entropy is typically used as the criterion for making decisions about splitting the training data in tree design. Decision trees are particularly useful for pronunciation modeling, because models can be built using both categorical and continuous attributes. This approach has been used successfully in [7] and [84]. Our approach for incorporating longer term information is to ask questions about utterance and word level context as well as the typical questions about phone based information.

2.3 Pronunciation Modeling

This section examines current approaches to pronunciation modeling, beginning with lexicon development, followed by comparisons of possible modeling decisions, and a discussion of prediction factors used in pronunciation models.

2.3.1 Lexicon Generation

The lexicon, or dictionary, is simply a list of the possible words that can be recognized, accompanied by the strings of phonemes that describe possible pronunciations of each word. The canonical pronunciation for a word, such as one would find in Webster's dictionary, is generally used. It is typically the pronunciation of the word spoken in isolation. A sample of a dictionary can be seen in Table 2.1. There are three issues in defining a basic lexicon: 1) which words are included, 2) the phone set, and 3) which pronunciations are included.

The first issue is usually defined by the task domain and computing constraints. The

Table 2.1: *Examples from the dictionary used in baseline experiments.*

Word	Pronunciation 1	Pronunciation 2
a	/ax/	/ey/
a's	/ey/ /z/	
aaron	/eh/ /r/ /ih/ /n/	
and	/ae/ /n/ /d/	/ax/ /n/ /d/
babies	/b/ /ey/ /b/ /iy/ /z/	
babies'	/b/ /ey/ /b/ /iy/ /z/	
baboon	/b/ /ae/ /b/ /uw/ /n/	/b/ /ax/ /b/ /uw/ /n/
baboons	/b/ /ae/ /b/ /uw/ /n/ /z/	/b/ /ax/ /b/ /uw/ /n/ /z/
coactive	/k/ /ow/ /ae/ /k/ /t/ /ih/ /v/	
coagulant	/k/ /ow/ /ax/ /g/ /y/ /uw/ /l/ /aa/ /n/ /t/	
coal	/k/ /ow/ /l/	

larger the size of the vocabulary, the larger the search space is during recognition and thus the larger the required computer resources, both time and memory. When dealing with an open vocabulary task, errors are introduced because some words will be out of vocabulary (OOV). On the other hand, including all possibilities can introduce errors coming from confusable words.

Phones are the sounds that are actually produced in speech. Since not all phones are used in every language, the choice of phone set can reduce the search space by excluding some sounds or merging extremely similar sounds. In American English, there are some sounds that are not distinguishable in most dialects. For example, the first vowel in the word "daughter" can be /aa/ as in "water" or a more rounded version /ao/, when produced by people native to the New England area. Since the majority of the time in databases covering the United States, the vowel is /aa/ and since /ao/ is similar to /aa/, /ao/ is a sound that may not be included in a given phone set. This highlights the difference between "baseform" phones, the phones from the lexicon, and "surface form" phones, the sounds actually produced during speech.

The third issue, that of which pronunciations to include for each word, involves the trade-offs between accurately describing possible pronunciations and increased confusability. Speech recognition systems generally work reasonably well describing most words with a single pronunciation. Hain showed that single pronunciation dictionaries with the single pronunciation chosen carefully give as good or better results as systems trained on a multiple pronunciation dictionary on Switchboard and Resource Management tasks [36, 114]. It is thus preferable to only have multiple pronunciations for the words where it will be most useful in correctly recognizing the words, and a single pronunciation for such words as "apple". Unfortunately, the words with the most pronunciation variability are shorter, high frequency words, in particular function words.[1] Multiple pronunciations of these words are often confusable with other short words that may or may not be function words. This has been examined more closely by Jurafsky *et al.* [44]. Their findings will be discussed further in Section 3.1.

[1]Function words are closed class words that do not necessarily have information content of their own but are important for grammatical reasons. They include such words as determiners, pronouns, prepositions.

Lexicons can be hand written or automatically generated. Hand written lexicons are generally based on the pronunciations in a traditional dictionary. Text-to-speech grapheme-to-phone systems are often used for automatic generation of new entries. Automatically generated dictionaries can also come from performing phone recognition, rather than word recognition, on a set of training data and then using the recognized phones to define the word pronunciations [80]. This method requires good acoustic models and, while describing some of the variability in speech production, may also introduce errors that come from misrecognized phones and incorrect word transcriptions. Holter and Svendsen generate lexicons jointly with the acoustic subword units (not necessarily phones) to optimize (in a maximum likelihood sense) the acoustic models and the lexicon [38, 39]. Automatically generated pronunciations can be a useful way to augment a hand written dictionary when multiple pronunciations are only added when they actually appear in training data a certain number of times [110]. This technique requires training tokens for each new pronunciation so it is not generalizable to unobserved vocabulary items.

Some systems include multiword sequences in the lexicon (e.g., [57]). For example, the phrase "you know" when used as a filler is often pronounced as "y'know". Additionally, word pairs such as "kind of", "should have", and "could have" are commonly used together and have distinct reduced pronunciations, often sounding like "kinda", "shoulda", and "coulda". While the concept of describing cross-word effects through the use of multiwords is appealing, Nock and Young showed that even including pronunciation weights in recognition experiments yielded very little change in results [66]. Thus, multiwords will not be used in this work. It is expected that cross-word boundary phenomena will be incorporated dynamically using word context in pronunciation prediction.

2.3.2 Frequent Cases vs. General Model

The issue here is whether the model requires observations, i.e., examples of a word with a particular pronunciation during training, or whether the model is generalizable to unseen possibilities. In a frequent case model, the probabilities for a particular pronunciation can be found by performing a phoneme alignment of a given speech signal and the associated

transcribed words using the static dictionary or by using hand-labeled phoneme transcriptions. By keeping a tally of how often a particular version of a word is used, a simple relative frequency distribution can be obtained, in which case, the sum of the probabilities for a given word would be one and would define $p(\phi|w_i)$. In these cases, examples of a pronunciation must occur in training data in order for it become a part of the model. While work to cull a wide range of pronunciations from training data has resulted in recognition improvement [2, 17, 57], the fact that it is not generalizable can cause problems when the recognition task vocabulary changes or when logical or even canonical pronunciations simply do not occur because of sparse training data.

There are two main ways to generalize models. One is to define rules about phone changes and the other is to predict phone changes explicitly. Rule based transformations have been implemented with the rules being hand specified with trained probabilities [11, 102] or automatically learned. Automatic learning can use either hand labeled data [57, 7] or recognized phones from training data [12, 28, 66]. Linguistic information has also been used to create rules that include higher level information [21, 20, 69]. Additionally, these rules can be augmented with probabilities to create a better model as in the work of Finke [21] or weights in the case of finite-state transducers [93]. Explicit phone prediction through phone confusion tables [84] and decision tree prediction [80] has also been used. These each require a large number of parameters to fully train and may be sensitive to a data sparsity problem as evidenced by poor performance when only a small hand-labeled subset is used to train pronunciation models [7].[2]

2.3.3 Static vs. Dynamic Modeling

In lexicons that have only a few alternate pronunciations, the probabilities of each are not necessarily represented, but when there are many alternatives a pronunciation model, $p(\phi|w_i)$, is needed to reduce confusability [79]. In static pronunciation modeling, the model does not change during the recognition process. Static modeling takes advantage of what

[2]It should be noted that there may be a mismatch between linguistically motivated assignment of phone labels by humans and the acoustic patterns learned automatically given a citation-form lexicon. This mismatch likely also contributes to the poor performance.

has been seen in the training data but does not allow the system to take advantage of what is currently happening in an utterance. In conversational speech, a speaker will not necessarily say the same word the same way all the time. The pronunciation can change depending on the context of the surrounding words, the overall message and the speaking style. With dynamic modeling, not all possible pronunciations are allowed at all times. This allows for more variability where it is pertinent while reducing the confusability problem that comes from having too many choices. The choices can be dependent on local word context, capturing phenomena like "gonna" without the added complexity of multiwords. Dynamical modeling can also include longer range effects associated with speaking style. The indicators of these effects will be discussed in Section 3.1. Fosler-Lussier has successfully implemented dynamic modeling and shown small improvements for spontaneous speech [25, 27]. Finke has also incorporated dynamic modeling with a feature based acoustic model [20]. This approach allows for portions of a phone to change rather than requiring an entire phone change in a pronunciation model.

While there are advantages to statically modeling pronunciation, such as reduced computation time during a recognition run, we will be focusing this work on dynamic modeling. The potential for reduced confusion that comes with dynamic modeling, in part because of the ability to take higher level context into account, make it a promising framework for pronunciation modeling.

2.3.4 Phone-based vs. State-based Modeling

Because not all pronunciation changes involve a phone transforming into an entirely different phone (as shown in [83]), state-based pronunciation modeling is currently being explored by several researchers. For example, a reduced vowel may not have the tongue fully in position for the full vowel but other aspects of the vocal tract will match the full vowel. By the end of production, the tongue may reach the position for the full vowel. In state-based modeling, the pronunciation model predicts changes at the level of different acoustic model states rather than phones. A model that changes only selected states of the acoustic model may better model the pronunciation change without requiring separate phone models for

all production variants. One example of this is the sharing of state-level Gaussian mixture distributions across acoustic models presented by Saraclar, Nock and Khudanpur [84]. A similar approach has been used for recognition of speech of non-native speakers [113]. Nock and Young present loosely coupled HMMs that allow asynchronous traversal of state-level model topology, which better represents the variation associated with coarticulation in the sense that asynchrony is observed in state-level alignments [65, 67]. Hain and Woodland present hidden sequence modeling which allows more than one model to be used for a phone in a particular context allowing the best model to be chosen locally rather than globally [37]. Eide automatically learns the HMM topology for context-dependent phones, allowing for substitutions and reductions by modeling sub-phonemic units as a single state [17].

While the state-based approach shows promise, we will restrict this work to the phone-based approach in the experiments reported here and describe how it can be extended to state-based modeling in Chapter 8.

2.3.5 Factors for Prediction

There are many factors that have been used in pronunciation prediction. For static modeling, information that can be encoded in the dictionary is available for use. This includes the phonetic context, where the identity of neighboring phones in a range of one to five phones on either side of the current phone are used, whether the phone is in a syllable with primary or secondary stress as labeled in the dictionary (or is not stressed at all), position of the phone both in the word and in the syllable (beginning, middle, end or only phone), and information based on individual word identity (e.g., part-of-speech tags).

For dynamic modeling, attributes that take into account information generated by the recognizer and hypothesis-independent acoustic cues have been used. This includes things like trigram probabilities of the words as given by the language model, word context (for example, by part-of-speech category), and speaking rate [21, 27].

Chapter 3

LINGUISTIC FOUNDATIONS

This chapter introduces the high level information such as prosody, syntax and discourse that are included in our pronunciation models, and also describes the linguistic representations of pronunciation variation and implications for computational models of pronunciation.

3.1 High Level Information

In Lindblom's theory of variation in speech [54], speakers adjust their articulatory effort to accommodate the listener and the importance of the information. Phonemes are hyper-articulated during points of emphasis or clarification and reduced at very predictable points. This can be seen in the reduction of such words as "to" and "of" to the point where it is difficult to associate a measurable segment duration with them. To some extent, reduction can be captured by word predictability as quantified by local n-gram language model scores, which analyses show to be a useful predictor of pronunciation variability [44, 26]. In addition, we hypothesize that this variation in articulation quality can be associated with the related phenomena of information structure, including syntax and discourse structure. This structure can be exploited by talkers in estimating the needs of their listeners, i.e., whether hyper-articulation is necessary or whether less careful articulation will do. More broadly, factors that have been shown to improve language models may be candidates for predicting pronunciation variation.

The variation in word pronunciation can be related to two different, although not independent, aspects of speech: 1) information structure as described by syntax and discourse, and 2) prosody. This section will first discuss information structure, followed by a section on prosodic events and their acoustic correlates. Throughout, examples will be given to connect this information with pronunciation variability.

3.1.1 Linguistic Structure

Syntax describes how words are put together to form phrases, clauses, or sentences. The particular words in a sentence can be described in terms of their part-of-speech (POS), e.g. noun, verb, adjective, etc. Syntax includes the bracketing of word strings into noun and verb phrases. The syntactic structure may have the effect of making the current word so obvious that it is barely necessary to speak the word. This can be seen in the reduction of such words as "to" and "of" to a single phone or even no measurable segment duration [107]. The same word can be quite reduced as a grammatical preposition but emphasized (and hence clearly articulated) as a directional preposition [73]. Words within the same noun or verb group may have a stronger effect on each other. Part-of-speech has been used successfully by Wakita *et al.* [108] to predict pronunciation changes in Japanese. However, they did not go beyond the POS of the particular word being recognized. The surrounding POS context has been used to improve language modeling and is a reasonable extension to the use of POS in pronunciation modeling. Neighboring POS is a low cost approximation to syntax and is easy to obtain automatically. For example, if "going to" is followed by a verb rather than a determiner, it may be reduced to "gonna". People often say "I'm gonna run to the store" but they never say "I'm gonna the house."

Jurafsky *et al.* [44] were able to predict reduction of vowels in function words using such things as word frequency and predictability from context. Additionally, word frequency has been shown to be useful in pronunciation modeling [84], but word frequency is correlated with content word/function word distinctions. The content/function distinction, which is related to POS, may (or may not) be more directly related to pronunciation.

Discourse describes how a conversation flows between one or more speakers. Grosz and Sidner [35] divide discourse into three components: 1) segmentation, which represents topic structure as a hierarchy; 2) intention, which describes what the speaker means to communicate; and 3) attention, or information focus. Segmentation divides a conversation into topics and sub-topics, as well as speaker turns and utterances. Topic segmentations for the Switchboard corpus are not available and would be costly to obtain by hand, so this work uses only speaker turns and utterances.

Intention is sometimes "coded" in terms of dialog acts. Dialog act labels describe the intention of the utterance. A general dialog act mark-up structure was devised by Allen and Core [1] and modified by Jurafsky *et al.* [45] for use with the Switchboard corpus of conversational speech. Examples of dialog acts can be seen in Figure 3.1. Work in automatic labeling of dialog structure has shown that both word content and prosody are useful for identifying acts [42]. Further, dialog acts have been shown to improve language modeling results [42, 56]. Because dialog act labels contain a great deal of information about what words utterances contain, they may also have an effect on how those words are pronounced. For example, if the speaker is simply murmuring a "yeah" as a backchannel response, it will have a more reduced vowel than when the same speaker is exclaiming an enthusiastic "yeah" or even an affirmative, but unemotional "yeah" in response to a question. These cases are in different contexts, specifically different dialog acts. A speaker who is trying to convey a new idea as in an opinion or statement is more likely to clearly and canonically pronounce a word. Lastly, the concept of "given" vs. "new" in a sentence, which is related to information focus or attention [74], can be used in either language modeling or pronunciation modeling. New information may be more likely to be canonically pronounced, or even hyper-articulated [54].

Meteer [43] has developed an automatic algorithm for detecting the pivot point in utterances. The pivot point is dependent on syntax and is basically the point after the main verb. Even though it is based on syntax, pivot information has been used to get at discourse information. The words before a pivot are understood to be given while following words are new. This point can be used to adjust language model probabilities or to detect and predict pronunciation changes. Pivot information was used in [42] for language modeling with some success.

In order to use syntax and discourse information, we need to have associated labels for the available pronunciation data. Both part-of-speech tags and pivot points can be automatically labeled and are thus easily available for training sets. In the context of speech recognition, given an N-best list of hypotheses, it is easy to get these labels for a second pass of recognition. Dialog acts have been hand labeled for a significant portion of the Switchboard corpus, as well as several test sets [42]. Lobacheva has developed an

Dialog Act	Utterance
Wh-Question	What kind do you have now?
Statement	*Uh, we have a, a Mazda nine twenty nine and a Ford Crown Victoria and a little two seater CRX.*
Acknowledge-Answer	Oh, okay.
Opinion	*Uh, it's rather difficult to, to project what kind of, uh, -*
	...
Turn-Exit	*So, uh, -*
Yes-No-Quest	And did you find that you like the foreign cars better than the domestic?
Answer-Yes	*Uh, yeah,*
	...
Statement	*And some of the fish were supposedly making a comeback.*
Backchannel	Uh-huh.

Figure 3.1: Switchboard conversation fragments labeled with dialog acts.

algorithm for automatically labeling sentences [56] which could be used in this work to label a larger portion of the training data if dialog acts prove to have a significant role.

3.1.2 Prosody

The word prosody is related to the Greek *prosōidia*, a song sung to instrumental music, and the Latin *prosodia*, accent of a syllable. Prosody is how a sequence of words is said, or poetically speaking, it is how speech is sung. Prosody is typically described in terms of symbolic events and acoustic correlates of these events. **Symbolic prosodic events** are associated with prosodic phrases and phrasal prominences, each with intonational markers. Prosodic phrases give structure to speech by grouping words and are related to syntactic bracketing, in that syntax affects the location of phrase boundaries, but reflect a flatter structure. Prosodic phrases are associated with intonational markers at the end of the phrase, often called boundary tones, that distinguish statements from questions or may indicate continuation. Prosodic prominence corresponds with an emphasized syllable, sometimes referred to as phrasal stress, and (in most symbolic labeling systems) is associated with a pitch accent marker. One widely used method for coding prosody is the TOBI labeling system [94], while contour-based models are described in [50] (for the British school of intonation) and (for the Dutch model of intonation) in [106]. Other intonation systems use a more continuous representation (e.g., Taylor's tilt model [104, 105]) but most retain some distinction between phrase and prominence markers. Useful overviews of prosody are presented by Cutler *et al.* [13] and Shattuck-Hufnagel and Turk [89].

It has long been clear that prosody is important in speech synthesis, as prosody is an important part of making speech sound natural. Because a great deal of information from the speech signal is included in prosody, it makes sense to take advantage of this in ASR as well. However, the symbolic events described in the previous paragraph are costly to hand label and are not widely available for speech corpora. **Acoustic correlates** of prosody are straightforward to extract from speech waveforms. The acoustic correlates of prosody include fundamental frequency (both local pitch movements as well as the pitch range over a phrase or utterance), duration at the segmental level and phrase level speaking rate, energy

of the speech, and articulation quality. Work has been done using acoustic correlates of prosody for various speech analysis tasks with success; in particular work using acoustic correlates in decision trees supports the approach taken in this work [91, 70].

It is known that fast and slow talkers are difficult to recognize [71]. Fosler-Lussier *et al.* have done a great deal of work developing automatic methods for measuring speaking rate and have shown that speaking rate in conjunction with dynamic pronunciation modeling can lead to improvements in recognition accuracy [26, 27, 63]. Additionally, Jurafsky *et al.* found that rate of speech and segmental context (following word beginning with a vowel or consonant) are correlated with reduction phenomena [44].

Fundamental frequency (F0), signal-to-noise ratio (SNR) and duration have been useful for dialog act and disfluency detection so may be useful for pronunciation modeling [42, 90, 100]. The connection between information structure and dialog acts is manifested through prosody as well as word choice. For example, intention affects tone: a question will often end in a rising tone. This can then be seen in changes in the fundamental frequency. F0 was not useful in the studies representing a hidden pronunciation mode, but that may be a normalization or measurement issue [69]. At issue when working with spontaneous conversational speech is the fact that crosstalk affects intonation and makes it less reliable. While cleaning would remove these effects, recent corpora which use echo-cancelling to deal with crosstalk make this less necessary in the long run. Another issue is that of glottalization, the effects of which can be countered by smoothing the F0 values. In addition, F0 cues to prominence must be normalized to account for differences in local pitch range and for speaker variability.

Jurafsky *et al.* [44] also found that planning problems as represented by repetitions, pauses and filled pauses ("um" and "uh") are strongly correlated with reduction of vowels and function words. Words coming after disfluencies are often words that have a lower language model probability and thus might fall under the context of being new information, contrastive with the information to be corrected. These words may be pronounced more carefully than words used to describe previously given information. Acoustic correlates of prosody have been used to identify disfluent regions in speech (i.e., repetitions, filled pauses) and sentence boundaries in speech [90, 100]. These results together suggest that prosodic

cues may be useful for modeling pronunciation variation.

We hypothesize that, while pronunciation is affected by many factors, it may depend more directly on prosody than word sequence characteristics. However, it should be noted that there are conflicting cues in prosody. A high value of F0 suggests emphasis while low F0 values can suggest either emphasis or mumbling. Wightman and Ostendorf [112] show that duration reflects both speaking rate as well as stress and phrase boundaries. Greenberg *et al.* have also presented work showing strong connections between stress and duration and pronunciation [34, 33] and stress information has been used successfully for clustering in acoustic modeling [86, 87]. Short durations can suggest either a fast speaking rate or reduction while long durations can suggest emphasis, phrase final lengthening or simply a slow speaking rate. For example, with phrase final lengthening, deleted phones are still possible. Any one cue is not reliable. We need to examine both local and longer term measures, as well as the interaction between F0, duration and energy in order to disambiguate their causes.

3.2 Distinctive Articulatory Features

The term "features" has been used to mean different things in the speech literature including continuous-valued articulatory parameters and acoustic measurements used in recognition. We will use it here to describe symbolic linguistic characteristics that distinguish a specific sound or phoneme, i.e., distinctive features. Groups of phonemes may share features but, by definition, no two will have the same combination of features. Additionally, not all possible combinations of features are used in the set of phonemes defined for American English. This allows some flexibility in the production of a phoneme. If most, but not all, of a phoneme's features are present, the sound is still likely to be identified as that phoneme.

The features will be closely related to those described by Stevens in [98]. The set of features that are "+" or "–" will completely define a sound and make it distinguishable from all other sounds. These features can be modified depending on the surrounding speech context. Features are sometimes organized in a hierarchy as seen in Figure 3.2. The features towards the top of the tree are more important for distinguishing between large classes of

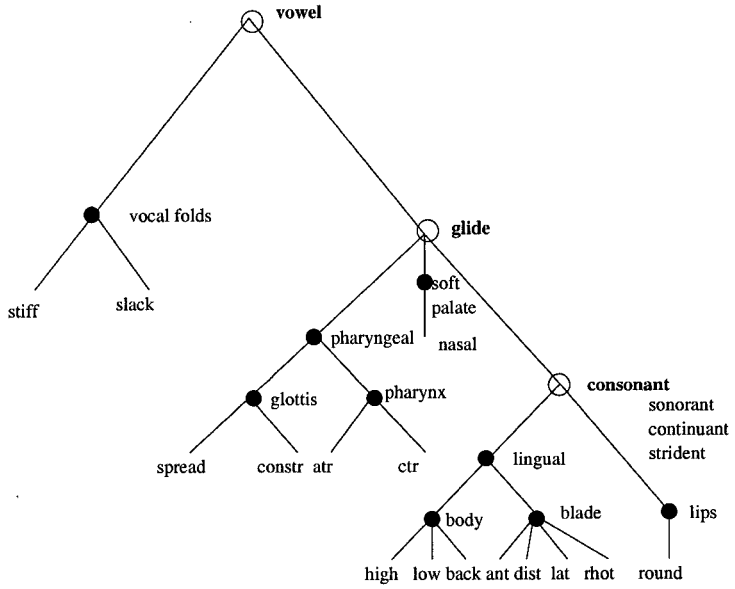


Figure 3.2: Hierarchical feature tree: open circles indicate articulator-free feature nodes, closed circles represent articulators (from [98]). This hierarchy does not necessarily represent dependencies between features.

sounds. The three kinds of features used in this labeling scheme are articulator-free features, articulator features, and articulator-bound features. At the highest level are the *articulator-free*, or 'manner', features that indicate whether the sound is a vowel, glide, consonant or syllabic consonant. The scheme further divides consonants into types (liquids, fricatives, stops, affricates) using binary categories of sonorant, continuant and strident. Lower in the tree are the *articulator features* that indicate which of the seven articulators is used in making a particular sound: the vocal folds, glottis, pharynx, soft palate, tongue body, tongue blade and the lips. The tongue and lips constitute the primary articulators for consonants.

The remaining articulators are secondary and are used to distinguish between different vowels and glides. The secondary articulators are not always explicitly labeled as the evidence of their use is seen in the third category, the *articulator-bound features*, which show how specific articulators are used to produce sounds and are seen as the leaves of the tree in Figure 3.2. "Stiff" and "slack" vocal folds are used to describe consonant voicing. A

"spread" glottis is open, while a "constricted" glottis is pressed together. This feature is only used in English for the /h/ sound (spread) and for glottal stops (constricted). Pharynx position is described with the terms "advanced tongue root", such as with non-back, tense vowels, and "constricted tongue root", such as with back vowels. The soft-palate articulator is associated with the "nasal" feature. Tongue body position in the mouth is described as being either "high", "low", or "back". The features associated with the tongue blade describe the position, shape and relative size of the constriction made by the tongue with the oral cavity. "Anterior" describes contact with the alveolar ridge while "distributed" shows how much of the tongue blade is involved ("+" denotes a broad portion). "Lateral" describes a tongue blade configured so that air can flow around the description and is canonically seen with /l/. If the tongue is rounded, as for /r/ sounds, it is described as "rhotic." The position of the lips is described as +/– "round", being + for /ow/ and – for /v/. Finally, there is a "delayed release" feature (not on the tree), which allows affricates to behave like both stops and fricatives. In total, we use 22 features in the distance, with the spread and constricted glottis merged into a single category.

These features are combined to create sounds as seen in Appendix A.2. Note that the diphthongs, as in the case of /aw/, are represented by two different feature sets over the course of production, which has implications for mapping features and feature changes in speech production to phoneme segments labeled simply as /aw/. Additionally, not all features are used for each phoneme. Some of the features that are unspecified can take on values depending on the surrounding speech context (e.g. vowels can become nasalized, consonants can be rounded); others are not defined in combination with specific features (e.g. vowels cannot be + or – strident). The coding system used is "+" for on, "–" for off, "X" for meaningless and "0" for unspecified but changeable. This coding system (i.e., including "X" and "0") is our own, but is somewhat similar to the work of Lahiri [52] in that while features are matched or mismatched, some features are allowed to be no-mismatch for particular phones.

Because the features described here correspond to the physical process of speech production, some instances of production may cause features to change in a systematic way. Since phonemes share features, these systematic changes may also occur across different

phonemes. Feature changes may be more general than phone changes in pronunciation variation. Examples of these feature changes include: the nasalization of vowels when followed by a nasal consonant, and the reduction of a vowel, as seen in the pharynx features "atr" and "ctr" when a syllable is not stressed. When a phoneme is "deleted" during the pronunciation of a word, such as the first vowel in the word "support" (pronounced more like "s'port"), features connected to that phoneme may still be present. In this example, the /p/ is aspirated as it would be at the onset of the second syllable and as it would not be as part of an /sp/ sound at the onset of "sport". The /s/ is also different as the vocal folds move toward the position necessary for /ax/ [58, 88]. This work will focus on the use of articulatory features to model pronunciation rather than modeling whole phone changes in order to reduce the number of parameters used in modeling. We hypothesize that modeling feature changes across multiple phones will reduce the effects of data sparsity. In Chapter 8, we will also describe extensions to this framework for modeling changes at a finer temporal scale.

The prior use of articulatory "features" in speech recognition appears on two fronts: symbolic and acoustic. Using features that symbolically map the human articulation system (whether distinctive or multi-valued articulatory) can be better suited than phones for multi-lingual ASR. While the phone sets of different languages vary significantly, they are still built with the same set of articulatory features. Building pronunciation and acoustic models using features may be better suited for recognition of languages or speech in domains with little available training data, as argued by Deng in [14]. When using features at the symbolic level, Deng uses rule-based systems to define multiple values for the articulators [15, 16]. The work here uses a distinctive (vs. articulatory) definition of features but another difference from Deng's work is the use of automatic methods rather than hand-written rules for characterizing feature changes. Other symbolic work with distinctive features has been done by Lahiri and Reetz [52, 77].

Prior work using acoustic manifestations of articulatory features involves detecting and using feature probabilities as observations in ASR. Eide *et al.* presented an articulatory representation for use in phoneme identification and word-spotting tasks in [19]. Eide has extended the work for use in ASR by combining traditional cepstra with a distinctive feature

representation generated in a first pass by a discriminative feature model [18]. Multi-valued articulatory observations have been used as the observation vector for the acoustic model. Kirchhoff showed that they were useful for creating robust models in noisy environments [47] and in language identification [48]. Stüker *et al.* have used binary representations of symbolic features based on acoustic observations for multilingual speech recognition [101]. Additionally, King *et al.* have shown that phonetic acoustic features can be successfully used for recognition by combining a neural network for detection of the features with an HMM decoding of the feature-based output into phonemes [46]. While these results differ from the use of symbolic features in pronunciation modeling, they can be seen to support our use in this work.

Chapter 4

THE SWITCHBOARD CORPUS

The Switchboard corpus is a collection of topic-directed telephone conversations between strangers. It has been widely used in the speech recognition community so a great deal of information and baseline results are available. Detailed descriptions of the collection methods and the corpus can be found in [30]. There were two separate collection phases for Switchboard. Phase 1 included data from many native speakers of American English. Phase 2 marked a concerted effort to collect data from specific regions of the United States and used echo-cancelling for a cleaner signal. The spontaneous nature of the speech in the corpus results in a great deal of pronunciation variability so the corpus is an appropriate one to use for this work. Additionally, with this corpus, work at JHU and SRI has shown that speaking style has a great impact on pronunciation and recognition results [84, 111].

While the focus of this work is on the Switchboard corpus, a smaller corpus, Callhome English, is also used in this work [53]. It was collected in the same manner as Switchboard but there is significantly less data than Switchboard. The conversations differ in two key ways: 1) the conversation partners know each other and 2) the conversations are not constrained to a particular topic. Past results comparing Switchboard and Callhome show that Callhome is a more difficult recognition task. See Table 4.1 (results marked II indicate that the test data is from Phase 2 of the Switchboard corpus). We conjecture that greater pronunciation variability is a key factor, since people can speak more casually when talking to someone they know, resulting in more reduction and deletion of baseform phones.

The results reported in Table 4.1 were the state of the art for their respective years.[1] The systems that were used for the evaluations included many different types of adaptation as well as multiple passes of recognition before declaring a final result. Adaptation methods

[1] Performance typically improves each year, but because the test sets are different each year the performance can degrade because of statistical variation.

Table 4.1: *Best reported results for Switchboard and Callhome annual evaluation sets [23, 59, 24]. Ranges of reported results span up to 20%. Results marked II indicate that the test data is from Phase 2 of the Switchboard corpus.*

Year	Switchboard	Callhome
1995	48.0%	NA
1996	38.8%	53.3%
1997	35.1% (II)	53.7%
1998	36.7% (II)	40.9%
2000	19.3%	31.4%
2001	19.8%/24.5% (II)	NA
2002	19.8%/24.3% (II)	NA

include such things as noise and channel adaptation, speaker adaptation of the acoustic models and vocal tract length normalization for the individual speakers. These adaptation methods typically require multiple passes to implement. Discriminative training [9], as opposed to maximum likelihood methods, for the acoustic models is another method that state-of-the-art systems may use. Also, Rover-style system combination is used [22]. Most recent developments in ASR systems yield small improvements, but gains from the advances tend to be additive with other advances. In particular, improvements in the first pass make unsupervised adaptation more successful. It should be noted that results without multi-pass scoring and adaptation are generally much worse. For example, the 19.8% result for the 2001 Switchboard test set started out as a 31.7% error rate for the first pass system with no adaptation [114]. The work reported here includes vocal tract length normalization but will not include any other model adaptation. We do this so that we can constrain the scope of this work and to maintain a fast turn-around time for experiments. A consequence of this choice is that the resulting performance gains may not be reflective of gains after adaptation. While adaptation may compensate for inter-speaker pronunciation variability, our focus is on intra-speaker (dynamic) variability, so we expect the methods to be complementary. It is also possible that pronunciation modeling gains are greater with more tightly tuned acoustic

models, i.e., that the problem of increased word confusability is lessened with pronunciation models used in conjunction with adapted acoustic models.

The corpus has been divided into training and test sets. The training set for acoustic modeling is expanded to include the Callhome English corpus as well. About 200 hours (excluding silences, noise and laughter) of speech are used for acoustic model training with about 12 of these being Callhome English. There are almost 400 speakers represented in this set. An 80 hour Switchboard training subset is used for training pronunciation models. This subset is being used because there are many different types of labeling associated with it, including syntax, disfluencies, parses and dialog act labels.

The development test set used in this work is the Dev '98 test set and is about 1.5 hours of speech from 14 conversations (about 14,500 words) between speakers not in the training set. The final evaluations are performed on Eval '00. There are a total of 40 conversations in this test set (about 27,500 words). Both test sets include conversations from Switchboard and Callhome English, making them more difficult than Switchboard only test sets. Our baselines for these test sets are in Table 4.2. These results include no adaptation except for vocal tract length normalization. The acoustic models are 3 state HMMs with diagonal covariance Gaussian mixtures. Gaussian distributions are shared by models in subtrees of the overall decision tree. Decision tree clustering included questions about syllable information [87]. The lattices generated from the baseline first pass of recognition will be used as the basis for second pass recognition for all the remaining experiments in this work. The lattices have oracle word error rates of 10.5% for Dev '98 and 9.9% for Eval '00. While these results are a little below state-of-the-art, they are reasonably close given the 12% difference noted in Table 4.1.

One particular aspect of the Switchboard corpus makes it most useful for the purposes of pronunciation modeling: a four hour portion of the training set has been phonetically hand-transcribed by Greenberg *et al.* [31, 32] at ICSI. The ICSI set will be used to initialize our pronunciation models and assess the role of different factors on pronunciation, with half an hour of data set aside for evaluation purposes. In this work, it is referred to as the "Held Out" set. The hand-transcribed corpus includes syllable alignments as well as phone labels. This set will be used exclusively for the analysis described in Chapter 5 and will be

Table 4.2: *Baseline word error results with no adaptation on the two test sets used in this work.*

Test Set	Portion	Baseline Result
Dev '98	Full Set	48.4%
	SWBD	41.9%
	Callhome	54.7%
Eval '00	Full Set	38.8%
	SWBD	33.8%
	Callhome	43.8%

combined with the rest of Switchboard and Callhome for the remaining chapters.

Of the 888 unique words in the Held Out set, 346 are seen less than three times in the ICSI training set and 211 are not seen at all. There is obviously a need for a generalizable pronunciation modeling approach. It is possible to generate optional pronunciation strings with decision trees trained on the ICSI training set for these words. For example, the word "above" is not in the training set but for use in recognition, six reasonable (for conversational speech) pronunciation strings are generated:

- /ax/ /b / /ah/ /v/ (canonical)
- /ax/ /b / /ah/
- /ih/ /b / /ah/ /v/
- /b / /ah/ /v/
- /ih/ /b / /ah/
- /b / /ah/.

Further pruning based on the likelihoods of pronunciation strings may mean that some of the generated phone strings are excluded from the dictionary. Because we have no fronted schwa in our recognition phone set, the acoustic model for /ih/ is used. Hence, /ih/ appears frequently as an option in our pronunciation model.

Chapter 5

FACTORS AFFECTING PRONUNCIATION VARIABILITY

This chapter describes the analysis of the high-level factors described in Chapter 3 and their connections to pronunciation variation. It begins with a description of the methods used to evaluate pronunciation variation, in particular the phonetic distance measure used to evaluate word-level pronunciation quality and to map between baseform and surface form phones. The questions addressed in this chapter are as follows.

1. What syntactic and discourse factors correlate with pronunciation variability? What prosodic factors?
2. Do local variables have substantially more impact than non-local ones (e.g., dialog act vs. neighboring POS for text, local vs. utterance measures of energy or F0 for prosody)?
3. Are certain factors more correlated with insertions than with substitutions and deletions?
4. Are certain factors more useful for predicting word-level variation or for predicting phone-level changes?
5. Is an intermediate variable (e.g., a reduction indicator) useful for predicting surface form realizations?
6. Are variables reflecting syntax and discourse correlated with prosodic cues to the point of not adding any additional information or are they both useful in combination?

Overall, this chapter will present the search for the combination of factors most useful for predicting pronunciation variation. Before concluding, significance testing results for key experiments will be presented. The best overall predictive models will be incorporated

into the recognition system. The options for incorporating the predictive models and the recognition results will be described in Chapter 6.

5.1 Analysis and Evaluation Methods

The goal of this work is to predict the change of a dictionary (baseform) phone to the realized surface form phone. To this end, we will predict measures of change that can either be used to predict pronunciation "quality" or as intermediate predictors in phone-to-phone prediction. By "quality" we mean closeness of the surface form phone to the baseform phone as defined by the dictionary. Pronunciations that are very different from the dictionary, particularly reduced pronunciations, can be said to be of lower quality. However, because the word is presumably still understood by the listener, we do not categorize this as a wrong pronunciation.

While an analysis of these predictor variables will be useful for understanding the style and quality of the speech, they are also potentially useful as an intermediate variable for prediction of phone-to-phone changes. Intermediate predictors allow for reduced dimensionality of information used in the decision tree pronunciation models. The goal for the intermediate factors is to predict a low-dimensional pronunciation variation factor based on high-level attributes[1] that can be used in combination with local phonetic and lexical stress context for surface form pronunciation prediction. Two different intermediate variables are considered: a word-level phonetic distance measure and a categorical indicator of the types of changes that can occur at the phone level. The types of changes and their predictors will be described below. The use of intermediate predictors can be thought of as the first stage of a two stage prediction algorithm as shown in Figure 5.1.

5.1.1 Phonetic Distance Measure

A phone distance matrix derived from articulatory features was developed to measure the distance between phones. We use the distance measure to align the baseform pronunciations of the transcribed words with the hand-transcribed surface form and to provide a single,

[1]We use attribute to mean information variables that will be used in a decision tree.

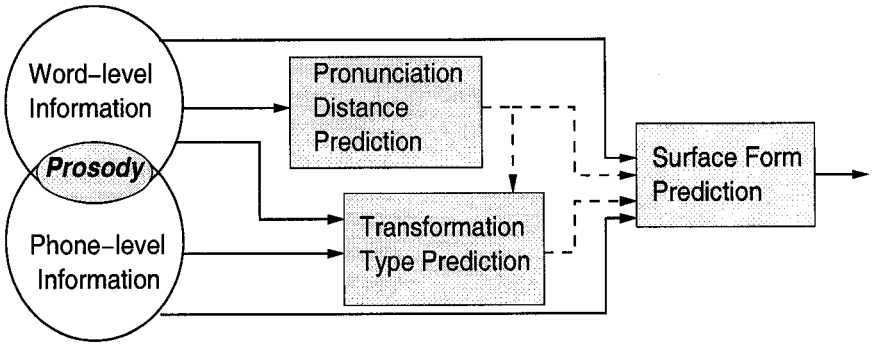


Figure 5.1: Block diagram for two stage pronunciation prediction.

word-level measure of pronunciation "quality".

A forced alignment between dictionary baseform pronunciations and the ICSI surface form phone labels was done using the finite state transducer tools developed at AT&T [62] with costs based on the phone distance matrix. This alignment gives both the actual phone strings used in the pronunciation of words as well as deletion, insertion and substitution statistics about particular phones and will be used for training our pronunciation models.

While the phone distance was first used as the cost matrix for alignments, it is also useful as a measure of how different a given pronunciation is from the baseform. The development of the distance measure will be described followed by the different ways it is used in this work. We use the term "features" to describe symbolic linguistic characteristics that distinguish a specific sound or phoneme. Groups of phonemes may share features (as described in Section 3.2) but, by definition, no two will have the same combination of features. Additionally, not all possible combinations of features are used in the set of phonemes defined for American English. This allows some flexibility in the production of a phoneme. If most, but not all, of a phoneme's features are present, the sound is still likely to be identified as that phoneme.

Following [98], the features are organized in a hierarchy (as in Figure 3.2 in Section 3.2), where those nearer the top of the tree are more important for distinguishing between larger classes of sounds. The distance measure is a sum of the distance between feature values (seen in Appendix A.2). A "+" feature has value 1; a "–" feature has value -1; an allowable

but unspecified feature has value 0; and an undefined feature ("X") has value -2. While many different combinations of weights were tried, with the goal of insuring that the biggest distances are between the phones in different manner classes, the best alignment results came when features were fully specified (including 0 and X) rather than combinations where some features were excluded or given different weights. The cost of substituting one phone for another ranges from 2 (for a single low-level feature difference, such as voicing in "p" vs. "b") to 72 for some vowel-consonant differences (e.g. "aa" vs. "t").[2] Deletion and insertion costs are proportional to the maximum substitution cost for a particular phone. Deletions are 0.6 times the maximum substitution value for the phone, and insertions are 0.85 times the maximum. Deletion costs for an unstressed vowel are further reduced (multiplied by 0.95). For consonants in syllable codas, substitution and deletion costs are also reduced (deletions multiplied by 0.9 and substitutions by 0.95). A grayscale plot of the distances (or costs) can be seen in Appendix A.3 and the distribution statistics for individual phones can be seen in Appendix A.4.

After performing an alignment between baseform and surface forms, each word has a total distance. This distance, which we use to describe pronunciation quality, is divided by the number of baseform phones in the word to give a normalized phone distance for the word. The average total cost (or pronunciation distance) of a word pronounced in the ICSI set is about 20, which is roughly 6.5 when normalized by the number of phones in a word. Excluding deletions, the average per phone cost is 2, so it appears that much of the distance is due to reduction phenomena (not surprisingly). Slightly less than half of the word tokens have a surface form that is identical to the baseform (zero cost). For the remainder of this work, the phrase *word distance* will be used to refer to the normalized pronunciation distance for a word. Sample alignments and word distances can be seen in Table 5.1.

There are several differences between these alignments and those reported in [7]. First of all, we allowed more than one pronunciation option in the alignment and recognition dictionary as baseforms. This was done in order to get the closest match possible between the surface form and an acceptable baseform and was seen particularly in words like "and,"

[2]Costs are scaled by a factor of 2 so that differences are retained for integer versions of the overall cost matrix.

Table 5.1: *Baseform and surface form phone alignments for an utterance in the ICSI training set with phone and word distance information.*

Word	Baseform	Surface Form	Phone Distance	Word Distance
yeah	/y/	/y/	0	0
	/eh/	/eh/	0	
that	/dh/	/dh/	0	13.33
	/ae/	/eh/	4	
	/t/	–	36	
is	/ih/	/ih/	0	0
	/z/	/z/	0	
an	/ae/	/ih/	8	4
	/n/	/n/	0	
easy	/iy/	/iy/	0	0
	/z/	/z/	0	
	/iy/	/iy/	0	
one	/w/	/w/	0	0
	/ah/	/ah/	0	
	/n/	/n/	0	

"them," and "to." Word fragments or phones not associated with any transcribed words were ignored in the alignments and for training pronunciation models, removing some insertion outliers that could change the reasonable alignments of the remaining phones. These changes generally led to higher quality alignments. For most regions other than these, the resulting alignments are quite similar. The main differences appear in severely reduced pronunciations. Example differences are seen in alignments for "everybody," and "going to" with extremely reduced pronunciations (see Table 5.2). Note that in "everybody" matching /w/ with /iy/ rather than /aa/ is due to more similarities in the tongue root and position

Table 5.2: *Comparison of baseform and surface form phone alignments for two phrases in the ICSI training set. Note that the baseforms are not always consistent.*

Phrase	Baseform	WS97	UW
everybody	/eh/	/eh/	/eh/
	/v/	–	–
	/r/	/r/	/r/
	/iy/	–	/w/
	/b/	–	–
	/aa/	/w/	–
	/d/	–	–
	/iy/	/ey/	/ey/
going to be	/g/	/hh/	–
	/ow/	–	/hh/
	/ih/	/ah/	/ah/
	/ng/	/n/	–
	/t/	–	/n/
	/uw/ (JHU)	–	
	/ax/ (UW)		–
	/b/	b	b
	/iy/	iy	iy

between /w/ and /iy/. This instance of "everybody" would have a word distance of 24. In "going to," with word distances of 28 and 25, our framework treats a glide like /hh/ as more closely related to a vowel than a consonant. Similarly, our approach assumes that it is more likely to have evidence of a syllable onset. The tongue position similarities between /t/ and /n/ combined with the syllable onset nature of the /t/ generate this alignment. While /t/ /uw/ is a valid "to" pronunciation in the dictionary used in this work, /ax/ was used in this particular alignment because there is a lower cost to delete /ax/ than to delete /uw/. Because utterance segmentations and transcriptions could be changed by the phonetic transcriber, the word transcriptions for a small percentage of the data differed depending on which release of the data was used. The data was organized with respect to the ICSI release along with utterance label in order to avoid word-level mismatch in the alignments.

5.1.2 Phone Transformation Categories

We examine two types of intermediate predictors for surface form phone prediction. The first is the word distance described above and the second is the transformation category of the phone change. The second predictor categorizes specific types of substitutions, insertions and deletions in terms of hypothesized cause (hyper vs. hypo-articulation). Phone insertions and substitutions of full vowels where a short or reduced vowel is expected suggest hyper-articulation. Phone deletions, substitutions of flaps for /d/, /t/, or /n/, substitutions of reduced vowels for expected full vowels and feature changes such as devoicing suggest reduction.[3] Substitutions can be classified as reductions of phones (hypo) or stress changes (hyper). In addition, there are other phone changes (e.g., from one full vowel to another) that could not be easily categorized as a reduction or hyper-articulation. Possibly these are associated with dialect differences (e.g., over 70% of "other" instances involve vowel changes)

[3]There are some contrasting examples however. Voicing can also represent hypo-articulation in the case of inter-vocalic consonants and some insertions can actually be related to hypo-articulation. Examples are excrescent stops which are present in some dialects, i.e., "else" pronounced as "eltse" in some dialects or a when a strongly devoiced vowel before a /t/ which can result in an /s/ being inserted before the /t/, but these cases are rare.

or assimilation.[4] Examples of the cases can be seen in Table 5.3. We divide our baseform and surface form phone pairs into 5 categories: same, hyper, reduction, deletion and other change. Predicting these categories allows examination of the factors that are associated with the specific types of articulation changes. Table 5.4 shows the relative frequency of these different types of phone transformations, compared to the frequency that the baseform phone is used. Not surprisingly because of the informal speaking style, phone transformations that may be associated with hypo-articulation (combining reductions and deletions) are more frequent than hyper-articulation and other transformations. Hyper-articulation is relatively rare, which may reflect the speaking style and/or may be a consequence of the particular baseforms used, but in any case it appears the least important source of variability in conversational speech. Appendix A.5 gives the details of the transformation types.

5.1.3 Evaluation Methods

Depending on what is being tested, different evaluation metrics are used. When the goal is word distance prediction, we predict a non-negative real number. In this case, we mainly use regression trees. In a few cases where the input variables are all numerical, we use generalized linear models (GLMs). For regression trees, root mean squared error (RMSE) of the distance is used to show performance gains. These can be compared to training set variance baselines. For GLMs, residual deviance which can be normalized to show RMSE and t values are also used for comparison of models [97].

When the goal is type prediction, whether a category or a surface form phone, we use classification trees. For classification trees, classification error rate (error divided by total) is used as one method of evaluation.

Our goal is to improve speech recognition. While reducing the phone error rate of prediction is a goal, when there is more than one likely pronunciation, error rate gives an incomplete picture. Hence, we also use entropy as a measure of improved output dis-

[4] In hindsight, it may have been useful to have an additional category for assimilation, because phone transformations of this type could have been in multiple categories. For example, the "find your" example in Table 5.3 has two phone transformations labeled as "other" and "deletion" that could more appropriately be described as assimilation. However, identifying these in general is somewhat complicated.

Table 5.3: *Examples of word pronunciations including hyper-articulation, other substitutions, reductions and deletions from the ICSI hand-labeled corpus.*

Word	Baseform	Surface Form	Type
and	–	/hh/	hyper
(in "smoker and a")	/ae/	/eh/	other
	/n/	/nx/	reduction
	/d/	–	deletion
reporting	/r/	/r/	
(in turn final phrase	/ih/	/ax/	reduction
"or the straight reporting?")	/p/	/p/	
	/ao/	/ao/	
	/r/	/r/	
	/t/	/dx/	reduction
	/ih/	/ih/	
	/ng/	/ng/	
	–	/k/	hyper
have	/hh/	/hh/	
(in many contexts)	/ae/	/ae/	
	/v/	/f/	reduction
find your	/f/	/f/	
(in many contexts)	/ay/	/ay/	
	/n/	/n/	
	/d/	/jh/	other
	/y/	–	deletion
	/uh/	/er/	other
	/r/	–	deletion
so	–	/t/	hyper
(in many contexts)	/s/	/s/	
	/ow/	/ow/	

Table 5.4: *Baseline distribution of phone transformations in the ICSI training set.*

Transformation	Percentage
Baseform	74.2
Hyper-articulation	1.7
Reduction	5.6
Deletion	9.7
Other	8.8

tributions of the pronunciation model as a way to assess the goodness of the predicted probabilities. Entropy gives a measure of how much the overall distribution is changing. Conditional entropy is defined as

$$E_{\phi,b}[\log p(\phi|b)].$$

We calculate entropy for the training set (using the Held Out test set to give an empirical estimate of $p(b)$) by summing the entropy for each phone transformation decision:

$$H_{train} = -\frac{1}{T}\sum_{j=1}^{T}\left(\sum_{i=1}^{N} p(\phi_i|b_j)\log p(\phi_i|b_j)\right) \tag{5.1}$$

where b_j is the baseform phone, ϕ_i is the predicted surface form phone, N is the total number of phones that can be predicted and T is the number of phones in the Held Out set.[5] Since we have hand-labeled data for the ICSI set, we can use the information about what is actually said to measure a Held Out entropy. Here we use only the empirical distribution of the baseform and surface form phones jointly in computing entropy:

$$H_{test} = -\frac{1}{T}\sum_{j=1}^{T}\log p_s(\phi_j|b_j). \tag{5.2}$$

Another useful measure is perplexity, which is related to entropy:

$$P_{test} = e^{H_{test}}. \tag{5.3}$$

[5]Note that context information may also be given in $p(\phi_i|b_j)$ but it is omitted here for brevity.

Roughly speaking, perplexity gives the effective number of alternatives at any point in the sequence. For direct comparison, the training set entropy is also presented as perplexity where

$$P_{train} = e^{H_{train}}.$$

In computing test set entropy, it is important that there are no zero probability events. Hence, the probability of the particular surface form ϕ_j given the baseform b_j is a smoothed distribution:

$$p_s(\phi_j|b_j) = \lambda_1 p_{DT}(\phi_j|b_j) + \lambda_2 p(\phi_j|b_j) + \lambda_3 p(\phi_j) \tag{5.4}$$

where $p_{DT}(\phi_j|b_j)$ comes from the decision tree and is context-dependent, $p(\phi_j|b_j)$ is the (context-independent) relative frequency that ϕ_j is the truth when the baseform is b_j, $p(\phi_j)$ is the overall relative frequency of ϕ_j in the training set, and $\sum_i \lambda_i = 1$. The λ_i values were chosen heuristically.

5.2 Baseline Attributes

Dictionary-based attributes have been used in the past for pronunciation modeling and we consider them to be our baseline attributes. From dictionaries, we get phonetic context, syllable position and stress information. For phone-related questions, we use categories describing the manner and place of the phone. Specifically, we use the following attributes of baseform phonetic context:

- consonant location: at syllable onset (before the vowel), or in the coda (after the vowel) (binary features);
- vowel phone in stressed syllable (either primary or secondary) (binary feature);
- phone at the beginning of a word (binary feature);
- vowel manner (6-valued feature) and place (11-valued feature);
- consonant manner (10-valued feature) and place (8-valued feature);

- manner and place of previous and following phones;
- phone category (vowel, glide, consonant, filled pause) of the previous and following phones (4-valued feature).

5.3 Analysis of Text Factors

In this section, we look at the usefulness of different high-level text factors for predicting the word distance, as well as a simple classification of phone transformation in terms of hyper-articulation vs. reduction. These factors have been useful for language modeling [42], speaker identification [109], utterance segmentation and disfluency detection [4], which have shown some correlation to pronunciation variation, so we hypothesize their usefulness in pronunciation modeling. We begin with a description of the text factors used and a distributional analysis of the factors. We then show how these can be used for predicting different intermediate variables.

5.3.1 Text Factors Used

Information variables that can be extracted from text or speech transcriptions include syntax and discourse information as well as information from dictionaries. In this work, attributes that are easy to obtain from text are examined. The attributes generated from word strings include:

- trigram language model scores,
- dialog act labels,
- syntax,
- part-of-speech (POS) groups,
- word category groups (content and function words, backchannels, filled pauses)
- word tags subsetting the function word category (used in surface form prediction only),

- pivot points in utterances,
- phone location in word (start or non-start), and
- word location in the utterance (beginning, middle, end, only).

The motivation for most of the factors was presented in Chapter 3. Dialog acts are clustered into 10 groups (including no available tag) based on words included in the acts (as in [56]). POS categories are based on an automatic POS tagger which takes into account word context. The multiple POS labels are clustered into 9 groups (as in [81]). The groups and the selection process are distinct from our word categories. Word categories use 5 classes aimed at conversational speech phenomena: backchannel, filled pause, content word, more accentable function word, and other function words. This division was motivated by observations of Jurafsky *et al.* [44] that planning problems – as represented by repetitions, pauses and filled pauses ("um" and "uh") – are strongly correlated with reduction of vowels and function words. Words are defined as members of classes regardless of their context. Word category requires nothing more than a table look-up to generate and, as will be seen below, is used differently than POS categories. Here, an analysis of all of these factors and their connections to pronunciation variability will be presented.

5.3.2 Distributional Analysis

In this section, we look at the relative importance of local words, syntax, and discourse cues to pronunciation variation, beginning with an illustration of the distribution of the word distance. Many have observed that function words have more pronunciation variation than content words, and indeed Figure 5.2 confirms this for the set of words that have non-zero cost. However, note that both word types have broad distributions, and the percentage of times that the two have non-zero cost is not so different: 54% of function word tokens (21K words) and 51% of content word tokens (29K words). In previous work, Fosler-Lussier showed a correlation between log trigram language model probabilities and word accuracy [26], providing support to the hypothesis that pronunciation variability is related to word predictability. Since word accuracy is only indirectly related to word variability (due to

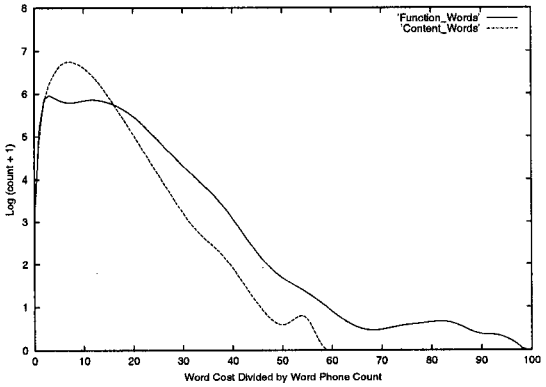


Figure 5.2: Smoothed content word and function word distance histograms for tokens with non-zero cost, showing the different distributions of pronunciation quality (word distance vs. log counts).

language model contributions in ASR decoding), we conducted a similar analysis using the word distance rather than word accuracy. The correlation between these variables is 0.1, and a scatter plot illustrating the relationship is given in Figure 5.3. While there is some information in the language model probability, especially for low probability words, this factor alone is not enough for useful prediction. Using the negative log trigram score in an analysis of variance gives a root mean-squared error (RMSE) of 9.5. For comparison, the standard deviance of the word distance (equivalent to RMSE) calculated using only the 5-class word categories for prediction is 8.4.

Next, we investigated the relationship between different high-level factors and the word distance. Figure 5.4 shows the distance distributions for the categorical factors described above (i.e., excluding trigrams). The word distance (or normalized phone distance per word) for the 9 POS groups used in this work differ as seen in Figure 5.4.a. The two groups defined by being more likely to be accented have the highest average pronunciation distance. While the pronunciation distance for dialog acts differs, as seen in Figure 5.4.b, the differences are smaller than that seen for the POS classes. The range of values is from 5.2 to 7.4, indicating that there are differences across dialog acts. Comparing the pronunciation distance for words before vs. after the pivot, we find an increased word distance before the pivot (7.4 vs. 5.8, respectively), which is consistent with the notion that the words before

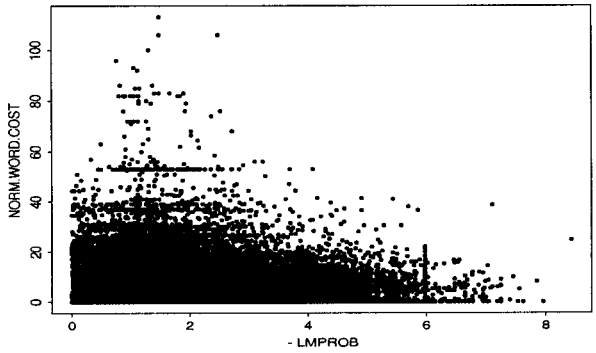
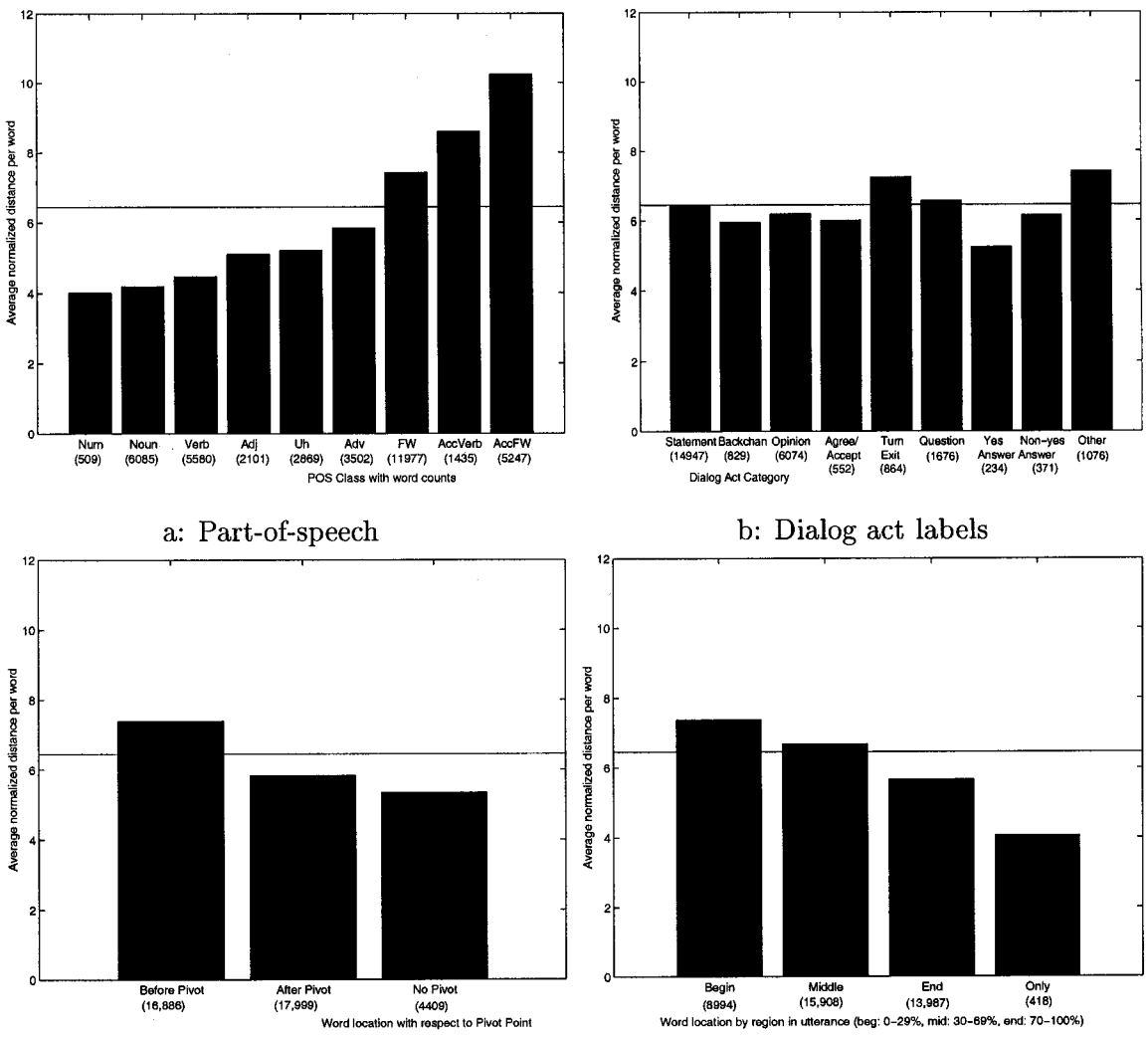


Figure 5.3: Relationship between word predictability and word distance: $-\log p(w_i|w_{i-1}, w_{i-2})$ vs. word distance.



a: Part-of-speech

b: Dialog act labels

c: Pivot point

d: Position of word in utterance

Figure 5.4: Relationship between text factors and word distance. Reference lines are the average for all classes. Numbers in parentheses are the number of instances of that type.

the pivot tend to be "given" and thus are more likely to be reduced. The word distance for words in utterances where there is no pivot is slightly lower at 5.3. (See Figure 5.4.c.) Since the automatic labeling of pivot points is based on the first verb phrase seen in an utterance and may not be accurate for compound sentences, another way to get at similar information is to look at the position of the word in the utterance (Figure 5.4.d) since the pivot point is

generally towards the beginning of the utterance. This gives a finer distinction between the regions of the utterance and can be calculated with very little processing. Again, words at the beginning of an utterance are less likely to match baseform pronunciations while words toward the end of the utterance will be closer to the baseform. Words that are the only word in the utterance tend to be greetings or closings ("hello", "good-bye") or backchannels ("uh-huh", "yeah", "okay") and are generally pronounced more like their baseform.

A limitation of using a single cost function is that it does not distinguish between hypo- and hyper-articulation. In an attempt to represent this difference, we characterized certain phone changes as associated with either hyper-articulation or reduction phenomena. As seen in Table 5.5, we also looked at the hyper-articulation vs. reduction categories as a function of part-of-speech label, with the hypothesis that POS classes that are more likely to be accented would tend to have higher percentages of hyper-articulation transformations and vice versa for POS classes that are not likely to be accented. We found that more accentable function words did indeed have a higher frequency of hyper-articulation phone transformations than other function words (2.3% vs. 1.6%, respectively), and a similar pattern for the more accentable verbs (1.8% vs. 1.1%). The highest rate of insertions (3.0%) was for the "uh" POS class, which includes exclamations like "really," "goodness," and "man" as well as backchannels and filled pauses. Interestingly, the more accentable words also had a higher rate of deletion compared to the corresponding words (11.0% vs. 5.1%, for function words, 15.0 vs. 7.5 for verbs), so the more accentable words seem to be exhibiting bimodal behavior.

The 5-class word categories are a different partitioning of words than POS and focus more on content/function word differences. These have a different distribution of hyper-articulation and reduction than the POS categories. The function word category is a closed set and includes "be" verbs which are not in the POS function word category. The word category of "filled pause" only includes "uh" and "um" so is different from the "uh" POS category which contains backchannels like "yeah". Additionally, the high level of insertions for the "filled pause" category may simply be a factor of the allowable baseform pronunciation. Our alignment dictionary allows for one filled pause phone (/fp/) for each word because that is the most common transcription for the word "uh" in the ICSI set. When the

Table 5.5: *The relative frequency of phone transformation types in the ICSI training set for text factors.*

Factor	Baseform	Hyper-artic.	Reduction	Deletion	Other Subs.
All Phones	74.2%	1.7%	5.6%	9.7	8.8%
POS: Num	80.8	1.1	5.2	7.4	5.5
POS: Noun	80.6	1.5	3.8	6.7	7.4
POS: Verb	80.3	1.1	6.1	7.5	5.0
POS: Adjective	76.9	1.8	2.8	9.3	9.2
POS: Uh	82.8	3.0	1.7	4.7	7.8
POS: Adverb	73.6	1.5	5.1	10.9	8.9
POS: Function Word	66.6	1.6	10.8	5.1	11.5
POS: More Accentable FW	64.3	2.3	10.4	11.0	12.0
POS: More Accentable Verb	67.6	1.8	7.3	15.0	8.3
WORDCAT: Content	77.2	1.7	4.5	8.8	7.8
WORDCAT: Function	68.8	1.2	7.3	12.4	10.3
WORDCAT: Accented Function	64.3	2.3	10.4	11.0	12.0
WORDCAT: Filled Pause	88.1	6.3	0.2	1.5	3.9
WORDCAT: Backchannel	86.7	3.4	0.2	2.0	7.7
Phone Location: Start of word	81.5	1.5	2.5	5.9	8.6
Phone Location: Non-start of word	70.8	1.7	7.0	11.6	8.9

transcription is /hh/ /ah/ or /ah/ /m/, the verb matches with /fp/ and the /hh/ or /m/ is labeled an insertion. Because the transcriptions are just as frequently /fp/ and because the acoustic model set has filled pause phones (/fpu/ for the vowel and /fpm/ for the nasal consonant), we chose to only use the /fp/ dictionary definition. This may artificially raise the number of insertions for this category only.

When looking at the location of phones in the word, we find a larger amount of variety for phones not at the start of a word than for phones at word onsets. This is consistent with Greenberg's work showing that: 1) phones at the onset of syllables are more likely to match the baseform and 2) Switchboard words are likely to be monosyllabic (over 80%) [116]. It should also be noted that when separating hyper-articulation into insertions and substitutions, word initial phones had more insertions (1.0%) than non-initial phones (0.7%) with the majority of the word initial insertions being glottal stops (/q/).

5.3.3 Prediction of Intermediate Variables

With the goal of predicting baseform to surface form phone transformations, we can take advantage of the connections shown in the previous section. We generated two types of intermediate predictor trees: word distance (normalized phone distance for a word) and transformation categories. In both cases, we looked at different syntactic and discourse cues for improving prediction using regression and classification trees respectively. The trees were built using different combinations of attributes in order to compare the usefulness of different types of variables. Trees were built using Splus with the splitting objective being to maximize the ability to distinguish between data observations. Tree growth continues until the leaf nodes are pure or until the data is too sparse to be split. Pruning is used to prevent overtraining on the data.

In order to easily combine the language model probability with the other attributes, a non-linear regression function (a generalized linear model or GLM) was used to predict the word distance with an RMSE of 9.5. This prediction only looked at word level information since the word distance is defined for a word rather than a phone. The value predicted by the trigram GLM is used as an attribute in combination with various categorical variables

Table 5.6: *Word distance prediction error using different attributes in a regression tree. RMSE = root mean squared error computed for the training set and using a held out portion of the ICSI held out set.*

		Distance (RMSE)		
			Held Out Set	
Expt	Factor	Train	Unpruned Tree	Pruned Tree
0	Baseline: σ of distance	9.5	10.8	–
1	trigram (GLM)	9.2	11.1	10.7
2	Dialog Act	9.5	10.8	–
3	current word category	9.4	10.6	–
4	current word POS	9.3	10.5	–
5	current word POS and category	9.3	10.5	–
6	POS window	9.2	10.5	–
7	POS & word category window	9.1	10.6	–
8	POS & word category window, + location of word in utterance	9.1	10.6	–
9	(1) + (8)	8.9	10.7	10.4
10	(1) + (2) + (8)	8.9	10.7	10.4
11	(2) + (8)	9.1	10.7	–

in a regression tree. The results are summarized in Table 5.6. The standard deviation (or average RMSE from the mean) of the word distance of the training set can be considered a baseline. For the training set, the RMSE is 9.5 and for the held out set, it is 10.8 (using the mean from the training set). While we expected a piecewise linear model to show a stronger connection between word distance and language model probabilities (as in Figure 5.3), it did not help at all (actually hurt) on the held out set. This is in contrast to the result reported by Fosler [27], which had the confounding factor of predicting a word recognition score, which includes language model information with the acoustic model score.

Part-of-speech is useful in the tree. In particular, a three word window of POS tags

brought down the RMSE for both the training set and the held out set. Additional attributes including a three word window of the 5-class word category, dialog act labels, location of the word in the utterance (beginning, middle, end) and location of the word with respect to the pivot point further reduced the RMSE for the training set but did not improve on the held out set. A sample tree can be seen in Figure 5.5. While additional information is useful for reducing the RMSE on the training set, POS and word category of the target word are the only useful attributes on the held out set.

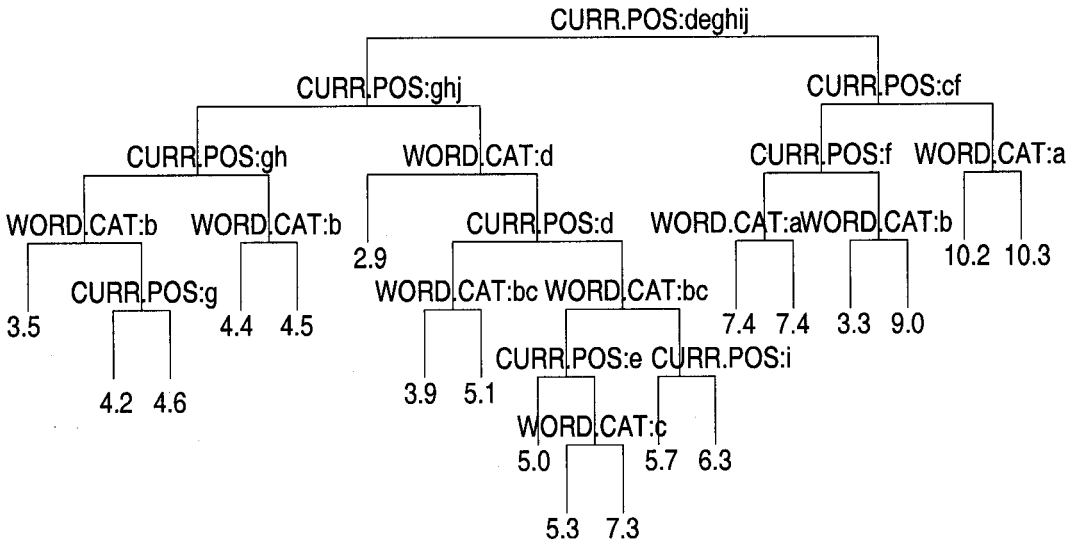


Figure 5.5: Regression tree for predicting word distance matching Expt. 5 from Table 5.6.

We also designed decision trees to predict five classes of phone transformations: deletion, reduction-related substitution, no change, hyper-articulation-related substitution or insertion, and other substitutions. In contrast to the experiments on predicting word distance, which gave word-level results, the experiments on phone transformation categories were at the phone-level, because we are moving towards prediction of actual phone transformations. Also, we hypothesize that some of our attributes may have better predictive effects at the word level and some may be more useful at the phone-level. Attributes used here included high-level variables as in the previous section as well as the baseline attributes described in Section 5.2 (excluding manner and placement features). Results are summa-

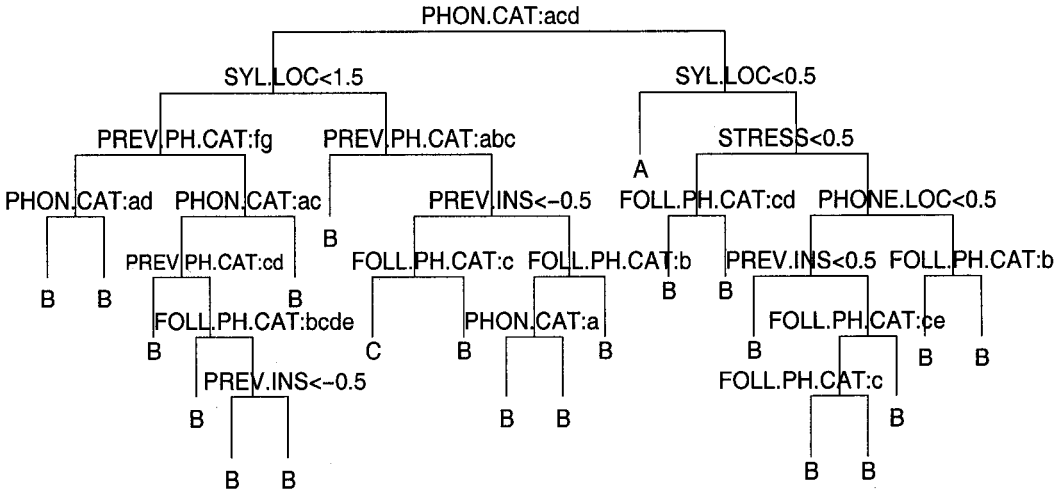
rized in Table 5.7. The experiments in Table 5.7 present the chance baseline (i.e., the result using priors) and the baseline using basic phone context information. Expts. 2a-c are based on the identity of the current word, while Expts. 3-9 ask questions about utterance level attributes.

Since the goal of the intermediate trees is to improve the prediction of lower level information, we also present in Table 5.7 the results of phone transformation type prediction using the word distance, both predicted (Expts. 10-12) and known (as an oracle experiment) (Expts. 13-15). The different versions of the word distance predictor, i.e., both the language model based GLM output and multiple decision tree (DT) outputs, were used as inputs to the transformation category trees. The best results of those experiments did not significantly change the results. However, the oracle experiments show that much more can be gained from using the word distance. The experiments provide performance bounds of 16.7% error on the training set and 21.7% error and perplexity of 1.80 on the held out set (Expts. 13-15). Note that the known word distance itself performs as well as the use of phonetic context and significantly better in combination. Hence, improved prediction of the word distance, and by extension other intermediate predictors, should help in this case as well as in surface form prediction.

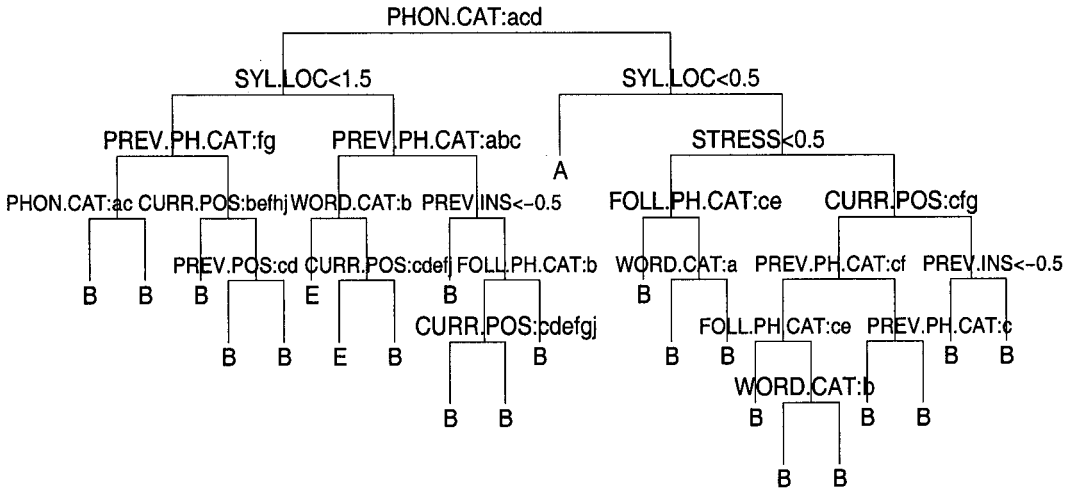
The best misclassification rate on a held out data set was 29.4% using the POS and word category windows, which can be compared to 31.9% error for assigning all cases to the "no change" class (an 8% reduction in error on 14,856 observations). Although we included dialog act labels in some experiments, they did not help. While the pivot feature was useful for prediction of the held out set (Expt. 4) when it was the only available non-phone level feature, it hurt the misclassification result when included with the other factors but matched the best perplexity reduction (from 2.55 to 2.3). Removing the pivot attribute from the combined results improved the prediction but slightly increased entropy. Even though the hyper-articulation class is infrequent (1.5%), the decision trees do find contexts for predicting it, in particular, when the phone is at the beginning of a syllable. This can be seen in the sample trees in Figure 5.6 (pruned for display), where "A" is the hyper-articulation class, "B" is the no change class, "D" is a reduction, "E" is a deletion and "C" denotes all other substitutions.

Table 5.7: *Phone transformation type prediction error using different text-based attributes in a classification tree for the training and held out portions of the ICSI set.*

		Transformation Type			
		% Error		Perplexity	
Expt	Factor	Train	Held Out	Train	Held Out
0	Baseline: context-independent	25.8	31.9	2.44	2.82
1	phone information	24.7	30.7	2.01	2.55
2a	(1) + POS	24.0	31.1	1.93	2.48
2b	(1) + word category	23.9	30.7	1.95	2.47
2c	(1) + POS + word category	23.5	30.2	1.92	2.40
3	(1) + word category window	23.9	30.2	1.95	2.38
4	(1) + Pivot	24.7	29.8	1.97	2.54
5	(1) + word location	24.6	30.7	1.99	2.57
6	(1) + Pivot + word location	24.6	30.1	1.95	2.53
7	(1) + POS window	23.8	29.6	1.93	2.40
8	(1) + POS & word cat. windows	23.4	29.4	1.95	2.34
9	(8) + Pivot	23.4	29.6	1.95	2.30
10	(8) + trigram GLM word distance predictor	23.4	29.4	1.95	2.33
11	(8) + DT distance predictor	23.4	29.5	1.95	2.30
12	(10) + (11)	23.4	29.5	1.95	2.30
13	(1) + word distance (oracle)	17.7	24.5	1.70	1.89
14	(7) + word distance (oracle)	16.6	21.7	1.65	1.80
15	word distance (oracle)	24.2	29.7	2.05	2.18



a. Tree designed using only phone information.



b. Tree designed using questions about phones, POS and word category windows, trigram GLM and the word distance predictor.

Figure 5.6: Classification trees for predicting phone transformation categories matching Expts. 1 and 10 from Table 5.7. (Only the top nodes of the trees are shown.)

5.3.4 Discussion

We have shown a connection between pronunciation variation and text based factors like part-of-speech, discourse, language model scores and location information, but POS and word category attributes are most useful of these in building intermediate predictors. There are clear differences in reduction behavior when looking at word and POS categories, with reduction happening more often on function words as would be expected. These factors are useful for predicting intermediate variables.

In general, local factors are more important for text attributes than utterance-level factors although pivot point and word location in utterance combined to reduce the perplexity and reduced the error rates. Word context, e.g., the POS window, improves the error rates further but gives somewhat mixed results in the amount of improvement between error rates and perplexity reduction. The largest improvement in error comes with the addition of POS values and word categories of the word and its neighbors. While this combination has an 8.2% reduction in perplexity, the perplexity is further reduced to 9.8% when the pivot information or decision-tree-predicted word distance are included. (These results are significant as shown in Section 5.6.)

There are trade-offs between improvement in percentage error and perplexity. Since our goal is to build pronunciation models predicting a variety of phone changes, reduction in perplexity, even when there are no changes in error, shows that the resulting distribution is improving. POS and word categories give some information about hyper-articulation according to our data analysis (Table 5.5 and Figure 5.6), but location of phone in syllable is most useful for prediction. These along with phone location in word are useful for predicting the chance of reduction. Using high level features, we are able to predict reduction sooner than using phone information alone. In looking at the types of phone transformations, phone location in word helps differentiate between the amount of reduction more than substitutions and hyper-articulation.

The use of word distance as an intermediate predictor in the pronunciation transformation type decision trees had a small effect when combined with text features. However, when the known word distance was used, there was a significant improvement. This suggests that

there is value to the use of the word distance as an intermediated predictor and to our two stage prediction strategy.

5.4 *Analysis of Prosodic Factors*

This section presents an analysis of acoustic correlates of prosody as seen in F0, duration and energy, which parallels the study of word-based cues. There will be a description of the factors, analysis of their relation to word distance and transformation types, as well as a discussion of their usefulness for predicting intermediate factors.

5.4.1 F0, Energy and Duration Factors

We use three types of **acoustic measures** of prosody: duration, energy and F0. The motivation for using prosody and acoustic correlates can be found in Chapter 3.

Duration values include length of the utterance, word and phone as well as the word duration normalized by (divided by) the utterance length and the phone duration normalized by the word duration. The first normalized number could be thought of as information about the relative importance of the word in the utterance. If it is a large percentage but the word count is high (or utterance duration is long), this may be a content word or emphasized. Similarly, it may be the only word in the utterance If the value is a small percentage, the word may be a function word or carry a small amount of information within the sentence.

Energy measures include mean, minimum and maximum energy values over: the utterance, word, and windows of 15 and 30 frames prior to the end of the word. These values were also normalized by the energy at the beginning of the conversation for an SNR-type measure by dividing the values by the average energy of the first ten frames of the conversation side. In some SNR cases, log values were also used.

F0 values were generated using `get_f0`, the ESPS pitch tracker [49, 103]. They were then processed using software developed at SRI [95]. The first step smoothed the F0 values to reduce doubling and halving artifacts. Piecewise-linear fits of the F0 values were then generated, creating slopes that concisely describe the F0 trends, i.e., falling and rising trends are more easily seen. An example can be seen in Figure 5.7. The piecewise-linear fits to

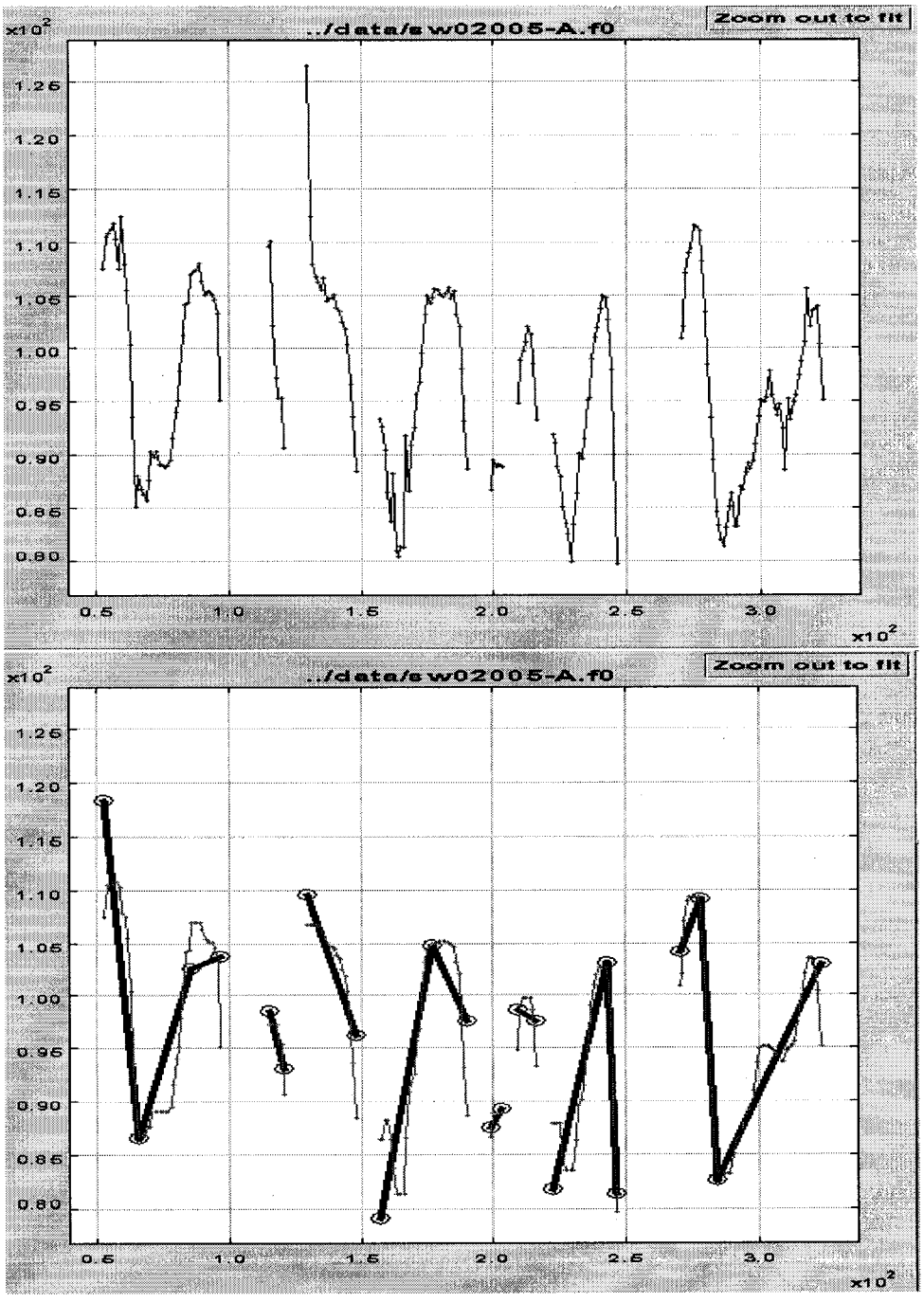


Figure 5.7: Processed F0 values: top figure shows raw F0 values, and bottom shows smoothed F0 contour (thick lines) and stylized slopes (thick lines).

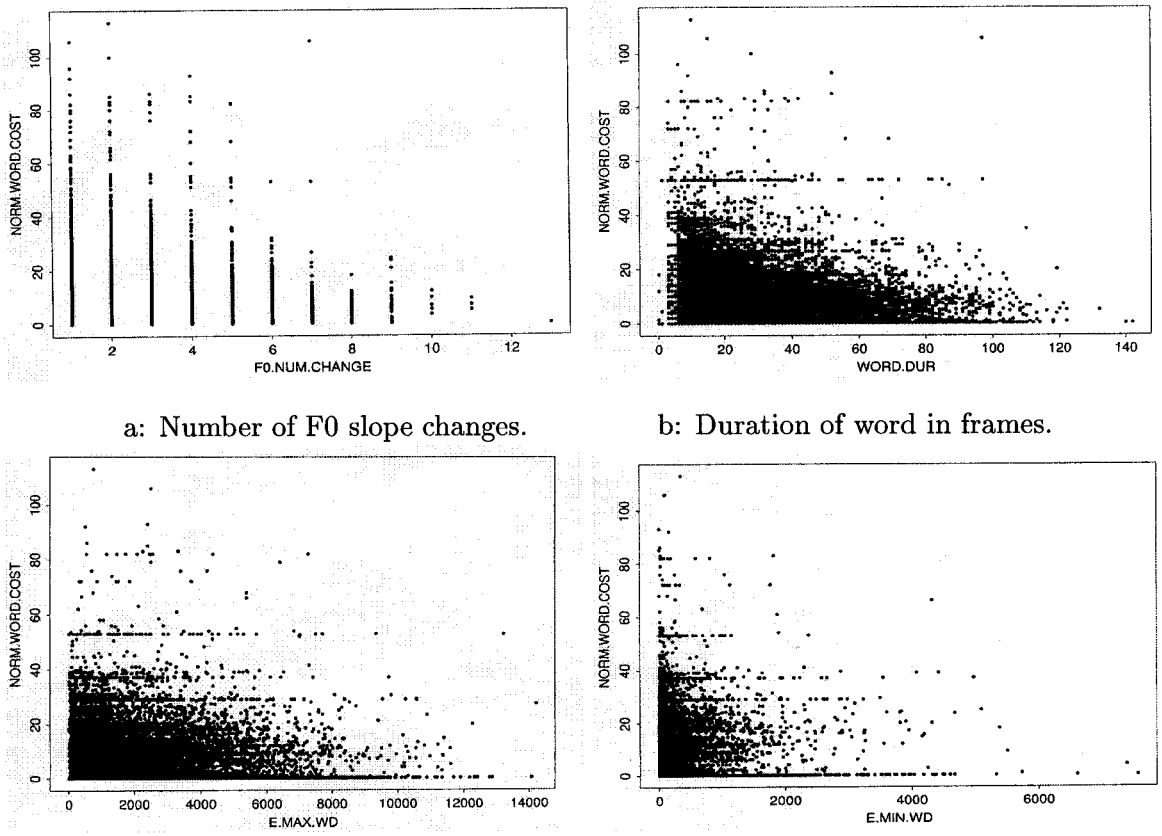
the F0 values are used in this work, facilitating extraction of slope and shape features.

These processed F0 values have been used with success in decision tree prediction, for such applications as speaker verification [96], speaker identification [109] and disfluency and punctuation detection [92]. The slopes of the piecewise lines and the line endpoints are used as attributes. Along with mean, minimum and maximum F0 values for the utterance, word and 15 and 30 frame windows moving back from the end of the word boundary, slope values and the number and value of slope changes in an utterance or word are used in our predictive trees. F0 values are normalized by the speaker baseline F0 in two ways; the baseline value can either be subtracted or used as a divisor. Normalizing F0 values is important to reduce speaker dependent variation. Counts of the number of frames estimated to have F0 halving or doubling are also used. All together there were 4 duration measures, 23 energy measures, and 32 F0 measures available for tree design.

In Chapter 3, we described the possible conflicting cues that can be presented with prosody. Low F0 could be a low pitch accent (representing emphasis), associated with no pronunciation changes or hyper-articulation, or mumbling, associated with reduction. Similarly, long duration could indicate emphasis, phrase final lengthening or a slow speaking rate. When the case is emphasis or slow speaking rate, pronunciations are more likely to be canonical, whereas phrase final lengthening will more likely generate phone deletions and/or reduction. The combination of different cues as well as normalizing can help address this.

5.4.2 Distributional Analysis

A visual inspection of the distributions of attributes extracted from the raw F0, energy and duration show that the shapes of the distributions are often similar to that seen in Figure 5.3 for the trigram scores. Example plots can be seen in Figure 5.8, where larger variation in all F0 variables is associated with lower word distance values. Because many of the prosodic attributes are numerical rather than categorical, we were able to include them in building generalized linear models (GLMs) for predicting word distance based on a Gamma distribution. The attributes that stood out as being most correlated (using a Pearson correlation test) with the word distance are shown in Table 5.8. The word distance predicted by GLMs can be used as an attribute in combination with various categorical and



a: Number of F0 slope changes.

b: Duration of word in frames.

c: Maximum energy in word.

d: Minimum energy in word.

Figure 5.8: Examples of relationships between prosodic values and word distance (labeled as NORM.WORD.COST).

Table 5.8: *Word-level prosodic attributes most correlated with word distance. Correlations of remaining attributes were between ± 0.03.*

Attribute	Correlation
Trigram LM Probability	0.10
Max. word energy	-0.04
Min. word energy	-0.08
Word duration	-0.16
Word dur/Utt. dur	-0.10
F0 num. changes in word	-0.11
F0 num. changes in word/dur	0.14

numerical variables in a regression tree. This will be described in the next section.

5.4.3 Prediction of Intermediate Variables

As with the text-based variables, the prosodic attributes are used in combination to generate predictors of word distance and phone transformation categories. The first modeling approach is to use GLMs to predict word distance, since the numeric values of the acoustic attributes of prosody are well suited to the GLM framework. The second approach is to build decision trees for predicting both word distance and transformation type using combinations of the raw prosodic features and the output of the GLM predictors.

Table 5.9 compares four different GLMs used for predicting word distance. The outputs of these predictors can then be used as inputs to the decision trees. Word duration is a better predictor of word distance than trigram language model probabilities. However, adding the LM value does improve the GLM slightly. Energy had a very small effect. This is not unexpected given the correlation results in Table 5.8. The RMSE on the held out set for all predictors is 13.1, indicating that this form of predictor is not very useful. Because there are some missing F0 values (in unvoiced regions), no F0 values could be used in building the GLMs.

Table 5.9: *RMSE and t values for generalized linear models predicting word distance with different word-level attributes. Values are reported on the training set.*

Attribute	RMSE	t value
Trigram LM Probability	9.5	57.7
Duration	9.4	43.0
Duration + Energy	9.4	39.2
LM + Duration + Energy	9.3	32.1

Multiple trees were built with different groups of attributes and the best results are summarized here. The results are reported for training and held out portions of the ICSI data set. Tree design is the same as used in Section 5.3.3. Because there are so many available attributes, especially for F0, those selected by the trees were labeled "best of" and used for experiments combining types of attributes.

The word distance results are summarized in Tables 5.10 and 5.11 with results given in terms of prediction RMSE. Trees built with the prosody variables directly (Expts. 1-4, Table 5.10) do as well as the LM GLM based tree (Expt. 1A, Table 5.11) on the training set and better on the held out set. When combined (Expt. 4), prosodic features give a better result than any of the trees based on GLM values alone (as in Table 5.11) and better than the text-based results presented in Table 5.6. From this, we can see that the interaction between duration, F0 and energy is important, as anticipated. The addition of word-based variables (POS and word category labels for the current, previous and next words) in the tree reduces the error on training but not test data. Pruning the tree helps, however, and yields a 2.9% improvement over the text-based tree, showing prosodic values do aid the prediction of word distance values.

While using the GLM as predictor variables in a decision tree is better than using them on their own as in Table 5.9, it is still not an improvement over the baseline variance of the test data alone. Including prosodic attributes directly (Expt. 5A), especially energy and F0 values gives improvement, consistent with Expt. 5. While duration is useful when not

Table 5.10: *Pronunciation word distance prediction error using prosody-based attributes in a regression tree. RMSE = root mean squared error computed using the training set and a held out portion of the ICSI set.*

Expt	Factors	Distance (RMSE)		
		Train	Held Out	Pruned Tree
0	Baseline	9.5	10.8	–
1	Duration	9.2	10.9	10.5
2	Best of F0	9.2	10.8	10.4
3	Best of Energy	9.1	11.0	10.6
4	Energy + F0 + Duration	8.9	10.5	10.5
5	(4) + word based	8.7	10.5	10.1

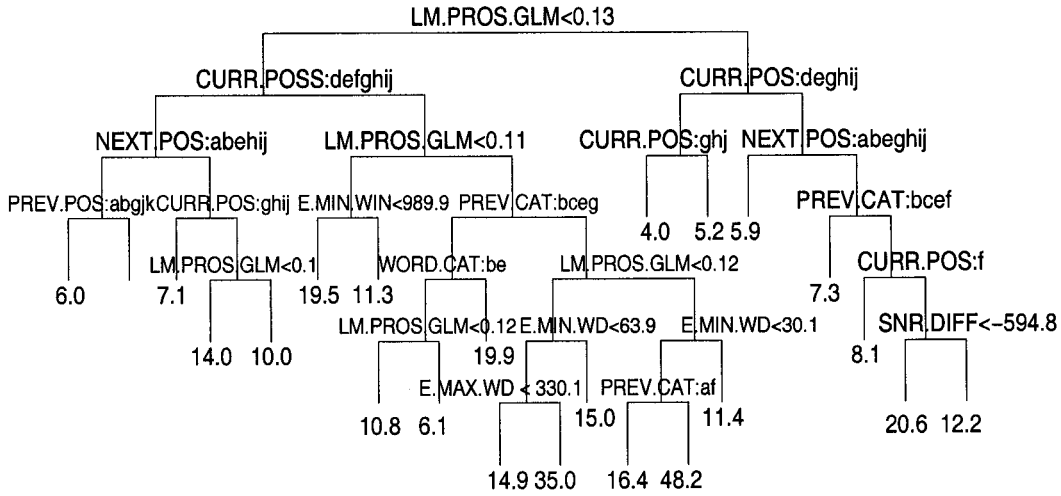
Table 5.11: *Pronunciation word distance prediction error using GLM-based and prosodic attributes in a regression tree. RMSE = root mean squared error computed using the training set and a held out portion of the ICSI set.*

Expt	Factors	Distance (RMSE)		
		Train	Held Out	Pruned Tree
0	Baseline	9.5	10.8	–
1A	GLM (LM only)	9.2	11.1	10.4
2A	GLM (prosody)	9.1	11.1	10.5
3A	GLM (LM, prosody)	9.1	11.1	10.5
4A	(3A) + word based	8.8	10.5	10.2
5A	(4A) + Energy, F0, Duration	8.7	10.5	10.1

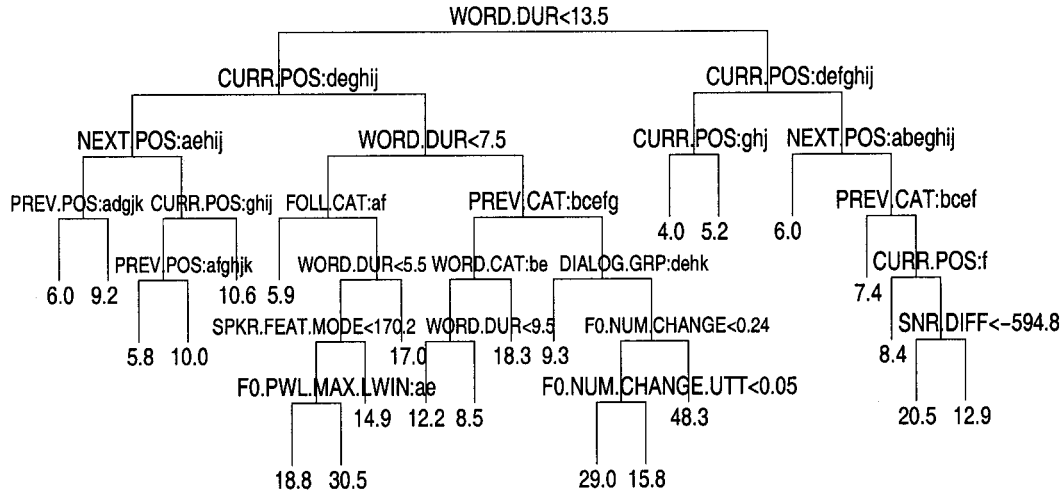
combined with word based or GLM values, its use did not have an impact on the results of combined models even though duration variables were chosen by the trees. This is likely because the GLM uses duration so any benefits are already incorporated. Examples of the trees used can be seen in Figure 5.9 (top nodes only). In either case, however, word duration is clearly a powerful attribute as it is represented at the tops of both trees, in raw form in Figure 5.9.b and through the incorporation of the prosody-based GLM in Figure 5.9.a.

The best starting point for predicting phone transformation categories was using word-based phone- and word-level variables; decision trees based on prosody alone were not better than chance. Prediction results are given in Table 5.12 with perplexity and percent error as evaluation criteria. On the held out set, text information had an error rate of 29.4% compared to 31.9%. The prosodic attributes used to build the combined trees were the ones chosen by trees built on text attributes plus energy, duration or F0 alone. Adding prosodic features did not reduce classification error rate, but it did reduce perplexity. The F0 values used in the tree include the number of F0 slope changes in the word (normalized by the word duration and unnormalized), the maximum and minimum F0 values of the word, the difference between the F0 slope at the end of the word and the slope at the beginning of the next word, and the count of the frames affected by pitch halving in the word. Minimum, maximum and mean energy values were used along with word duration. These predicted transformation categories, along with a confidence score based on the difference between the most likely and second most likely categories, are used in the surface form prediction.

The addition of various F0 values to the text-based trees reduced the perplexity of the model with respect to text variables alone but increased the error rate. Energy attributes had little effect on the result on their own or in combination with F0 or duration. While duration had a negative impact on the results for the training set, it reduced the perplexity on the held out set more than any other attribute. When all three types of prosodic information were available for the trees to use, energy, duration and F0 values were chosen, but there was no additional performance gain over duration only. The trees with duration had an 8.4% reduction in error with a perplexity reduction of 2.4% (from 2.34 to 2.26). The mixed results seen in the training misclassification rates when adding information may be due to the greedy growing nature of the trees. The trees look very similar to those for



a. Tree designed using questions about text attributes, GLM and energy.



b. Tree designed using questions about text attributes, GLM and energy, F0 and duration. (Expt. 5A from Table 5.11.)

Figure 5.9: Top nodes of regression trees for predicting word distance.

Table 5.12: *Phone transformation type prediction error using different prosodic attributes in a classification tree for the training and held out portions of the ICSI set.*

		Transformation Type			
		% Error		Perplexity	
Expt	Factor	Train	Held Out	Train	Held Out
0	Baseline: context-independent	25.8	31.9	2.44	2.82
1	Text variables	23.4	29.4	1.95	2.34
2	(1) + F0	23.3	29.5	1.93	2.29
3	(1) + duration	23.5	29.4	1.97	2.26
4	(1) + F0, energy	23.4	29.6	1.95	2.28
5	(1) + F0 + duration	23.5	29.4	1.97	2.26
6	(1) + F0, duration, energy	23.5	29.4	1.97	2.26

text-based variables as the text attributes are generally high in the tree. Pruned versions of these trees make it difficult to see which prosodic attributes are more important. Since this is a primary goal of this section, pruned trees are not used here.

5.4.4 Discussion

The prosodic features most correlated with word distance are word-level minimum and maximum energy values, word duration (alone and normalized by utterance duration) and the number of times the F0 curve changes in the word (alone and normalized by utterance duration). Prosodic values improve slightly the prediction of word distance and phone transformation categories over just using text-based variables. Again, word-level attributes were used by the trees more often than longer range values. Differences in energy across words and utterances were not chosen attributes although attributes local to the words were used. Prosodic information did improve the perplexity for transformation type prediction but because the gains were small, the question of whether prosodic cues helps remains an open question as we investigate its usefulness for surface form prediction. It should be noted that

word and utterance-level attributes are used for word distance and phone transformation category prediction. Phone-level information is also used for phone transformation category prediction.

5.5 Surface Form Prediction

This section examines the prediction of surface form phones. Previous work [7, 78] uses a set of contextual factors (described in Section 5.2) in decision-tree-based predictors of surface forms. These will be used as our baseline. Surface form prediction using these baseline attributes with text information and/or prosodic attributes will be described in this section. Note that, while improving pronunciation modeling through better surface form prediction is our goal, better phone prediction does not guarantee better automatic speech recognition results. ASR performance will be examined in Chapter 6.

5.5.1 Text Experiments

In order to assess whether high-level attributes have a significant impact on the pronunciation model – and whether an intermediate predictor is useful – we conducted experiments predicting the surface form phones from the baseforms. Several new trees were then grown adding higher-level attributes to the baseline set: log trigram score (or the same processed by a GLM), POS tags for the word and its left and right neighbors, word location in the utterance, word location relative to the pivot, and dialog act. For the cases using the phone transformation type predictor, two attributes were used: the most likely transformation class and the probability difference between this and the next most likely class. Decision trees are built for each baseform phone using Splus with different sets of possible questions for each experiment. Trees were built with different levels of pruning (using different cost-complexity parameters, k, in cross-validation on the training data) based on the size of the tree. Results presented here all have a pruning level of $k = 10$. Discussion and presentation of the pruning experimental results can also be found in [5]. We compared the direct use of the different factors to using an intermediate predictor (word distance or phone transformation type), and to the combination of these approaches by comparing classification error

rate and perplexity.

The results are summarized in Table 5.13.[6] The best result represents roughly a 1.3% reduction (30.7% to 30.3%) in phone prediction error on the held out set relative to the baseline using phonetic context and other word-internal variables. However, none of the high-level attributes led to a reduction in perplexity. The predicted phone transformation type was more useful than the predicted word distance for the held out set, and had slightly better perplexity than the case where all the text-based attributes were used directly. Although the inclusion of the LM GLM and predicted word distance reduced training set misclassification and perplexity, they only had a negative effect on the held out set. For the intermediate predictor-based trees in Table 5.13, long distance attributes (i.e., beyond the current word) are incorporated only in the intermediate predictors (trigram GLM, predicted word distance and predicted transformation type) and are not directly available to the tree. However, in the trees with text-based attributes (both alone and with intermediate predictors), text-based attributes chosen by the trees include such long distance attributes as dialog act, word categories and POS for a three-word window, and pivot information. When the predicted transformation type is used, more trees choose the confidence measure (showing how certain the predicted type is) as a feature than the predicted transformation type.

Our use of the intermediate predictors is supported by the cheating experiments at the bottom in the Table 5.13. When the true word distance and true phone transformation types are used as attributes in the decision trees, the results are surprisingly good, 13.7% error rate on the held out set. This suggests that our hypothesis that using the intermediate predictors will improve the surface form prediction is reasonable, and it supports the need for further examination of the prediction of intermediate variables.

[6]Insertions are included in misclassification and perplexity values. We include them here (but not in later chapters) because they are one of our five transformation type categories. While current recognition systems are typically not implemented to allow for insertions, they may be in the future.

Table 5.13: *Misclassification rates and perplexity for surface form phone prediction on trees with pruning (cost-complexity parameter k=10) using text attributes only. The average number of nodes per phone tree is also listed.*

Type of Attributes	Avg. Nodes per Tree	% Error		Perplexity	
		Train	Held Out	Train	Held Out
Context-independent	NA	26.8%	33.0%		3.22
0. Baseline: phone context	20.4	21.9%	30.7%	1.92	2.94
1. (0) + text based	41.9	20.4%	30.6%	1.77	3.05
2. (0) + trigram GLM	36.1	21.5%	31.1%	1.84	3.05
3. (0) + predicted word distance	34.0	21.6%	31.0%	1.86	2.99
4. (0) + predicted transform. type	25.5	21.4%	30.6%	1.86	2.95
5. (0) + all of above	47.7	20.1%	30.3%	1.73	3.07
6. (0) + oracle word distance	47.4	10.9%	19.6%	1.53	2.08
7. (0) + oracle transform. type	11.0	5.6%	14.8%	1.28	1.51
8. (1)+ oracle word distance + oracle transform. type	14.5	3.8%	13.7%	1.18	1.47

5.5.2 Prosody Experiments

In order to assess whether prosodic variables have a significant impact on the pronunciation model we conducted experiments predicting the surface form phones from the baseforms. Decision trees were built using individual sets of prosodic attributes for each baseform phone. The results are summarized in Table 5.14. Expts. 1-4 present prosodic variables alone. F0 and duration perform better than chance. When prosodic attributes are included with the phone based information, training misclassification hinted at a win but this did not hold for the Held Out set (Expt. 5). Including prosodic attributes helped in many of the experiments compared to the text-based baselines presented in Section 5.5.1. The intermediate transformation category predictor also improved the result, more so than the text-based case. Note that, here, intermediate predictors were allowed to include prosodic information

Table 5.14: *Misclassification rates and perplexity for surface form phone prediction using prosodic attributes with pruning (cost-complexity parameter k=10). Held out set baselines for text-based attributes only are included in parentheses.*

Factors	Avg. Nodes per Tree	% Error		Perplexity	
		Train	Held Out	Train	Held Out
Context-independent	NA	26.8%	33.0%		3.22
0. Phone Context	20.4	21.9%	30.7%	1.92	2.94
1. F0	60.7	23.7%	31.6%	2.36	3.07
2. Energy	64.2	23.2%	33.7%	1.92	3.56
3. Duration	52.1	24.1%	32.4%	2.03	3.38
4. All Prosody	65.7	22.3%	32.4%	2.14	3.19
5. (4) + phone info (30.7)	59.1	19.0%	30.8%	1.80	2.98
6. (5) + word info (30.6)	59.3	18.4%	30.4%	1.77	2.98
7. (5) + LM/pros. GLM (31.1)	59.8	19.0%	30.7%	1.80	2.98
8. (5) + predicted distance (includes prosody) (31.0)	59.5	19.0%	30.8%	1.80	2.99
9. (5) + predicted transform. types (30.6)	58.1	18.8%	30.0%	1.80	2.92
10. (5) + 3 intermed. predictors	59.7	18.9%	31.1%	1.80	2.95
11. (10) + word info (All Attributes) (30.3)	59.9	18.3%	30.9%	1.77	3.01

as well as text. As with the text-based experiments, the use of the LM and prosody-based GLM and the predicted distance do not improve the results (except when compared to the text-based case). The best result, using prosodic information, phonetic context and the predicted transformation type (Expt. 9), represents a 2.0% relative reduction in phone prediction error relative to the text-based case (30.6% to 30.0%) and a 2.3% relative reduction relative to the baseline using phonetic context and other word-internal attributes (30.7% to 30.0%), but only a small reduction in perplexity compared to the baseline (2.94 to 2.92). These results are significant, however, as shown in Section 5.6.

An examination of the effects of pruning shows that it has a larger impact on the trees built with prosodic attributes, possibly because the trees generally have more nodes than the text only trees. Using a cost-complexity parameter of k=10 for the trees built with prosody, reduces the tree size considerably with, generally, a very small impact on the entropy and the error rate of the Held Out set. In all cases, pruning reduced the misclassification error rate on the Held Out set.

5.5.3 Discussion

Surface form prediction using prosodic and text attributes leads to a 2.3% improvement over just using dictionary information. Prosodic variables did not help when used directly for surface form prediction, but they did help when used in intermediate predictors. While some of the information variables used may be correlated, it is generally of value to use both types of information (word-based and prosodic) in pronunciation prediction. F0 is the single most important prosody variable (as seen in Expts. 1-4 of Table 5.14). Predicted transformation types are the most useful intermediate predictor (as seen in Expts. 4-9). Note that the finding that F0 is useful for prediction is in contrast to previous findings.

5.6 Significance Testing and Error Analysis

While the prosodic and text-based attributes vary in their contributions to either changes in prediction accuracy or reduction of entropy, it is important to analyze whether changes using the attributes described here are significant or not. We use two basic tests (available through

Table 5.15: Significance testing for transformation category prediction error and perplexity calculated using t-tests for error and perplexity results. (Values of 0 are less than 10^{-5}.)

Result Table	Experiment Pair	Significance Level (p)	
		Error	Perplexity
5.7	0 vs. 1 (Baseline vs. phone info.)	0.02	0
5.7	1 vs. 9 (phone info. vs. word-level text info)	0.02	0
5.7	7 vs. 14 (POS window vs. POS window + oracle distance)	0.00001	0
5.12	1 vs. 6 (text-based vs. text + prosody)	No difference	0

Splus): 1) a Gaussian-based t-test for measuring the significance of overall prediction error for transformation categories and for phone-level transformations, and 2) a paired t-test for measuring the significance of perplexity changes (using the negative log probabilities used in the perplexity calculations). There are 14,856 samples in the held out set.

Significance results for selected phone transformation type prediction experiments are in Table 5.15. For transformation categories, all text-based experiments are significantly better than the baseline with $p < 0.02$. There is no significant difference when adding prosody to the text-based system when examining the error rate, but the reduction in perplexity is significantly different.

Another way we can analyze the differences in the systems is to look at confusion matrices. In Table 5.16, we see that adding more information, particularly prosodic information, results in the ability to predict and, thus differentiate, all five categories. Correctly predicted hyper-articulation types in these experiments include reduced vowels going to full vowels as well as the insertions of phones like /q/ and /hh/. The reduction category is the most difficult to predict, even though it is more likely than other substitutions or hyper-articulations. Some of the "Other" category can be explained by assimilation, as in the case of /t/ going to /ch/ when followed by an /r/, and could possibly be re-categorized as

Table 5.16: Confusion matrices for transformation category prediction experiments. A: Phone information alone. B: All text-based information (matching Exp. 11 in Table 5.7). C: Both text-based and prosodic information (matching Exp. 6 in Table 5.12).

A (30.7% error)	Hypotheses				
Truth	Hyper	No Change	Other	Reduced	Deleted
Hyper	173	162	0	0	0
No Change	0	10119	0	0	0
Other Substitution	0	1954	0	0	0
Reduced	0	979	0	0	0
Deleted	0	1469	0	0	0

B (29.5% error)	Hypotheses				
Truth	Hyper	No Change	Other	Reduced	Deleted
Hyper	180	149	6	0	0
No Change	11	9971	33	0	104
Other Substitution	16	1815	107	0	16
Reduced	0	948	0	0	31
Deleted	4	1197	31	0	237

C (29.4% error)	Hypotheses				
Truth	Hyper	No Change	Other	Reduced	Deleted
Hyper	180	146	7	1	1
No Change	11	9879	39	64	126
Other Substitution	16	1761	126	14	37
Reduced	0	905	0	39	35
Deleted	4	1140	37	20	268

reduction phenomena using transformation rules that incorporate phone context.

For surface form phone prediction, the differences in error rates are not significant. However, using the paired t-test for the perplexity values does show significant differences between the systems examined. For text-based results (seen in Table 5.13), including text information alone has no significant difference from phone context alone. Although the perplexity values are significantly different, they are worse. Including oracle transformation type or word distance results is significantly better than text-based prediction alone ($p < 10^5$ for both prediction error and perplexity). When prosody is included for surface form prediction (Table 5.14), the differences in error rates are not significant and the perplexity values do not decrease. When prosody and predicted transformation types are used (Expt. 9), we have the largest reduction in error rate compared to phone context alone (30.7% to 30.0%), but the reduction in perplexity is not significant.

5.7 Summary

Overall, our analysis showed that text and prosodic attributes are useful in prediction of pronunciation variation. While additional research may lead to a deeper understanding, some answers have been obtained to the questions raised at the start of this chapter.

1. What syntactic and discourse factors correlate with pronunciation variability? What prosodic factors?

   There are correlations between text-based attributes and pronunciation variability for some of the factors investigated here. Part-of-speech and word category information are the most useful attributes for predicting intermediate predictors of surface-form pronunciation variation. Discourse features were not useful for the intermediate stages although they were used for surface-form phone prediction. The use of prosodic attributes does improve prediction slightly, with duration values being the most useful individual predictors for transformation type prediction and F0 values being most useful for surface-form prediction. The combination of F0 and duration attributes is used often and, we suggest, helps to disambiguate between conflicting prosodic cures.

2. Do local variables have substantially more impact than non-local ones (e.g., dialog act vs. neighboring POS for text, local vs. utterance measures of energy or F0 for prosody)?

   Local word-level values are typically more useful than utterance-level values. For text cues, immediate word context is most important, as seen in the use of POS and word category windows. However, the relationship of words to dialog act labels, an utterance pivot point and the position of the word in an utterance do have some correlation with pronunciation variation. For prosodic features, word-level attributes are used more frequently for predicting variation though these are normalized by either utterance or speaker-level values. For example, word duration and the number of F0 changes in the word were among the prosodic attributes most correlated with word distance.

3. Are certain factors more correlated with insertions than with substitutions and deletions?

   Within-word text-based attributes, i.e., phone context and dictionary information, are useful at this intermediate level of prediction (i.e., transformation types). While higher-level attributes are useful for this task in general, phone location in the word, i.e., whether at the start of a word or not, is a major factor for differentiating phone transformation types, especially for reduction and deletion. Phones at the start of a word are much less likely to be reduced or deleted. Insertions are more likely to happen at the beginning of a syllable and this feature is used at a high-level in the decision trees when predicting insertions. Part-of-speech and word category labels also have striking differences between the amounts of deletion and reduction, with accented function words more likely to vary from the baseform than other types of categories. (See Table 5.5.) Prosody attributes were typically calculated at the word-level rather than at the phone-level to reduce reliance on word identity. While this makes it harder to say if prosodic attributes are more useful for particular types of transformations, we see in Table 5.12 that the use prosodic attributes reduces the overall perplexity when predicting transformation types, with duration measures giving the largest reduction over the text-based case (2.34 to 2.26).

4. Are certain factors more useful for predicting word-level variation or for predicting phone-level changes?

   For the intermediate prediction stage, both word-level (word distance) and phone-level (transformation type) use word context information along with current word information for improved prediction. The intermediate predictors used different attributes for prediction than phone-level surface form prediction. However, all classes of attributes examined, whether text-based or prosodic, were useful for both word-level and phone-level variation.

5. Is an intermediate variable (e.g., a reduction indicator) useful for predicting surface form realizations?

   Intermediate predictors give a small gain when included in surface form prediction. Using the text- and prosody-based transformation type predictor helped more than using the text-only transformation type predictor and resulted in the best phone prediction error rate. Using the known intermediate values rather than predicted values has a large impact on transformation type and surface form prediction accuracy and suggests further exploration in predicting intermediate variables.

6. Are variables reflecting syntax and discourse correlated with prosodic cues to the point of not adding any additional information or are they both useful in combination?

   While some of the text-based variables may be correlated with prosodic cues, experimental results show that the combination of the two types of variables (word-based and prosodic) helps. As seen in Table 5.13 (1) and 5.14 (5), neither help alone but Table 5.14 (6) shows that together, they do help improve the error rates for surface form pronunciation prediction.

Chapter 6

COMBINING PRONUNCIATION MODELS WITH ASR

Since improving phone prediction does not guarantee reduced word error rate, this work evaluates the effects of including our phone-based pronunciation models in an ASR system. The analyses performed in Chapter 5 used a small hand-labeled subset of Switchboard for training the pronunciation models. Before incorporating models into an ASR system, we examine the effects of training pronunciation models using a larger set of automatically aligned data. There are three key questions addressed in this chapter:

1. What effect does a larger training set have on the effectiveness of the pronunciation models and the likely word-level pronunciation strings?

2. Do improvements in static pronunciation models (e.g., adding part-of-speech information) lead to ASR improvements?

3. Do the recognition results improve by implementing dynamic pronunciation models for ASR?

A secondary issue will also be examined. We introduce a new ASR framework that includes a penalty term for hypothesis-dependent features. Since the prosodic attributes used in Chapter 5 were generated using word boundary (and sometimes word identity) information, incorporating a dynamic PM based on these hypothesis-dependent features will require compensation. We derive a method for compensation, but do not present experimental results within this framework since the surface-form prediction experiments with prosody did not show significant performance gains.

In summary, this chapter presents the mathematical framework for combining pronunciation and acoustic models, the methods for building the pronunciation models and performing

recognition experiments, and our recognition results using static and dynamic pronunciation models compared to single and multiple pronunciation baselines. We first present this for static pronunciation models, then dynamic pronunciation models and conclude with a summary of our work implementing and evaluating pronunciation models in ASR.

6.1 Static Pronunciation Models

Pronunciation strings can be encoded in a static dictionary and used to constrain the pattern search of the acoustic space. These dictionaries can be as simple as a single string of phones per word as in most ASR systems, or phones enhanced with information about syllables and stress [87]. The dictionary can be deterministic or it can include probabilities of different pronunciations for a given word. When multiple pronunciations are then included, the dictionary becomes a pronunciation model (PM). This section presents the implementation and experimental results for three forms of PMs used in this work: the base dictionary (most words have a single pronunciation), a static baseline (representative of prior work), and a static extension that incorporates word-level information. The following portions of this section will describe how static PMs are incorporated into an ASR system in theory, followed by a description of the training methodology and comparison of perplexity results for the various pronunciation models, followed by a description of implementation details of the recognition process and recognition results using the models. Results will be presented for pronunciation models trained on the ICSI set (phone-level training) and on the 80 hour set (word-level training).

6.1.1 Review of Theory

Returning to our equation that splits up the original acoustic model into a pronunciation model $p(\phi|W)$ and a subword unit model $p(x|\phi)$,

$$p(\mathbf{x}|W) = \sum_{\phi} p(\phi|W)p(\mathbf{x}|\phi), \tag{6.1}$$

the pronunciation model for $W = w_1, \ldots, w_n$ can be represented in different ways. We will describe the pronunciation of word i as $\underline{\phi}_i$, or

$$\underline{\phi}_i = \phi_{i,1}, \ldots, \phi_{i,l(i)}$$

where $l(i)$ is the number of phones in the word.

We can describe the **static baseline** pronunciation model in terms of the local context

$$p(\underline{\phi}_i|W) = p(\underline{\phi}_i|w_i), \tag{6.2}$$

reducing the effects of the entire word string on the pronunciation of word i to the current word. For decision tree models, this equation can further be expanded to

$$p(\underline{\phi}_i|w_i) = p(\underline{\phi}_i|b_{i,1}, \ldots, b_{i,k(i)}) \tag{6.3}$$

where $\underline{b}_i$ is the baseform pronunciation of word i, or

$$\underline{b}_i = b_{i,1}, \ldots, b_{i,k(i)}$$

and $k(i)$ is the length of the phone string in the dictionary for word i. At this point, we are just looking at the pronunciation of an individual word. Assuming $\phi_{i,j}$ are conditionally independent given $\underline{b}_i$, the equation now reduces to

$$p(\underline{\phi}_i|w_i) = \prod_{j=1}^{l(i)} p(\phi_{i,j}|\underline{b}_i). \tag{6.4}$$

We can further reduce the context of the baseform pronunciation to the neighboring phones with a window of $\pm m$ and get

$$p(\underline{\phi}_i|w_i) = \prod_{j=1}^{l(i)} p(\phi_{i,j}|b_{j-m}^{j+m}) \tag{6.5}$$

dropping the i index on b for brevity. While also conditioning the model on $\phi_{i,j-1}$ is proposed and implemented elsewhere with a slight improvement [78], it is left for future work in this framework since it significantly increases implementation complexity and the focus of this work is on the incorporation of higher-level information. In addition, the sequential dependence is incorporated when we move from phone-based models (generated

by the decision tree) to word-based models (generated by forced alignments of a larger training set). In other words, performing word-based pruning gives a model of the form in Equation 6.2, without the conditional independence assumptions of Equation 6.5.

In training, the mapping between $\underline{\phi}$ and $\underline{b}$ is aligned using dynamic programming, and $\underline{\phi}_i$ is redefined so that some $\phi_{i,j}$ may be insertions or deletions and $l(i) = k(i)$. Insertions are rare and are more complex to model, so only deletions will be allowed in the recognized surface-form phone sequence. In this case, each baseform symbol is associated with exactly one surface form symbol (including deletions). Note that this model is described for phonemes, but it can be extended to work for coded phones, i.e., phones described in terms of an articulatory feature vector. The extension to coded phones will be addressed in Chapter 7.

Because of the original assumption (Equation 6.2) that $\underline{\phi}_i$ depends on the word, we can now include word-level cues such as part-of-speech information in this model. Equation 6.5 is expanded to represent our **extended static model** with local word information as follows:

$$p(\underline{\phi}_i|W) = \prod_{j=1}^{l} p(\phi_{i,j}|b_{j-m}^{j+m}, w_i) \tag{6.6}$$

where w_i can be any available information about the current word.

6.1.2 Training Pronunciation Models

Our approach to training the pronunciation models is the same as previous work at JHU and WS '97 [7, 78]. Decision-tree-based pronunciation models are initially built using the 3.5 hour ICSI training set as described in Chapter 5. Phones that are hand-labeled (e.g., /ix/, /ux/, /q/, and /em/) but are not in the acoustic model phone set are mapped to the closest phone in the set for recognition purposes using the articulatory feature distance (in Appendix A.3). The trees are pruned using 10-fold cross validation as well as thresholds on the output probabilities and are limited in their growth based on minimum sample counts at the nodes. The trees are more aggressively pruned than the trees built in Chapter 5, but part of the goal there was to see which types of attributes were chosen and whether low-probability instances of production, such as insertions in the transformation type prediction, could be predicted at all.

While recognition can be done using these pronunciation models, previous work at JHU and WS '97 showed that it was important to build pronunciation models on more data than the ICSI training set alone to reduce the effects of outlier pronunciations. Thus, we use expanded dictionaries generated using these heavily pruned (ICSI-based) PMs to perform Viterbi alignment on an 80 hour training data set which is then used to re-estimate the costs of the pronunciations generated by the original pronunciation model. The acoustic model used for this alignment is based on 143 hours of training data. It has a five-state model topology, uses mean normalized MFCCs and derivatives, shared covariances and clustering based on information about syllable and stress.[1]

After alignment, two options are available for incorporating information from the larger training set: phone-level and word-level probability training. (At this point, we no longer use the hand-labeled data.) In the phone-level option, the output phone alignments of the 80 hour PM training set are considered as surface-form phones. We then align these automatically generated surface forms with the associated dictionary baseforms using weighted FSMs and the pronunciation distance matrix as transformation costs, as for the original ICSI-to-baseform alignment. New decision trees can then be generated to create a new pronunciation model. The word-level option involves using counts of the pronunciations represented in these alignments to generate relative-frequency-based costs for the dictionary that replace the costs generated by the original decision tree. The second option is used here since it avoids potential problems with overfitting the pronunciations to the training data. Additionally, pronunciation strings generated by the ICSI-based trees that are never seen can be culled from the resulting dictionary, even though they may have had a reasonable cost using the original transformation probabilities. The new PMs created with a larger training set have two advantages: 1) better statistical modeling with the larger training set and 2) closer matches between the surface form output and the acoustic models used for recognition. However, they do not allow for modeling pronunciations of words unseen in training.

[1]This acoustic model used for forced alignment of the training data is different from the acoustic model used in recognition experiments, because the signal processing used for training the models for recognition is based on proprietary software that is no longer available.

As described in Chapter 5, perplexity is used to evaluate the phone-based pronunciation models prior to recognition.[2] Perplexity will be calculated on the ICSI Held Out set in order to have a comparison between the two sets of training data. Perplexity is calculated as described in Chapter 5, but the perplexity results are a little different here than those reported in Chapter 5, for three main reasons. First, the hand-labeled set included phones that are not part of the recognition phone set as mentioned above. These phones were used in the Chapter 5 models but not here. The test set tokens are mapped to the closest phone in the acoustic model set for the perplexity calculations. Also, the different stopping criteria (i.e., limits to the size of a data set that can be split at a node) and pruning approaches result in different trees. Finally, insertions are not included in the prediction trees or in the perplexity measures here, but they were in Chapter 5.

In the static models, attributes included are: manner and placement information of the baseform phone context for a 7 phone window (± 3 phones) up to word boundaries; whether the word is an interjection or not (based on whether the phone is a filled pause phone); whether the current and left or right phone is stressed; and the distance of the left and right word boundaries. The models are trained on the ICSI set (3.5 hours) and costs are retrained based on word-level frequency on the 80-hour set. We present experiments using expanded pronunciations for all words in the dictionary and for a 150 word subset. The 150 words were the most frequent words, selected based on unigram word frequencies from a language model trained on Switchboard and Callhome data.

The extended static model was designed using local part-of-speech (POS) and word category (WordTag) information as additional attributes. Word categories are defined by the word identity alone so there is only one category per word. Words may be assigned multiple part-of-speech categories. The POS labels for words in the dictionary are chosen based on frequencies of POS labels in the Switchboard training set, and are clustered into groups as in Chapter 5. As with the static PM, the dictionary is expanded with alternate pronunciations generated by the PM tree trained on the ICSI data. When multiple categories are assigned

[2]Pruned word-level pronunciation costs do not cover all variations in the hand-labeled set, so perplexity is not a useful tool for evaluating these.

to a word, the categories that occur more often than a threshold[3] are *all* used to tag the word. Approximately 775 out of 30k words have multiple POS tags. These are treated as multiple pronunciations when expanding the dictionary since the POS tag is an attribute for every phone in the baseform pronunciation string. POS information is not preserved after generating the pronunciation string and the initial pronunciation cost.

Table 6.1 summarizes the perplexity results for comparing versions of the static and extended static PMs. The perplexity of the models is measured on the hand-labeled ICSI held out set for comparison purposes. Perplexity numbers are presented with and without filled pause phones because there is labeling mismatch between filled pause words in the training set (typically a "filled pause" phone) and the test set (typically a combination of /ah/ and /hh/ or /m/). Incorporating word-level information reduces the perplexity but, unlike results in Chapter 5, there is not a win from including both word category and part-of-speech information.[4] As seen in Chapter 5, part-of-speech and word category attributes are still used and are not pruned away though more aggressive pruning is used here. The average nodes per phone subtree for trees pruned with the same criteria increases slightly as more information is available via the attributes. The POS and word category attributes are used in many of the trees, with POS being used for 30 phone trees (out of 46) by itself and 28 when word category is used. Word category is used in 18 phone trees but only 10 when it is included with POS information.

The baseline recognition dictionary is the "single" pronunciation dictionary. A 34k word dictionary was available for the first pass of recognition with 1.07 pronunciations per word. For the second pass of recognition, the word lists for the two test sets are used to extract words from this dictionary. The dictionary subsets also have an average of 1.07 pronunciations per word.

The pronunciations per word for dictionaries generated by the static PM and the extended static PM (static + part-of-speech information) vary widely. They decrease consid-

[3]For words with more than 3 tags, if the percentage of POS tags seen in the training data is greater than 1 over the total number of tags for that word, the tag is included.

[4]For calculating held out set perplexity on the POS case, the true POS tag is used. This gives a somewhat optimistic perplexity value.

Table 6.1: *Perplexity of static and extended static pronunciation models built using word categories and/or part-of-speech information. (Results are presented without the filled pause phone.)*

DT Attributes	Average Nodes per Tree	Perplexity	
		Train	Held Out Set
Static	5.98	3.68	4.77
+ WordTag Category	6.00	3.66	4.75
+ POS Category	6.13	3.68	4.75
+ both	6.15	3.66	4.75

erably when constrained by the 80-hour test set as shown in Table 6.2. In the recognition experiments described below, two results for the ICSI-set and 80-hour models are presented. For the ICSI-set model, the first has up to ten pronunciations per word with an average of approximately 6.2. The second has up to ten pronunciations per word for the 150 most common words (based on unigram language model probability) with an overall average of 1.09. For the 80-hour set, a word-level pronunciations is pruned if there are less than 20 occurrences of that string actually seen in the training data. If there are less than 20 occurrences of all pronunciations of a word in the training data, the pronunciation string reverts to the canonical string with a cost of 0. After pruning, the costs for each string are adjusted based on frequency of occurrence.

Using the POS-based PM reduces the number of pronunciations in the final ICSI-based dictionary. The less-likely pronunciation strings generated by this PM have a higher cost and are not within the probability cutoff range.

6.1.3 Recognition Implementation

For recognition, this work is implemented in a multi-pass framework. First-pass results require a great deal of computer resources, so recognized word lattices expanded into pronunciation networks using the pronunciation model are used in a second-pass framework

Table 6.2: *Average pronunciations per word for static and extended static dictionaries generated with either the ICSI training set or the 80-hour training set. Dictionary wordlist is that used for the Eval '00 test set. The canonical dictionary has 1.07 pronunciations per word.*

	Pronunciations per word					
	Static		+ POS category		+ WordTag	
Conditions	Full	150 word	Full	150 word	Full	150 word
ICSI set	6.18	1.09	6.10	1.09	6.21	1.09
80-hour set	1.09	1.08	1.09	1.08	1.09	1.08

to make decoding time reasonable and to reduce memory requirements. Useful information such as the class of the hypothesized word after the target word can be handled in this framework by using information from the first pass output rather than modifying the recognizer.

First pass recognition results were generated using a three-state HMM triphone system, with diagonal covariance Gaussian mixtures and vocal tract normalized MFCCs, derivatives and delta-derivatives, as described in Chapter 4. Trigram language model scores are generated with the SRI language modeling toolkit [99].

The implementation of all static models in the second pass lattice rescoring requires dictionary expansion and pruning. For decision tree models trained on the ICSI data, each word in the sub-vocabulary dictionary is processed by the prediction trees to generate alternate pronunciations. The most likely pronunciation is normalized to have a cost of 0 by subtracting the value of the cost (negative log probability) associated with that pronunciation. This means there is no penalty for the single most likely pronunciation, while less likely transformations add higher costs to the path. All of the other pronunciations also have that cost subtracted from their costs as a scaling factor. The pronunciation cost is then given the same weight as the language model scores. Individual phone transformations with a probability below a given threshold ($p < 0.1$) are discarded, as are pronunciation strings with $p < 0.1$ (except where indicated). The set of generated pronunciation strings for a

given word is further pruned to result in no more than 10 possible strings per word. The pruning is done primarily to keep the memory requirements reasonable during decoding, but also to exclude unlikely possibilities. For word-level models trained on the 80 hour set, the expanded dictionary is converted to a finite state machine (FSM) using AT&T tools [61] and then composed with the first-pass word lattices during decoding. A lattice-based search for the most likely word string is performed. The POS-based extended static model is like a knowledge-driven mixture, since POS tags are not being used explicitly in recognition.

Recall that clustering of the acoustic model distributions includes questions about syllable and stress information as well as phonetic context, so some information used in our pronunciation models is also used in the acoustic models. The phones of the baseline dictionary are thus enhanced with syllable and stress information to match the acoustic model units. A mapping is performed between untagged predicted surface-form phones from the pronunciation model and the phones used in the dictionary. This creates some mismatches since a predicted phone transformation from a stressed vowel to a reduced vowel, for example, cannot simply have the same tag as the baseform phone and still match the available acoustic models. Here, tags appropriate for the predicted phone are used.

A second pass of scoring is done on the lattices generated in the first pass using the same acoustic and language models, now including the static pronunciation models built using either the ICSI set or the 80-hour set.

6.1.4 Results

Recognition performance is evaluated using percentage word error which is found by dividing the number of word errors by the total number of words in the correct transcriptions. The number of word errors is a sum of the number of substitutions, insertions and deletions. The test sets used here are Dev '98 and Eval '00 as described in Chapter 4. Dev '98 and Eval '00 are evaluated first with the canonical dictionary, and then dictionaries expanded using the static and extended static pronunciation models. The dictionaries have expanded pronunciation options for either all words or the 150-word subset, with and without pruning pronunciations based on total cost. If word-level pronunciation pruning is not used, the

average number of pronunciations per word increases slightly for the 150-word set (from 1.09 to 1.10) but increases substantially for the full vocabulary (from 6.18 to 8.84). Results are also presented for dictionaries with the pronunciation costs trained with the 80-hour set. As stated earlier, the maximum number of pronunciations per word is ten.

Results for the static case are given in Table 6.3. Results using the ICSI-trained models alone are worse than the canonical dictionary baseline, but retraining with a larger data set yields a statistically significant improvement over the baseline. The poor performance using the ICSI-trained models is most likely due to a few issues: confusability introduced by the large number of available pronunciations, poorly trained pronunciation costs, and phone-level (vs. word-level) models.

The results are mixed with respect to the problem of confusability. On one hand, performance is improved when when the expanded pronunciations are limited to 150 words, for both training sets. On the other hand, when no word-level cost-based pruning is used with the ICSI-trained models and the number of available pronunciation strings increases (substantially for the full vocabulary), the results are not significantly different, but possibly better. However, we did not investigate reducing cost-based pruning in further experiments because of the high implementation costs associated with a much larger dictionary. The performance differences between training with the 80-hour set and the 3.5 hour ICSI set reflect both differences in data and word-level vs. phone-level costs so it is difficult to separate these factors. In summary, while the pronunciation models trained on a small corpus hurt performance, there is an overall gain of 0.8% absolute word error rate with retraining on the 80 hour set. This difference is statistically significant on both test sets.

Table 6.4 summarizes the recognition results using the extended static pronunciation models with cost-based pruning. When a small amount of training data is available, as with the ICSI set, it is useful to incorporate higher-level information about the word. This effect disappears when more data is available, as with the 80-hour case. Comparing results between POS tags and word category information show mixed results, with none of the differences being significant. While we do not see a significant improvement incorporating POS information, it should be noted that POS is not actually being used in recognition, only in pronunciation model training. Including POS explicitly, e.g., via separate lexical entries

Table 6.3: *Baseline recognition results.*

PM Training Set	Probability Pruning	Expanded Words	Test Set WER	
			Dev '98	Eval '00
Canonical Dictionary	–	–	48.4	38.8
ICSI Static PM	Yes	All	49.9	39.5
	No	All	49.8	NA
	Yes	150	49.6	39.3
	No	150	49.6	39.3
80-hour Static PM	Yes	All	47.8	38.2
	Yes	150	47.6	38.0
Oracle best-path	–	–	10.8	9.9

Table 6.4: *Recognition results (word error rate) using extended static PMs which include part-of-speech category. Training data set is used to describe the pronunciation models.*

Extended with:		Part of Speech		Word Category	
PM Training Set	Expanded Words	Test Set		Test Set	
		Dev '98	Eval '00	Dev '98	Eval '00
ICSI Data	All	49.8	39.5	49.7	39.6
	150	49.4	39.1	49.5	39.1
80-hour Set	All	47.7	38.3	47.8	38.2
	150	47.5	38.1	47.6	38.0

in the language model may be a way to incorporate some of the benefits seen in Chapter 5. This will be explored to some extent in the next section, where the implementation of dynamic PMs will incorporate text-based information in the lattices.

6.2 Dynamic Pronunciation Models

The next step in pronunciation modeling is to include a broader context. While some word-level information can be included in static models, the surrounding word-context also has an effect on pronunciation variation. This section first develops the theory of dynamic pronunciation modeling, then describes how dynamic PMs are trained, the implementation of dynamic PMs into an ASR system, and experimental results.

6.2.1 Theory

Equation 6.5,

$$p(\underline{\phi}_i|w_i) = \prod_{j=1}^{l(i)} p(\phi_{i,j}|b_{j-m}^{j+m}),$$

expands further for **dynamic models**. The two cases we look at include 1) information about the words in the utterance; and 2) information about the words and prosodic attributes within the utterance.

For the first case, the pronunciation model $p(\underline{\phi}_i|W)$ can be described as

$$p(\underline{\phi}_i|W) = \prod_{j=1}^{l(i)} p(\phi_{i,j}|b_{j-m}^{j+m}, W). \tag{6.7}$$

Dynamic pronunciation models require information about the context of the word rather than the current word alone so are implemented during recognition by expanding the network of possibilities during a second pass of recognition. A new issue that arises here is that lattice rescoring is not an option because utterance-level information is incorporated. When utterance-level text variables are used, the word sequence for the complete utterance needs to be available. This would come from an N-best list. When information from the word and its left and right neighbors are used, we can use the trigram lattice:

$$p(\underline{\phi}_i|W) = \prod_{j=1}^{l} p(\phi_{i,j}|b_{j-m}^{j+m}, w_{i-1}, w_i, w_{i+1}). \tag{6.8}$$

In either case, a dynamic decision-tree-based pronunciation model requires that the composition with the pronunciation model occur after the canonical dictionary and word lattice have been composed rather than having one expanded dictionary available.

Many of the prosodic cues used in pronunciation prediction are dependent on the hypothesized word sequence, whether the word identity or the word boundaries alone are used for the calculation of acoustic correlates of prosody. In Chapter 5, the prosodic cues used for building predictive trees were all constructed using word boundaries. Although some attributes used time windows, at least one end of the window was set at a boundary. This dependency of the cues on word-based information means that independence assumptions typically used to tease apart the aspects of word strings for modeling will not hold when using prosody.

This is addressed in the following manner. When including prosody in the model of speech, the initial recognition equation (Equation 2.2) is expanded to

$$\underset{W}{\operatorname{argmax}}\, p(W|\mathbf{x}, Y) = \underset{W}{\operatorname{argmax}}\, \frac{p(\mathbf{x}, Y|W)p(W)}{p(\mathbf{x}, Y)} \tag{6.9}$$

where Y represents a lattice of prosodic attributes y_i when the prosodic attributes are calculated using word boundary and or word identity information. We can ignore p(**x**,Y) if Y describes the entire lattice (as opposed to hypothesis-dependent features $Y(W)$), resulting in

$$\underset{W}{\operatorname{argmax}}\, p(W|\mathbf{x}, Y) = \underset{W}{\operatorname{argmax}}\, p(\mathbf{x}|W, Y)p(Y|W)p(W). \tag{6.10}$$

We partition Y into the sequence $Y(W)$ corresponding to hypothesis W and $Y(\bar{W})$ for all other y_i. This distinction is important since different word hypotheses give different values for the prosodic attributes. Since $Y(W)$ depends on W,

$$\underset{W}{\operatorname{argmax}}\, p(W|\mathbf{x}, Y(W)) = \underset{W}{\operatorname{argmax}}\, \frac{p(\mathbf{x}, Y(W)|W)p(W)}{p(\mathbf{x}, Y(W))} \neq \underset{W}{\operatorname{argmax}}\, p(\mathbf{x}, Y(W)|W)p(W). \tag{6.11}$$

The model of the prosody given the words, assuming conditional independence of y_i, is

$$p(Y|W) = p(Y(W)|C)p(Y(\bar{W})|E),$$

where C indicates the word hypothesis is correct and E indicates the word hypothesis is

incorrect. The prosody-dependent acoustic model is rewritten as

$$p(\mathbf{x}|W,Y) = \sum_{\phi} p(\mathbf{x}|\phi,W,Y)p(\phi|W,Y) = \sum_{\phi} p(\mathbf{x}|\phi)p(\phi|W,Y(W)).$$

Assuming that $\mathbf{x}$ is conditionally independent of Y and W given ϕ, and that ϕ is conditionally independent of $Y(\bar{W})$ given W and $Y(W)$, replacing the expanded models into Equation 6.10 yields

$$\begin{aligned}\underset{W}{\operatorname{argmax}}\, p(W|\mathbf{x},Y) &= \underset{W}{\operatorname{argmax}} \left(\sum_{\phi} p(\mathbf{x}|\phi)p(\phi|W,Y(W))\right) p(Y(W)|C)p(Y(\bar{W})|E)p(W) \\ &= \underset{W}{\operatorname{argmax}} \left(\sum_{\phi} p(\mathbf{x}|\phi)p(\phi|W,Y(W))\right) p(Y(W)|C)p(Y(\bar{W})|E)p(W)\frac{1}{p(Y|E)} \end{aligned} \tag{6.12}$$

since $P(Y|E)$ is constant for all W. Using

$$p(Y|E) = p(Y(W)|E)p(Y(\bar{W})|E),$$

we then simplify the equation to

$$\underset{W}{\operatorname{argmax}}\, p(W|\mathbf{x},Y) = \underset{W}{\operatorname{argmax}} \left(\sum_{\phi} p(\mathbf{x}|\phi)p(\phi|W,Y(W))\right) \frac{p(Y(W)|C)}{p(Y(W)|E)}p(W) \tag{6.13}$$

$$\approx \underset{W,\phi}{\operatorname{argmax}} \left(\max_{\phi} p(\mathbf{x}|\phi)p(\phi|W,Y(W))\right) \frac{p(Y(W)|C)}{p(Y(W)|E)}p(W). \tag{6.14}$$

The likelihood ratio, $\frac{p(Y(W)|C)}{p(Y(W)|E)}$, serves as a penalty function that gives a lower score for features that are computed at errorful word hypotheses. There are also more complex penalties we could use such as the near-miss model suggested in [29]. Such a model could be useful because our assumption that the values associated with C and E are independent is clearly invalid, since both sets are calculated on the same stream of acoustic attributes (such as F0). The likelihood ratio can be trained using the ICSI Held Out set. Utterances are first recognized and then aligned with the reference text to get the C and E values. The underlying distributions of the individual prosodic attributes depends on the type of values representing the attribute. Continuous variables, such as F0 ranges, may be represented by Gaussian distributions while integer values, such as the number of F0 slope changes, may be represented by a Poisson distribution.

The pronunciation model is $p(\phi|W, Y(W))$ and the alternative to Equation 6.6 is

$$p(\underline{\phi}_i|w_i, y_i) = \prod_{j=1}^{l} p(\phi_{i,j}|b_{j-m}^{j+m}, w_i, y_i) \tag{6.15}$$

where y_i represents the prosodic attributes calculated with respect to hypothesized word w_i. Assuming conditional independence of y_i given a correct word segmentation, then

$$p(Y(W)|C) = \prod_i p(y_i|C).$$

The context of the pronunciation model can be broadened easily. For example, neighboring word or prosodic context can be added:

$$p(\underline{\phi}_i|W, y_i) = \prod_{j=1}^{l} p(\phi_{i,j}|b_{j-m}^{j+m}, \phi_{i,j-1}, w_{i-1}, w_i, w_{i+1}, y_{i-1}, y_i, y_{i+1}). \tag{6.16}$$

6.2.2 Training Dynamic Pronunciation Models

The dynamic pronunciation models presented here include the attributes in the baseline and static pronunciation models as well as text-based attributes discussed in Chapter 5. Perplexity results are summarized in Table 6.5. The addition of broader context for POS or word category information reduces perplexity from 4.75 for the extended static model to 4.57 for the best case dynamic model. This is a significant reduction in perplexity given the significance results from Chapter 5. While the pronunciation model built with all available information had similar perplexities as POS window alone, the average number of nodes per tree is somewhat reduced.

6.2.3 Recognition Implementation

The framework for performing recognition using dynamic pronunciation models is necessarily different from the method used for static PMs. In order to dynamically incorporate information from word context, we use an N-best rescoring framework. Given the word lattices generated in the first pass of recognition, the canonical dictionary without expanded pronunciations, and phone-based decision trees using word-level context, the algorithm for recognition using dynamic PMs is as follows:

Table 6.5: *Perplexity of dynamic pronunciation models using higher-level information built on different size training sets.*

DT Attributes	Average Nodes per Tree	Perplexity	
		Train Set	Held Out Set
Static	5.98	3.68	4.77
+ POS window	4.98	3.53	4.58
+ WordTag window	5.84	3.55	4.57
+ Location in Utterance	5.60	3.60	4.67
+ All	4.80	3.54	4.58

1. Generate word-level N-best in FSM form.

2. Compose the N-best hypotheses with the canonical dictionary to generate N phone-level lattices (with output word labels and LM scores preserved), where lattices are needed because of multiple pronunciations in the canonical dictionary.

3. Generate the dynamic context for each word, i.e., baseform phone, POS label (from [76]), word category, neighboring phone and word context information. (Prosodic information can be included here as well).

4. Predict the surface form distribution for each phone in the phone-level lattice using the decision trees to generate phone-level costs. Pruning of low likelihood transformations happens at this stage, where phone transformations with $p < 0.05$ are excluded.

5. Generate a phone-level latticelet with associated costs for each phone in the lattice. These are then combined into an utterance-level phone lattice, and all of the N-best phone lattices are combined into a single compacted lattice.

6. Rescore the recombined utterance-level lattice. The final output is the best path through the utterance lattice, i.e., the single hypothesis with the lowest combined

acoustic, language and pronunciation model cost.

While it would be nice to implement Steps 3, 4 and 5 using the FSM framework applied to our other recognition experiments, it would require limiting the types of questions asked by the decision trees as well as requiring that all continuous attributes be quantized. Instead, calls to the tree prediction software are made for each baseform phone.

6.2.4 Results

Table 6.6 summarizes the results using text-based dynamic pronunciation models. Since we are limiting this task to an N-best ($N = 2000$), we also present results using the N-best possibilities for the cases of the ICSI-trained "All" model. For both the static and POS-extended static models, the error rate is lower than seen in Tables 6.3 and 6.4. The results are actually comparable to the 150-word case, suggesting that reducing the word possibilities by using an N-best list reduces the problem of confusability. We find that word error rate decreases when more high-level information is added, but overall results are not improved, probably because of differences in pruning associated with the phone-level implementation of the dynamic model. The static and extended static experiments use models with pronunciation strings pruned based on word-level costs and limited to a maximum of ten pronunciations per word. While limited experiments with cost-based pruning suggest this is not a factor, we believe the word-level maximum is. The dynamic models have a much higher maximum because all paths through the phone latticelets are allowed.

6.3 Summary

This chapter presented work incorporating three types of pronunciation models into an automatic speech recognition system. While further experimental work is warranted, we have addressed the introductory questions in the following manner.

1. What effect does a larger training set have on the effectiveness of the pronunciation models and the likely word-level pronunciation strings?

Table 6.6: *Recognition results (word error rate) using dynamic text-based PMs for rescoring the 2000-best hypotheses. The ICSI set is used to train all the models. Static and extended static results on the N-best are presented as baselines.*

Decision Tree Attributes	Test Set	
	Dev '98	Eval '00
Static (Phone-based)	49.4	39.0
POS-extended static	49.5	39.1
Dynamic (POS window)	49.9	40.1
Dynamic (POS window, WordTag window and utterance location)	49.9	39.7

Performing alignments of a larger training set is critical to the success of pronunciation modeling, primarily because of the reduction of the number of available pronunciation strings. Reducing the number of available pronunciations, whether only including expansions for a 150 word set or by using forced alignments to generate pronunciations has a positive effect on recognition results. Using word-level training information for the costs associated with pronunciation strings leads to significantly improved recognition results.

2. Do improvements in static pronunciation models (e.g., adding part-of-speech information) lead to ASR improvements?

   The results here do not show that extending static models to include part-of-speech tags improves recognition results when using a hidden variable (mixture) framework. We do see that when limited data is available, there is a small gain from including part-of-speech and word category information. However, it is no longer useful when other word-level information, in the form of aligned pronunciation strings, is available for re-training pronunciation costs. Of course, it may be that a word-level version that is combined with a POS-dependent language model will lead to improvements.

3. Do the recognition results improve by implementing dynamic pronunciation models for ASR?

   At this point, dynamic prediction of phone-level transformations does not improve recognition results. Although we see improvements within the dynamic framework when higher-level information is included, these results still do not meet the static model baselines. Training the decision trees on a larger set of data, such as automatically aligned baseform-surface form pairs from the 80-hour set, coupled with explicit modeling of phone sequence dependence (including pruning) may improve results since both more training data and word-level constraints improved recognition results using a static prediction framework. In addition, it would probably be better to implement the dynamic pronunciation model with word-level pronunciations, which would be relatively straightforward for frequent words but more difficult for infrequent words.

Chapter 7

ARTICULATORY FEATURES FOR PRONUNCIATION MODELING

This chapter addresses the question of whether articulatory features are useful for pronunciation modeling and whether they give improvements over phone-based modeling. Linguistic analysis now relies on feature-based changes to describe pronunciation variation (e.g., [51][1]). Incorporating these features in speech recognition may result in better modeling of the pronunciation variation. Here, we investigate the trade-offs between the use of these features vs. phones for modeling pronunciation variation. The potential advantages of modeling feature level changes are 1) the lower dimensional model that is shared across phones which can take advantage of the relatively small amount of training data available, 2) the ability to capture segment changes to non-canonical feature combinations, and 3) the ability to capture sub-phone (e.g., state-level) effects if implemented at a state level, which also allows for asynchronous feature changes.

In this study, the focus is on the first two issues: the advantages of low dimensionality and the ability to capture segment changes. We design decision trees to predict feature changes rather than phone changes. This allows us to take advantage of the fact that the same feature changes are shared by many different phones, thus reducing problems with sparse data and yielding more robust models. An example of this is the reduction of vowels when they are in unstressed syllables. We show that by using articulatory features we can reduce the number of parameters needed to model pronunciation variants. Additionally, because the hand-labeled ICSI set has diacritic notations attached to some phones, showing changes in such features as voicing or nasalization or to groups of features representing frication or approximation, we are able to model a wider variety of productions.

[1]Note that while features are typically used to describe variation, there is still debate about appropriate feature values. This work uses a different feature set than Ladefoged [51], although the approach is typical of the field.

Because a feature change does not always result in a different phone in our acoustic model set, we can map collections of features to phones in our set. We are, however, able to predict a broader range of possible articulations. This can result in a better representation of what is actually produced by a speaker. One example is the nasalization of the vowel /ow/ in the word "don't" when the /n/ and /t/ are deleted. When this change occurs, we are able to express the change in the feature representation of the /ow/, rather than simply having two deletions and an unchanged vowel. When this representation is used for both pronunciation and acoustic modeling, we retain the benefits of low dimensionality but also have no mismatch between predicted changes and the acoustic model.[2] The use of features allows the modeling of the coarticulation and allophony via modeling the distributed and temporally asynchronous nature of information in the signal. Additionally, by modeling feature transformations in both the acoustic and pronunciation model spaces, we may be better able to avoid catastrophic errors or recover from a degraded environment. If we can predict many of the correct features, we may be able to recognize the correct word in context, even if individual phone recognition rates are not high.

This work is tested through prediction experiments similar to those presented in Chapter 5 and addresses the following questions:

1. What intermediate dictionary representation (phones, features or feature groups) is most useful for predicting pronunciation changes?

2. Are high-level conversational attributes used more often in feature prediction than in phone prediction?

3. Do hierarchical dependencies between features offer significant advantages over independent feature prediction?

This chapter presents the mathematical framework for using features in pronunciation modeling, beginning with the theoretical framework for pronunciation modeling using features. Articulatory features are examined as independent features, in feature groups, and

[2]In this framework, sets of articulatory features can replace phones as symbolic labels not only in the pronunciation model but also for indexing the acoustic model.

with hierarchical dependencies. The experimental method is described, followed by an evaluation of the feature-based models. After presenting and discussing pronunciation modeling results, the framework for incorporating articulatory features into an automatic speech recognition system will be presented.

7.1 Theory: Features in Pronunciation Modeling

Beginning with Equation 6.6, but dropping the word index i for brevity and defining the phone window to be only the left and right context,

$$p(\underline{\phi}|W) = \prod_{j=1}^{l} p(\phi_j|b_{j-1}^{j+1}, w). \tag{7.1}$$

This model can be expanded to work for coded phones, i.e., phones described in terms of a feature vector. When we use feature vectors instead of phones, Equation 7.1 expands to

$$p(\underline{\phi}|W) = \prod_{j=1}^{l} \prod_{k=1}^{N_f} p(f_{j,k}|b_{j-1}^{j+1}, w) \tag{7.2}$$

where N_f is the number of features in the articulatory feature vector, $f_{j,k}$ is the value of the k^{th} feature for phone j, and b represents a coded version of the phones in $\underline{b}$. In Equation 7.2, the individual features at a given time are assumed to be conditionally independent given the baseform and the previous value of the feature being examined. This independence assumption is not necessarily valid, so we will also examine bundles of features, using independent subsets and hierarchical groupings of the features. Assuming that there is dependence within groups of features but that groups are conditionally independent of each other, then Equation 7.2 becomes

$$p(\underline{\phi}|W) = \prod_{j=1}^{l} \prod_{k=1}^{N_s} p(s_{j,k}|b_{j-1}^{j+1}, w) \tag{7.3}$$

where N_s is the number of feature subsets, s represents the feature sets, and b represents the feature subsets of and/or phone information about the baseform phone string $\underline{b}$. Here, the baseform can represented as a function of the feature clusters. When hierarchical groupings of features are used, a dependency on the parent nodes of the feature being examined is

included. The resulting equation for individual features is

$$p(\underline{\phi}|W) = \prod_{j=1}^{l} \prod_{k=1}^{N_f} p(f_{j,k}|F(b_{j-1}^{j+1}), f_{j,\pi(k)}, w), \tag{7.4}$$

or, for sets of features

$$p(\underline{\phi}|W) = \prod_{j=1}^{l} \prod_{k=1}^{N_s} p(s_{j,k}|F(b_{j-1}^{j+1}), s_{j,\pi(k)}, w), \tag{7.5}$$

where $\pi(k)$ are the parents of the feature set in a dependence tree. The parents of a given feature for the features used here are motivated by the shape of the articulatory feature tree in Figure 3.2, though we use a somewhat different topology because of the reduced set of features used here. This work, in part, examines whether the dependencies described in phonological trees can translate into conditional dependencies in a probabilistic model since they are not equivalent. In this work, we examine feature sets with dependencies defined in the next section. Additionally, because of the original assumption that $\underline{\phi}$ depends on the entire word string, we can include word-level cues such as part of speech and lexical stress in this model.

7.2 Method: Building and Evaluating Feature-based Models

7.2.1 Data and Experiment Plan

The hand-labeled data used for prediction work in Chapter 5 is also used here. However, because we are now modeling feature level changes, we incorporate the diacritic notations available with the ICSI hand-labeled set for some surface-form phones that map to feature changes within the phone. The baseform and surface-form phones (with diacritics) are converted to articulatory feature sets as described in Appendix A.2, which includes a description of how diacritics are mapped to feature changes (Tables A.6 and A.8) as well as the distribution of diacritics in the data (Table A.7). The features can then be used as attributes in decision trees along with the phone context, word context and prosody attributes used earlier. Three types of feature groupings will be examined compared to the baseline of phone-based modeling which can be thought of as a single group with all features. Thus,

results from Chapter 5 will be used for comparisons. In addition, we compare the different assumptions in Equations 7.2, 7.3 and 7.5.

When the features are assumed to be modeled with independent groups, i.e., no dependencies across groups, we examine three groups: individual features (21 features) and two articulator-based groupings (8 and 9 groups). For individual feature modeling, 21 decision trees are built.[3] For easier comparisons with phone-based modeling, attributes based on the phonetic context are used in decision trees rather than attributes matching individual features, though these are effectively equivalent. Baseform feature values are used as attributes in the tree, effectively given a top-level split between the most likely outputs for that feature. For the articulator-based case, the features are grouped if they have the same articulator and/or are at the same level of the dependence tree. Since the articulatory features are part of a single production system, along with assuming dependency within groups, it also makes sense that there would be dependencies across groups of the features. Thus, the third case is a hierarchical grouping, generating 9 groups, where baseform and surface-form information about the parent of the feature can be included in the set of attributes that the tree can use. The groupings are shown in Table 7.1, and the dependencies for the third case are shown in Figure 7.1. The topology of the dependence tree reflects the fact that some nodes in the Figure 3.2 tree are grouped (vowel and glide, lingual and lips, glottis and pharynx). In addition, our feature set includes delayed release which allows affricates to behave like both stops and fricatives. The one real departure from the topology in Figure 3.2 is that the tongue body feature is attached to sonorant/continuant/strident node rather than the lingual (dorsal/coronal/labial) node. This was kept separate primarily because the high, low, and back features are used somewhat differently for vowels than for glides and consonants.

In this chapter, decision trees are built for the three cases using different sets of text-based and prosodic attributes. The groups of text-based and prosodic attributes are those that generally yielded the best results for phone-based prediction in Chapter 5. Specifically,

[3]The two glottis features (spread and constricted) are merged into a single category as in the feature distance. The feature round is excluded in this model since, in American English, it is either redundant with the labial feature or not used as a distinguishing feature.

Table 7.1: *Feature groupings for pronunciation prediction experiments.*

Experiment	Features in Groups
Phone-based	All features
Independent Features	21 individual features (as in Appendix A.2) with a trinary feature for spread vs. constricted glottis and excluding round
Grouping I (articulator-based) (8 groups)	1. syllabic, consonantal, sonorant, continuant, strident, and delayed release 2. voicing 3. glottis (trinary feature for spread vs. constricted) 4. advanced tongue root, constricted tongue root 5. nasal 6. dorsal, coronal, labial 7. high, low, back 8. anterior, distributed, lateral, rhotic
Grouping II (Hierarchical articulator-based) (9 groups)	1. syllabic, consonantal (vowel, consonant, glide) 2. sonorant, continuant, strident *given (1)* 3. delayed release *given (1)* 4. voicing *given (1)* 5. glottis (spread vs. constricted), advanced tongue root, and constricted tongue root *given (1)* 6. nasal *given (1)* 7. dorsal, coronal, labial *given (1) and/or (2)* 8. high, low, back *given (1) and/or (2)* 9. anterior, distributed, lateral, rhotic *given (7)* *and/or (1) and/or (2)*

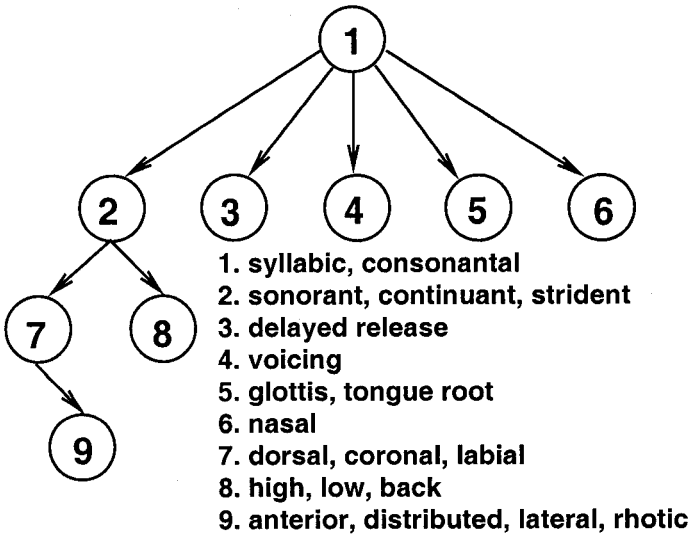


Figure 7.1: Dependence tree for hierarchical feature groups presented in Table 7.1.

the attributes made available to the trees are shown in Table 7.2. In some cases, attributes not selected in Chapter 5 were chosen here and some chosen in Chapter 5 were not. Combinations of the attributes are direct combinations of the three sets described in the table.

7.2.2 Evaluation Methods

There are three methods of evaluation used in this chapter: individual feature and group misclassification rates, held out set perplexity, and distance measurements.

Misclassification rates are calculated both at the individual feature level and for whole groups of features. Error rates are reported both for individual features and the average over the 9 or 10-dimensional set for the feature groupings. Misclassification rates allow for analysis of the groupings and attributes but may not be the best measure of the usefulness of a particular approach for recognition. Therefore, we also examine perplexity results for selected experiments, since the overall goal is to use the distribution in recognition rather than to use these models for prediction. These results are tested for significance with a Gaussian-based t-test.

Table 7.2: *Feature groupings for pronunciation prediction experiments.*

Factor Description	Attributes Used
Phone context	previous/following phone category (vow/gli/cons)
	lexical stress, previous phone deleted or not
	previous/following manner and placement categories
	phone location in word/syllable
Word Info	previous/following/current part of speech
	pivot point, dialog act label
	current word group (to further separate function words)
	previous/following/current word category
Prosody	Energy (20), F0 (32) and duration (4) features

Entropy and **perplexity** are calculated as in Chapter 5. To simplify the notation in the discussion here, we ignore the dependence on the word or prosodic features. Perplexity on the held out set is based on Eqn. 5.3 where the probability of the surface form phone is expanded to

$$p(\phi_j|b_j) = \prod_{l=1}^{N_f} p(f_{j,l}|b_j) \tag{7.6}$$

yielding

$$H_{test} = -\frac{1}{T}\sum_{j=1}^{T}\sum_{l=1}^{N_f} \log p_s(f_{j,l}|b_j) \tag{7.7}$$

where N_f is the number of features or the number of groups, T is the number of tokens in the held out set, and N is the number of phones in the set and where

$$p_s(f_{j,l}|b_j) = p(f_{j,l}|b_j) * 0.95 + p(f_{j,l}) * 0.05. \tag{7.8}$$

The perplexity is then simply

$$P_{test} = e^{H_{test}}. \tag{7.9}$$

Note that calculating perplexity values when diacritics are included requires a non-zero probability $p_s(f_{j,l}|b_j)$ for diacritic cases. For the cases represented in the held out set,

there are not always instances in the training set for a non-zero $p(f_{j,l}|b_j)$. In these cases, we could use the back-off that is the probability of the feature $p(f_{j,l})$ since the features themselves will have a probability over the entire set (i.e., voicing added to /f/ results in the same characteristics as /v/). Note also that for each time segment, in phone-based prediction, one decision with 47 options is being made while in feature-based prediction, 21 decisions with up to 5 options are being made. In this work, we will compare phone-based perplexity results with phone-level perplexity results for features modified and not modified by diacritics. The feature based results can then serve as a comparison for the diacritic-modified modeling.

In order to answer the question of whether feature modeling is more efficient than phone modeling, the feature prediction results need to be comparable to the phone prediction results. While the held out set perplexity value lets us answer this to a certain extent, it is more difficult to assess the phone-based system for phone labels using diacritics. Instead, we use the phone-level distance shown in Appendix A.3 which shows how different the predicted features are from the surface-form features. The fact that we can predict a broader range of "phones", means that we may capture more information about actual speech production. While our current phone-based recognition systems may not be able to use this additional granularity, recognition systems that are feature-based or incorporate feature information at some point (such as in clustering of HMM states) should be able to take advantage of the information.

The articulatory feature-based distance measure is used to calculate the distance between predicted features and the features corresponding to the surface-form phone. The average distance over the Held Out set between predicted features and surface-form features is compared to the distance between the surface form and baseform phones, calculated by summing the feature-based distance between the two phones. This distance can also be used for comparing feature groups.

7.3 Analysis of Feature-based Pronunciation Models

This section examines three things. The first is an evaluation of articulatory features as predictors of pronunciation variation. The second examines the interactions between groups of features. The third presents discussion of the prediction differences between features and phones.

7.3.1 Role of Attributes in Feature Prediction

The first case examined is the prediction of individual features. Recall that we are modeling features at the phone-segment level modified by diacritics. Because the most common phone transformations are between phones that share some features, the average context-independent (chance) levels for feature prediction are at 13.5% and 14.9% but they range from 10.9 to 17.3 and 11.5 to 21.4, for training and held out sets respectively. The results presented here will show that this baseline can be improved upon. All of the results described here are based on 21 features. There are 321,846 test samples in the held out set, or 15,326 samples per individual feature or group.

The decision trees are better than chance at modeling individual features. Table 7.3 shows the average results for predicting individual features using the collection of attributes described in Table 7.2. For comparison, classification results for phone-based features (i.e., without diacritics) are presented. There is less variability in the phone-based features, resulting in lower misclassification rates as well as, on average, smaller decision trees for modeling feature transformations.

In the experiments shown in Table 7.3, text-based information about the word and word-level context is the most useful for modeling. The trees do not do a good job of predicting deleted features even with the inclusion of the predicted phone transformation type as seen in preliminary experiments.[4] The poor prediction of deletions may be an artifact of the phone-level segmental labeling of the data, since it may be that the features themselves are

[4]Predicted phone transformation type was not useful for modeling feature transformations for non-diacritic experiments. However, using the oracle phone transformation type did make it easier to predict deletions and the results improve dramatically, from the 14.9% baseline (without diacritics and separate glottis features) to 3.9% (a 74% improvement). This is consistent with the surface-form prediction experiments.

Table 7.3: *Average feature-based pronunciation prediction misclassification results for individual features with no pruning (pruning did not have a positive effect on the Held Out set results). Groups of attributes match the best results from Chapter 5.*

	Phone-based Features			Features with Diacritics		
	Avg. Nodes	Feature % Error		Avg. Nodes	Feature % Error	
Factors	per Tree	Train	Held Out	per Tree	Train	Held Out
Chance	NA	13.2	14.6	NA	13.5	14.9
0. Phone context	77.7	12.2	14.7	78.6	12.4	15.0
1. (0) + word info	62.6	12.1	13.8	66.4	12.3	14.0
2. (0) + prosody	62.8	12.2	13.8	64.1	12.4	14.1
3. (2) + word info	62.5	12.1	13.8	63.4	12.3	14.1
4. Word info alone	82.0	13.1	14.6	84.3	13.3	14.9
5. Prosody alone	87.7	13.2	14.6	89.2	13.5	14.9

not actually deleted but are reflected in changes to neighboring features. Prosody and text-based (word) information both lead to gains but they appear to be somewhat redundant since there is no additional gain from using both types of features together. As also seen in Chapter 5, word category, location of the syllable in the word, and POS windows are frequently used when available (Expts. 1, 3 and 4). Questions were asked about dialog act categories in two of the 21 feature trees for Expt. 1, three of 21 for Expt. 3, and all 21 for Expt. 4. While the dialog act attribute was used in most surface form prediction trees in Chapter 5, it was consistently used in the feature trees for the feature "strident". For trees built using prosody (Expts. 2, 3 and 5), word duration was used as an attribute in all 21 trees. As seen in surface-form phone prediction, a variety of energy and F0 features were used with minimum, mean and maximum energy values in a word-sized window used most frequently. The word-level and prosodic attributes are used at high nodes in the trees, so their benefits will still be present even with extensive pruning.

The individual features vary in their prediction rates as seen in Table 7.4. The tongue-

Table 7.4: *Misclassification rates by feature for the held out set for the diacritic encoded feature Expts. 0, 1 and 2 in Table 7.3. Boldface is used to show the best result between Expts. 1 and 2 when it is significantly better than phone context. An * is used when there is a significant difference between Expts. 1 and 2.*

	Feature % Error			
Feature	Context Independent	Phone Context (Expt. 0)	Text (Expt. 1)	Prosody (Expt. 2)
All Features	14.94	15.02	**14.08**	14.10
Syllabic	12.31	12.31	11.54	**11.44**
Consonantal	15.84	16.77	**14.88**	14.92
Sonorant	12.55	12.53	11.78	**11.50**
Continuant	14.40	14.31	**13.39**	13.44
Strident	11.46	11.38	11.02	**10.52***
Delayed release	13.53	13.42	12.63	**12.45**
Voicing	14.18	14.38	13.85	**13.75**
Glottis	13.79	13.58	13.33	13.27
Advanced tongue root	17.15	17.16	**16.27**	16.50
Constricted tongue root	15.86	15.86	**15.04**	15.05
Nasal	17.95	17.96	**17.11**	**17.11**
Dorsal	13.52	13.52	12.65	**12.55**
Coronal	13.74	13.74	**12.90**	12.93
Labial	14.07	13.99	13.02	**13.00**
High	21.38	22.77	**20.01**	20.33
Low	17.00	17.00	16.12	**16.08**
Back	19.28	18.94	**17.85***	18.28
Anterior	13.92	13.92	**13.07**	13.08
Distributed	14.24	14.26	**13.39**	13.40
Lateral	13.77	13.79	**12.83***	13.39
Rhotic	13.81	13.82	**12.87**	13.12

associated features high, low and back, also vary more than most of the other features and are thus more difficult to predict. The syllabic feature, i.e., whether a phone is a vowel (or syllabic consonant) or not, has less variation. Prosodic attributes result in better prediction for syllabic, sonorant, strident, and voicing. Since voicing results in F0 values and is also "+" for vowels, this result is not surprising for distinguishing between vowels and consonants (i.e., the syllabic and consonantal features). Tongue-associated features such as constricted tongue root, dorsal, anterior, distributed, and lateral are predicted slightly more accurately with text-based attributes.

7.3.2 Feature Independence Assumptions

As discussed earlier, the articulatory features are physically related to each other so it makes sense to examine groups of features. This section examines the two sets of groups shown in Table 7.1. We present misclassification results for both groups (in terms of grouped feature prediction and individual feature prediction) and then examine the use of dependency information for the hierarchical groups (Grouping II).

The average feature prediction results for Grouping I are in Table 7.5. Again, the prediction of deleted features is not frequent in this experiment. (Preliminary results using the oracle transformation type show that better prediction of phone deletions improves the result.) Contrary to phone-based prediction and the results predicting individual feature transformations, the incorporation of prosodic attributes does not improve the prediction results on the held out set. The use of text-based information does significantly ($p < 0.0005$) improve the prediction rate.

In a similar way, Grouping II (consisting of 9 groups) can be treated as a set of independent groups by not including group dependency information as attributes in the trees. (Results with or without the baseform information about the parents are similar.) Results for Grouping II with no baseform or surface form dependencies are shown in Table 7.6. Again, prosodic attributes do not result in an improvement in the held out set but including text-based attributes significantly ($p < 0.002$) improves the result.

We measure individual feature prediction to see which of the groupings (I or II) results

Table 7.5: *Feature-based pronunciation prediction results for Grouping I with no pruning. Results describe misclassification of feature groups with respect to the hand-labeled phones with diacritics, and are not comparable to average feature prediction.*

Factors	Avg. Nodes per Tree	Group I % Error	
		Train Set	Held Out
Context-independent	NA	15.0	17.2
0. Phone context	68.5	13.9	16.8
1. (0) + word info	58.5	13.8	16.3
2. (0) + prosody	57.4	13.9	17.1
3. (2) + word info	56.3	13.8	17.0
4. Word info alone	75.5	14.7	17.0
5. Prosody alone	72	15.0	17.2

Table 7.6: *Feature-based pronunciation prediction results for Grouping II using no baseform or surface-form dependencies. Results describe misclassification of feature groups with respect to the hand-labeled phones with diacritics, and are not comparable to average feature prediction.*

Factors	Avg. Nodes per Tree	Group II % Error	
		Train Set	Held Out
Context-independence	NA	14.9	16.9
0. Phone context	66.9	13.9	16.5
1. (0) + word info	59.1	13.8	16.1
2. (0) + prosody	58.1	13.9	16.7
3. (2) + word info	56.6	13.8	16.7
4. Word info alone	75.3	14.7	16.8
5. Prosody alone	74	14.9	16.9

Table 7.7: *Individual feature prediction error rates for the three levels of grouping with no dependencies between groups on the Held Out set. Groups of attributes match the best results from Chapter 5. Results are misclassification error rate of individual features with respect to the hand-labeled phones with diacritics.*

Factors	Individual Features	Independent Groups	
		Grouping I	Grouping II
Context-independent	14.9	14.9	14.9
0. Phone context	15.0	14.7	14.3
1. (0) + word info	14.1	14.4	14.1
2. (0) + prosody	14.1	14.9	14.5
3. (2) + word info	14.1	14.9	14.7
4. Word info alone	14.9	15.2	14.9
5. Prosody alone	14.9	15.3	14.9

in better prediction of individual features rather than prediction of group categories. The group error rates are not directly comparable since a grouping could have 2 of 3 features correct but still be considered incorrect overall. Results for individual feature prediction are shown in Table 7.7. While there is a win for grouping features when only phone context is used, there is either no win or a loss as higher-level attributes are added. However, as we shall see later (in Table 7.14), there is a significant improvement in perplexity from the groupings. Grouping II consistently gives better prediction results than Grouping I, though again, we will see later that it has worse perplexity results. Table 7.7 give results for predicting diacritic coded features, but the conclusions are similar when training and testing without diacritics.

We can also examine specific differences in prediction of the different groups for different high-level attributes. Table 7.8 presents the results for individual groups for Expts. 1 and 2 using Grouping II. The average of the individual feature misclassification error rates is also included for comparison. The two values are equal for groups with one feature member. The differences between prosody and text-based features vary somewhat but the differences

Table 7.8: *Misclassification rates by group for the Held Out set for Expts. 0, 1 and 2 in Table 7.6 (Grouping II with independent groups), in terms of exact match for all features in the group (group % error) and matching at the feature level (individual feature % error). Boldface is used to show the best result between Expts. 1 and 2 when it is significantly better than phone context. An * is used when there is a significant difference between Expts. 1 and 2.*

	Group % Error			Individual Feature % Error		
	Expt. 0	Expt. 1	Expt. 2	Expt. 0	Expt. 1	Expt. 2
Feature	(Phone Context)	(Text)	(Prosody)	(Phone Context)	(Text)	(Prosody)
Group 1	16.7	**15.7***	20.3	14.1	**13.2***	15.6
Group 2	13.9	14.0	14.1	11.9	12.1	12.1
Group 3	13.4	12.6	**12.4**	13.4	12.6	**12.4**
Group 4	14.4	13.9	**13.8**	14.4	13.9	**13.8**
Group 5	19.0	18.9*	19.6	15.2	15.1*	15.8
Group 6	18.0	**17.1**	**17.1**	18.0	**17.1**	**17.1**
Group 7	15.7	15.5	15.9	13.0	**12.8**	13.2
Group 8	23.9	23.7	23.9	18.1	18.0	18.1
Group 9	13.6	13.4	13.5	13.2	13.0	13.1

across groups are more striking. The one case where prosody offers an advantage over text is Group 3 (delayed release). The useful effects of prosody for some individual features may be lost when groupings are used. As with the individual features, the tongue features (high, low, back) represented by Group 8 are the most difficult to predict.

Abandoning our assumption of independence allows us to examine a more realistic model of speech production. The decision trees for dependent subgroups predict the feature groups shown in Table 7.1 (Grouping II), but the attributes used by the tree now include the baseform and surface form feature information for the given feature groups that are only one level above the group, i.e., for group (7), the given group is (2) and for group (9), only (7) is given. (Results using more dependency levels will be discussed further below.) We present an upper bound on results for dependent hierarchical features using the hand-

Table 7.9: *Feature-based pronunciation prediction results on the held out set for Grouping II using various dependencies. Results describe misclassification of feature groups scored with diacritic encoded phones as truth. The first column matches Table 7.6.*

Factors	Independent Groups	Tree Dependence on Baseform		
		alone	+ surface form with diacritics (S1)	+surface form with no diacritics (S2)
0. Phone context	16.5	16.5	9.4	6.0
1. (0) + word info	16.1	16.0	9.3	5.9
2. (0) + prosody	16.7	16.7	9.8	6.3
3. (2) + word info	16.7	16.7	9.8	6.3

labeled surface form values for both phones and phones with diacritics. The compilation of results using various levels of dependencies is in Table 7.9.

The first two columns of Table 7.9 basically show that there is no need to refer to the baseform hierarchy in terms of prediction accuracy, though in fact the trees do sometimes choose these features. Also, there is little difference in the size of the two sets of trees (roughly 55-67 nodes per tree). When the higher level features are taken from the surface form phones, with (S1) or without (S2) the diacritics, the prediction error drops significantly and the average size of the trees is roughly half (21-25 nodes). One reason for this improvement is that the dependency information allows better prediction of deletions, since a parent feature labeled as deleted can be used to correctly predict a deleted child. Surprisingly, it appears to be better to train the models without the diacritics (S2) in that prediction error is roughly 30% lower. Note that both the surface form experiments are "oracle" experiments but these are most comparable to the cases of interest in perplexity scoring. In all four cases, note that the best performance is obtained with word (text-based) information.

Comparing Tables 7.8, independent groups, and 7.10, dependent groups given baseform and surface form with diacritics (S1), shows that there is a large difference in the prediction

Table 7.10: *Misclassification rates by group for the held out set for Expts. 0, 1 and 2 in Table 7.9 (Grouping II with tree dependencies on known surface form values (S1)), in terms of exact match for all features in the group (group % error) and matching at the feature level (individual feature % error). All results are significantly better than the context-independent case. Boldface is used to show the best result between Expts. 1 and 2 when it is significantly better than phone context. An * is used when there is a significant difference between Expts. 1 and 2.*

	Group % Error			Individual Feature % Error		
Feature	Expt. 0 (Phone Context)	Expt. 1 (Text)	Expt. 2 (Prosody)	Expt. 0 (Phone Context)	Expt. 1 (Text)	Expt. 2 (Prosody)
Group 1	16.7	**15.7***	20.3	14.1	**13.2***	15.6
Group 2	2.4	2.4	2.4	1.3	1.3	1.3
Group 3	1.4	1.3	**1.2**	1.4	1.3	1.2
Group 4	3.3	3.3	3.1	3.3	3.3	3.1
Group 5	6.3	6.0	6.3	2.4	2.3	2.4
Group 6	6.9	6.9	6.9	6.9	6.9	6.9
Group 7	21.6	21.6	21.6	15.4	15.5	15.5
Group 8	25.0	25.0	25.0	19.2	19.2	19.2
Group 9	1.3	1.3	1.3	0.5	0.5	0.5

rates for groups and features when dependencies are included. As expected, we see no improvement in Group 1 since it has no parent feature groups. Unexpectedly, Groups 7 and 8 actually do at chance or worse (16.5% and 25.0%) when the dependent information is included. One possible reason for this is that in Group 7, the tree does not use the baseform value of Group 2 and instead chooses the diacritic-modified surface-form. In a test situation, it may not be as powerful for accurate prediction. Similarly, for Group 8, while the baseform value of Group 2 is selected (which might be more robust in testing), it is not high in the tree. While the prosodic features do not predict Group 1 very well, they do relatively well compared to text features on the remaining groups.

This unexpected result drives the next experiments on the usefulness of levels of depen-

dency in modeling feature transformations. In building our dependent trees, we initially hypothesized that one level of feature dependence is sufficient for dependent feature group prediction results. This was further explored by examining the trees and results for the pertinent groups of features:

- Group 7: dorsal, coronal, labial
- Group 8: high, low, back
- Group 9: anterior, distributed, lateral, rhotic

Since Group 8 is the group that has consistently been most difficult to predict, this group has the most room for improvement by including more information. Groups 7 and 8 could have up to two levels of dependence while Group 9 can have up to three levels of dependence. Table 7.11 shows the misclassification rate differences for the known surface-form prediction case, using diacritic enhanced values for the dependencies. Results with no dependencies are a subset of those in Table 7.6 and results with one level of dependence are a subset of those in Table 7.9 (S1). An additional column gives results when one additional level of baseform dependency is included without the surface form value (Table 7.9, baseform alone). As expected, the results are between the "None" and "One" dependency level results. These results, modeling diacritic encoded features using surface form features that are also diacritic encoded, do not show that there are positive effects when more levels of dependence are included in the trees. However, this may be due to the increased amount of variability when using the diacritic encoded features. Table 7.12 shows the results for trees trained with phone-based surface form features, i.e., assuming the canonical form of the features rather than modifying with the diacritics (Table 7.9 (S2)). The predicted values are still diacritic encoded features. These trees do a significantly better job predicting the surface form values of Groups 7-9. The perplexity differences for these results are in Table 7.13. While we still see no gain from including prosodic features for misclassification, there are slight improvements from prosody in perplexity for the case of no dependencies. The overall improvements in perplexity for increased levels of dependencies mirror the improvements in misclassification rates.

Even with the mismatch between training and testing, i.e., training without diacritics and predicting the diacritics, we see improved misclassification rates when adding a second

Table 7.11: *Levels of dependency and feature prediction: Held Out set group misclassification rates for Groups 7-9 of Grouping II with varying levels of dependencies. Oracle diacritic encoded features are used in training. The context-independent baseline is 18.6%.*

	Levels of Dependence			
Factors	None	One	Two	Base: one, Surface: none
0. Phone context	17.7	16.0	19.2	17.6
1. (0) + word info	17.5	16.0	19.2	17.5
2. (0) + prosody	17.8	16.0	19.2	17.6
3. (2) + word info	17.8	16.0	19.2	17.6

Table 7.12: *Levels of dependency and feature prediction: Held Out set group misclassification rates for Groups 7-9 of Grouping II with varying levels of dependencies. Oracle phone-based surface form features are used in training. The context-independent baseline is 18.6%.*

	Levels of Dependence		
Factors	None	One	Two
0. Phone context	17.7	7.0	6.3
1. (0) + word info	17.5	7.0	6.2
2. (0) + prosody	17.8	7.1	6.3
3. (2) + word info	17.8	7.1	6.2

Table 7.13: *Levels of dependency and feature prediction: Held Out set perplexities for Groups 7-9 of Grouping II with varying levels of dependencies. These results match Table 7.12.*

	Levels of Dependence		
Factors	None	One	Two
0. Phone context	6.17	2.35	2.06
1. (0) + word info	6.02	2.34	2.05
2. (0) + prosody	5.93	2.35	2.06
3. (2) + word info	5.90	2.34	2.05

level of dependencies. It should be noted, however, that the dependency features of the levels are consistently used as attributes in the decision trees for all dependency experiments. Although we see significant differences when adding a second level of dependence (Table 7.12), there is no significant difference between including the second and third levels of dependence for known surface form-based prediction. Adding a third level of dependence did not change the results for Group 9, because the attributes used for prediction were the same as those with two levels of dependency, giving 1.1% error rate for the training data.

7.3.3 Features vs. Phones

It is difficult to compare the results to phone-based prediction for two reasons. First, the feature model allows more possibilities for a fine grained representation of phone specification. Secondly, the level of detail associated with features was not fully annotated in the corpus (though encoding the diacritics does include some of the differences). With the caveat that any measurements are biased to favor phone-based prediction because of the phone-based labeling (i.e., the segment based structure and not all possible feature changes are labeled or labeled consistently), there are two types of information we look at to compare feature prediction performance with phone prediction results. The first is a phone-level perplexity measurement. The second looks at the feature-based distance presented in Chapter 5 to show the distance between surface form phones and the predicted set of features.

Table 7.14 shows the held out set perplexities for modeling of features with diacritics, effectively a phone-level measurement. While including the text-based information gives the best result for misclassification, including prosodic information reduces the perplexity on the Held Out set for individual features and independent groups (an improvement of up to 19% over the baseline phone context case for groups and 71% for individual features). The assumption of independent features gives a very high perplexity result,[5] and it is not a good model of pronunciation variability. It allows unrealistic combinations of features such as having a spread glottis and voicing. However, the use of groups reduces the perplexity significantly, and adding hierarchical dependence gives significant further gains.

[5]Although the independent feature perplexity looks too high to be possible, recall that if each feature has N possibilities and there are M features than it could be as high as N^M or 5^{21}.

Table 7.14: *Feature-based pronunciation prediction perplexity results representing a phone-level measurement for prediction of diacritic encoded phones. Phone-level chance perplexity on the Held Out set is 38.3.*

Factors	Individual Features	Independent Group I	Independent Group II	Dependent Group II
0. Phone context	34,172	90.7	142.4	33.4
1. (0) + word info	12,380	74.4	114.6	31.8
2. (0) + prosody	10,509	74.2	114.3	32.9
3. (2) + word info	9,806	73.1	112.5	32.7

Table 7.15 compares phone-based prediction with feature-based prediction of surface form phones without diacritics. While the phone-based prediction numbers have lower perplexity, feature-based prediction is tending towards the phone-based numbers.

Table 7.16 shows the percentage of surface-form phones that are predicted completely based on the combination of the predicted articulatory features, first comparing phone-based features and then diacritic encoded features. The rest of this section examines the remaining incomplete matches using the feature-based distance. For feature-based prediction, a complete phone level match is typically lower than chance and lower than the phone-based prediction of Chapter 5. When diacritics are included in scoring, the percentage of matches for the phone-based system drop by an additional 5% (real).

The features that are predicted are rarely completely wrong for the surface-form phone. We use the feature-based distance measure to examine how different the predicted "phones" or combinations of features are from the surface form. Using the distance allows us to compare phones vs. features on the data with diacritics and to examine the range of phone instances that may be produced. In the case of a feature that is unspecified but changeable (such as "nasal") which has values of -1, 0 and 1, the fact that the prediction was incorrect counts as one error. When measuring the distance, any error is half the distance from

Table 7.15: *Feature-based pronunciation prediction perplexity results representing a phone-level measurement. Feature and phone results are based on modeling without diacritics. Phone-level chance perplexity on the Held Out set is 38.3.*

No Diacritics Predicted		
Factors	Phone-based	Dependent Group II
0. Phone context	2.94	7.45
1. (0) + word info	3.05	7.05
2. (0) + prosody	2.98	7.34
3. (2) + word info	2.98	7.26

Table 7.16: *Complete phone-level matches for phone-based, individual features and Grouping II (with oracle surface form information) experiments. Chance for phone-level prediction without diacritic notation on the Held Out set is 67% and with diacritic notation is 62%. Results are presented with no diacritics in training or prediction and with diacritics used in both training and prediction.*

	No Diacritics			Diacritics Included		
Factors	Phone-based	Individual Features	Dependent Features	Phone-based	Individual Features	Dependent Groups
0. Phone context	70.0	66.0	69.5	65.3	59.6	58.3
1. (0) +word info	69.9	67.0	69.2	65.2	61.2	58.0
2. (0) +prosody	69.7	66.9	64.3	65.1	60.5	53.6
3. (2) +word info	70.1	66.3	64.4	65.4	60.0	53.7

the true value of 0, as it would be for a feature that did not allow a 0 value.[6] Ideally, misclassifying this feature would not hurt a recognition system as much as misclassifying the syllabic feature, for example. When a feature is deleted, the added distance is half the maximum possible distance, or 1. (This is slightly different from the method used in Chapter 5, which weights the overall cost of phone deletion, but is implemented here for simplicity and ease of comparison.) When a feature is incorrectly predicted to be "X," or is "X" and is misclassified, the added distance is the maximum possible distance, or 2. Unlike the distance used in Chapter 5, the distance is not multiplied by 2, presenting a more intuitive average of how many of the 21 features match the labeled surface form.

The average distances from the surface-form phone set and diacritic encoded feature set are shown in Table 7.17. Phone-based prediction comes from Chapter 5 with the distances recalculated to match the metric used here (rather than the costs used for alignments) and with insertions excluded (as opposed to Chapter 5 where they were included). While phone-based predictors are more likely to predict the correct surface-form phone than feature-based predictors, even with diacritics included in scoring (e.g., 65.3% correct vs. 59.6% and 58.3% using phone context), the errors that are made are further from the surface form value. Feature-based predictors are more likely to generate a closer representation of what is actually produced than phone-based predictors. When diacritics are not predicted, the distance is further away. A large portion of this difference, however, can be explained by the incorrect prediction of deleted phones. Because the phone-based predictors are more likely to predict a full deletion than the feature-based predictors, the average distances are also presented for prediction without including incorrectly predicted deletions (Table 7.18).

As seen in Table 7.19, glides typically have a higher distance from the average for feature results because of the way diphthongs are converted to feature sets. (Note, however, that the use of dependent groups reduces the distance significantly when text and prosodic information is included.) When a diphthong is misclassified as a vowel, the glide portion of the diphthong is considered deleted, raising the overall error for glides. This also creates a

[6]Ideally, the distance should be zero if the feature shows up on certain neighboring phones, but the implementation of such a distance is complex and so not included here.

Table 7.17: *Predicted phone-level distance between: 1) hand-labeled surface form and predicted phones and individual features and 2) hand-labeled surface form with diacritics and predicted individual features and dependent groups.*

	No Diacritics Predicted			Diacritics Predicted		
Factors	Phone-based	Individual Features	Dependent Groups	Phone-based	Individual Features	Dependent Groups
0. Phone context	7.6	11.4	10.6	11.8	10.1	8.9
1. (0) +word info	7.8	10.8	10.5	11.9	9.7	8.9
2. (0) +prosody	8.1	10.7	9.4	11.8	9.5	8.3
3. (2) +word info	7.9	10.5	9.4	11.8	9.4	8.3

Table 7.18: *Predicted phone-level distances matching Table7.17 with deletion distances excluded.*

	No Diacritics Predicted			Diacritics Predicted		
Factors	Phone-based	Individual Features	Dependent Groups	Phone-based	Individual Features	Dependent Groups
0. Phone context	5.1	4.3	5.0	11.1	4.1	4.2
1. (0) +word info	4.9	5.4	5.1	11.1	5.1	4.3
2. (0) +prosody	4.9	5.5	4.6	10.8	5.1	4.1
3. (2) +word info	4.9	5.5	4.6	10.9	5.1	4.1

Table 7.19: *Predicted phone-level distance for vowels, consonants and glides between: 1) hand-labeled surface form and predicted phones and individual features and 2) hand-labeled surface form with diacritics and predicted individual features and dependent groups.*

0. Phone context						
	No Diacritics Predicted			Diacritics Predicted		
Category	Phone-based	Individual Features	Dependent Groups	Phone-based	Individual Features	Dependent Groups
Vowels	4.9	5.1	5.7	7.2	4.3	4.3
Consonants	11.0	17.2	14.4	16.9	16.3	15.3
Glides	6.2	14.4	19.1	19.4	13.6	15.8
3. (0) + word info + prosody						
	No Diacritics Predicted			Diacritics Predicted		
Category	Phone-based	Individual Features	Dependent Groups	Phone-based	Individual Features	Dependent Groups
Vowels	5.2	5.1	5.8	7.2	4.3	4.3
Consonants	11.3	14.4	13.4	17.1	14.0	14.5
Glides	8.5	16.5	9.3	19.4	14.5	9.2

slight mismatch in the results presented in Table 7.16, since diphthongs are counted as a single phone in the phone-based prediction, "artificially" reducing the error count of phone-based prediction with respect to feature-based. Nevertheless, modeling with articulatory features and higher-level information reduces the distance by up to 6% when predicting features without diacritics and 62% when predicting features with diacritics.

If minimum phonetic distance is the criterion, then using articulatory features is the best option for modeling pronunciation variation, which includes a wider variety of production than the typical phone set used in ASR. In addition, large reductions in perplexity are obtained by representing feature groups and dependent feature information compared to modeling individual features. Modeling with dependent feature groups is also the best

Table 7.20: *Number of terminal nodes used to model surface-form pronunciation variation.*

Factors	Individual Features	Independent Group I	Independent Group II	Dependent Group II	Phones (Pruned)
0. Phone context	1651	548	602	229	899
1. (0) + word info	1395	468	532	217	1884
2. (0) + prosody	1347	459	523	211	2717
3. (2) + word info	1331	450	509	213	2726

option for reducing the complexity of the pronunciation models. When modeling surface-form phone transformations, we need decision trees for 47 phones. When modeling surface-form feature transformations, we use 21 decision trees for individual feature prediction, 8 for Grouping I and 9 for Grouping II.

The total numbers of terminal nodes for the combined set of trees used in the experiments are in Table 7.20. The best feature-based approach uses roughly 10% of the number of terminal nodes as the phone-based approach. In addition, phone prediction involves using approximately 47 trees (one per phone), while feature predictions involve fewer (8-9), smaller trees that can be run in parallel, so it is possible to leverage this structure for faster prediction the dynamic modeling case. Including prosodic attributes along with phone context reduces the number of nodes used to model feature variation (but not phone variation), even though many more attributes are available for questions. This suggests that prosodic attributes are able to encode information about variation more efficiently even though they may not lead to gains in performance.

7.4 Implementation Options for Features in Speech Recognition

While low test perplexity of features (and to a lesser extent accurate prediction) is certainly a goal, and the prediction experiments with dependent feature groups show promising results, we still face the problem of incorporating feature prediction into a phone-based recognition

system. Within existing systems, one approach is to map the predicted feature vector to the closest phone using the feature-based distance. Alternatively, a complete system (i.e., both acoustic and pronunciation models) can be built with feature bundles as the primary acoustic unit. The advantage of the first approach is the ability to work with existing systems, while the second approach avoids the approximation of the feature mapping. This work has shown that deleted features are hard to predict when dependencies are not included. In this case, the ASR implementation would need to account for deleted features in the neighboring segments in the acoustic model. Part of the motivation for features is to get more granularity in the representation of speech production, allowing features to be deleted or shifted as happens in conversational speech, so we present two ways to use it more directly.

The first option is state-level clustering of acoustic models based on a feature representation of the context rather than phone context. A set of pilot recognition experiments on Dev '98 (using a five state model topology with shared covariances) used articulatory features for clustering. This produced a comparable word error rate performance to clustering with traditional phone-based information. This approach could be combined with state-level feature prediction to capture production changes such as nasalization at the end of a vowel but not the beginning. While nasalization is captured by triphones when the neighboring /n/ is present, it is not captured when a phone-based pronunciation model predicts deletion of the nasal segment.

Another option that is a more significant departure from current systems would be to use a factorial HMM as in [65] with different state chains corresponding to different feature groups. These groups could then be connected to acoustic feature observations representing the articulation production as in [47]. Dynamic Bayesian networks are also suggested as a way to implement a feature-based recognition system [55].

7.5 Summary

This chapter presents several findings related to modeling pronunciation variation through the use of articulatory features. We have addressed the questions raised in the introduction

as follows.

1. What intermediate dictionary representation (phones, features or feature groups) is most useful for predicting pronunciation changes?

   While models built using either phones or features will allow the prediction of pronunciation changes, feature-level modeling allows a representation of sounds that may better match actual speech production. Because this work is based on hand-labeled phones rather than hand-labeled features, there is a bias towards better phone prediction. Despite this, the feature-based models are more efficient and result in feature combinations that are close to the actual surface form phone in the sense of smaller distances.

2. Are high-level conversational attributes used more often in feature prediction than in phone prediction?

   As with phone-based modeling, local word and phone attributes were most commonly used. For independent (grouped or not) feature prediction, the prosody and word-based decision trees used dialog act, the average energy level over the utterances, the number of F0 slope changes over the utterance, the utterance duration as well as word category and POS windows. However, when dependence information is included, the number of high-level attributes used by the trees drops and local word and phone information is used rather than utterance or conversational level attributes, consistent with what we find for phone-based modeling. The root node feature trees (Group 1) in the dependency experiments still use the high-level features, thus allowing the effects of the high-level features to be incorporated in the dependent variables. In general, prosody was not as useful for feature-based prediction. However, this is largely due to disappointing results on the syllabic/consonantal features.

3. Do hierarchical dependencies between features offer significant advantages over independent feature prediction?

   While all of the decision trees built included the same phone context information used in Chapter 5, it was not until we included information about the hierarchical depen-

dence relationships between articulatory features that results were in the same range of perplexity and classification as phone-based modeling. Information that is inherently coded in the phone needs to be explicitly modeled in feature-based prediction. Independent groupings of features performed slightly better than predicting individual features when only phonetic context was available but actually did worse when higher-level information was used. Including dependencies improved performance significantly, reducing both perplexity and distance. While a single level of feature dependency clearly improves the results, further improvements come from increasing the levels of dependency in the trees with reductions in perplexity and misclassification rates on the order of 13% for including a second level of dependency when baseform and phone-based surface form features are used in training.

Feature-based pronunciation modeling shows promise for pronunciation modeling, and can be incorporated into existing ASR systems with only slight modifications. Future possibilities for this work are on two fronts: further exploration of different feature hierarchies and feature-based recognition.

The method used here can be applied to testing the applicability of various models of articulation in order to determine which linguistic theories best model what is observable in speech data. Given that feature-based analysis is common in the field of linguistics, exploring the modeling of articulatory features for potential use in recognition systems is an important task. Suggested starting points for further feature exploration include examining other models of articulation (e.g., [82]) and their suggested articulator dependencies as starting points in the feature hierarchy model.

Further exploration of the use of feature-based pronunciation models within ASR systems is clearly necessary. While some aspects of the feature-based models, such as increased granularity, may be ignored by existing phone-based systems, this may be addressed by using features as the symbolic labels for acoustic model indexing and clustering. This would address the question of which predictive models will yield the best results for recognition, feature-based or phone-based, and which type of recognition system will best leverage the flexibility of articulatory features.

Chapter 8

CONCLUSIONS

This chapter presents a summary of the main contributions and implications of the work. It closes with suggestions for continued work in the areas of analysis, prediction and recognition, with an emphasis on feature-based work, which seems to be the most promising aspect for future work.

8.1 Summary and Impact

Intra-speaker variation remains a problem for speech recognition systems. Pronunciation modeling allows speech recognition systems to deal with more of the inherent speaking variability that comes with spontaneous speech. The main contributions of this work have been an analysis of pronunciation variability in terms of high-level conversational factors including prosodic factors (F0, energy and duration measures) and text-based information (word categories, dialog act labels, and part-of-speech tags), a new recognition framework for pronunciation modeling that includes these factors, and an evaluation of articulatory features for pronunciation modeling. The following three subsections will summarize the contributions.

8.1.1 Analysis of Factors Affecting Pronunciation Variation

As a framework for modeling pronunciation variation, we introduce a two-stage prediction model that involves first predicting either a word-level pronunciation quality measure or a phone-level transformation type or both. The phone-level transformations included five broad categories: hyper-articulated, reduced, deleted, remaining the same, or other change. The predictions from these models are then used as attributes in a surface form transformation model. The advantage of the intermediate predictors is that they represent broad

categories of changes (rather than detailed phone-based changes) that are easier to train given the sparsely represented higher-level cues. While intermediate predictors give a small gain when included in surface form prediction, using oracle information (i.e., the known intermediate values), gave large wins in predicting variation. Hence, improved modeling of the intermediate values could yield further improvements in pronunciation modeling. While the two-stage model did not result in large gains for surface form prediction, the intermediate variables proved to be useful for analyzing the utility of different high-level attributes at the word level as well as the phone level.

Using intermediate predictors, we showed that word-level variation is correlated with such factors as part-of-speech and word categories, word predictability and position of the word in an utterance. We used these attributes to predict the distance between the surface form pronunciation of a word from the canonical baseform. For phone transformation type prediction, we showed that within-word text-based attributes, i.e., phone context and basic dictionary information were useful. While higher-level attributes are useful for this task in general, phone location in the word, i.e., whether at the start of a word or not, is a major factor for differentiating phone transformation types, especially for reduction and deletion. We confirm that phones at the start of a word are much less likely to be reduced or deleted, matching earlier findings by Greenberg [32]. The behavior of hyper-articulations (including insertions) is distinguishable enough that we predicted it with our trees, even though it accounts for only 1.7% of all phone transformations. The category "reduced" was a difficult category to predict. Reduction was only predicted when prosodic attributes were included in the model.

The investigation of the effects of higher-level information is an important contribution both to speech recognition and the understanding of factors related to pronunciation variability in conversational speech. In this area, we have shown that there are correlations between pronunciation variability and such text-based attributes as part-of-speech and word category information although discourse features were not particularly useful. Using prosodic attributes does improve prediction slightly and duration is one of the single most important cues available, with different representations of F0 values being used frequently as well. While we explored various utterance-level attributes, word-level and local

context values were typically more useful. Windows of POS and word categories were useful and appeared frequently in the decision tree predictors. For prosodic features, word-level attributes normalized by either utterance or speaker-level values were the most frequently used. While some of the text-based variables may be correlated with prosodic cues, experimental results show that the combination of the two types of variables (word-based and prosodic) does help surface form prediction.

It should be noted that having a large variety of available attributes is important as different F0, energy and duration features are used for predicting different baseform-to-surface transformations. Similarly, intermediate predictors used different attributes for prediction than phone-level surface form prediction. All classes of text-based (except discourse features) and prosodic attributes examined were useful for both word-level and phone-level variation.

Overall, we showed that higher-level information is useful for predicting pronunciation variation, with significant perplexity reductions coming primarily from text-based attributes, and that a two-stage model of prediction has promise as a framework for prediction of surface form phone transformation, but work is needed to improve the first stage before the upper bound shown with oracle results is reached.

8.1.2 ASR with Pronunciation Models

In this work, we implemented various types of pronunciation models into an ASR system, focusing on text-based cues since these gave the biggest gain in surface form prediction experiments. The first type was a static model using phonetic context and word-based lexical information. The second type was an extended static model that incorporated either part-of-speech information or a five-class word category label. Both of these types resulted in transformed dictionaries, with new pronunciation strings and cost-based weights for the strings. When less training information is available, it is beneficial to include word-level information. However, the best recognition results come when the dictionary weights are retrained using word-level frequencies from a forced alignment of training data, which is consistent with work reported by others. The word error rate was significantly reduced by

0.8% (absolute) using a simple word-based pronunciation model, but not further reduced with the extended static model.

The framework for and implementation of dynamic modeling in the context of N-best rescoring is a contribution of this work. This work presented recognition results for two text-based versions of dynamic PMs, with attributes consisting of phone-level context plus part-of-speech windows or part-of-speech windows, word category windows and word location in utterance. Although phone-level perplexity is significantly reduced with the dynamic PMs, and word error rate is reduced with more high-level information within the dynamic framework, we find that this does not translate into improved recognition results over the static models. The current dynamic framework has no way to include word-level information about a particular pronunciation string, although word-level information is included for individual phone prediction. Future improvements could come from intelligent pruning of word-level phone paths in the phone-level latticelets, phone-sequence dependence modeling, or a word-level implementation. Of couse, a larger data set for training the decision trees would give improved phone-level costs which should result in recognition improvements based on results for the static case. We anticipate the larger training set is even more important for the the dynamic model because the higher-level information is more sparsely represented in speech.

There are also possibilities for further investigation of incorporating PMs within the current ASR framework. Including prosodic attributes in the dynamic PMs as well as a likelihood ratio to compensate for the hypothesis-dependent values is an important next step. While current work gives the language model and pronunciation model scores the same weight when combined with acoustic model likelihoods, splitting up the LM and PM weights may be a better way to take full advantage of the pronunciation model.

8.1.3 Articulatory Features for Pronunciation Modeling

The examination of articulatory features for pronunciation modeling resulted in several contributions. First of all, the features were used to create a feature-based distance used for phone string alignment and as an intermediate variable for predicting phone-level pro-

nunciation variation. Articulatory features were then used for prediction of pronunciation variation to compare feature-based vs. phone-based modeling. We expect that the articulatory features are better suited to taking advantage of higher-level information since they are low dimensional and share training data across phones. After developing the framework for using features and feature groupings in pronunciation prediction, we found that some features (or groups of features) were significantly more difficult to predict than others. In particular, features associated with the tongue (high, low, back) consistently had higher prediction error rates than other features. We also introduced the use of dependencies based on a hierarchical model of articulation. The use of these dependencies greatly improves prediction as well as reduces perplexity. We found that prosodic attributes are used for feature prediction and, for independent individual feature prediction, prosody alone did as well as phonetic context information in some cases.

In this work, we took advantage of the fact that the ICSI hand-labeled data includes a large number of diacritics noting feature changes of produced phones. When mapping the surface-form phones to features, including the diacritics allows for a more accurate joint representation of feature changes, as well as clarifying the picture of what feature combinations are actually produced by speakers. While we showed that including diacritics makes it much more difficult to predict variation at the phone-level, they are likely to give information about, for example, whether the following phone is deleted or reduced in some way. Further examination of dependencies in time (as well as feature dependencies) would take advantage of this aspect of labeling.

We showed that the use of features for pronunciation modeling reduces the (feature-based) distance between the predicted and the hand-labeled surface form vector of features over phone-based prediction, even though the feature prediction was trained to optimize likelihood and not distance. In addition, the feature prediction model is a more efficient representation of pronunciation variation in the sense of requiring at least 75% fewer parameters for similar or better prediction performance.

8.2 Future Directions

There are several areas associated with this work that could benefit from further exploration. This section will raise questions in the areas of analysis, prediction and recognition.

8.2.1 Furthering High-level Information Analysis

While this work explored the connections between pronunciation variation and syntax, discourse and prosody, some potential attributes remain untested. The first acoustic correlate of prosody that could be examined is the use of speaking rate. While duration measures may capture some speaking rate information, measures such as mrate, an energy based speaking rate measure [63], has been shown to be useful in other work ([26, 27]) and is likely to be useful in the paradigm presented here.

The energy attributes used here were not particularly helpful for surface-form phone prediction but this may be a function of our choice of word-level time windows and windows calculated using word boundaries. Syllable-specific measures rather than word-level measures may be better acoustic correlates for prosodic stress and emphasis.

In this work, syntax has been represented simply by part-of-speech labels and a three word window of part-of-speech information. Since syntax is often structured with hierarchical dependencies (as in parse trees), including information about these dependencies in decision trees would test whether syntax, beyond part-of-speech, is useful for prediction.

8.2.2 Pronunciation Prediction

As was shown in Chapter 7, increasing the level of dependencies across the subgroups of features improved prediction accuracy. Given the framework to evaluate different feature hierarchies, presented here, an exploration of different groupings and hierarchies of features has the potential for improving prediction as well as leading to a better understanding of proposed feature hierarchies within the articulatory system. Exploring existing feature hierarchies could be augmented with automatic learning methods for finding hierarchical feature dependencies. Adding temporal dependencies of the surface-form realizations would also be a possible way to improve prediction, especially in the case of cross-word pronun-

ciation prediction and in light of the importance of word-level pruning in the phone-based models.

Another way to expand feature-based prediction would be to explore finer temporal categories. State-based pronunciation modeling using features would allow for models that show how a phone may specifically change in context. While triphone acoustic modeling with Gaussian mixtures captures a lot of this variability, feature-based modeling may provide a more systematic description that will lead to tighter acoustic models. Since the dynamic pronunciation results suggest that word-level or sequential dependencies are useful, it would be important to extend the dynamic framework to include these.

8.2.3 Feature-based Speech Systems

The most obvious extension of this work is a feature-based recognition system. Some possibilities are described in Chapter 7. One of the limitations of feature-based prediction is the mismatch between predicted feature combinations and existing acoustic models. Acoustic models that describe the feature acoustics in combination with models that predict variation in the features are likely to give the biggest improvements for speech recognition.

Additionally, the use of an articulatory-feature-based system may be better suited for language-independent recognition systems than phone-based systems. The articulatory feature combinations that may result in phones not used for distinction in English may be useful in recognition of other languages. Languages with small amounts of training data can also be bootstrapped from feature-based models.

BIBLIOGRAPHY

[1] J. Allen and M. Core. Coding dialogs with the DAMSL annotation scheme. In *Working Notes of AAAI Fall Symposium on Communicative Action in Humans and Machines*, 1997.

[2] M. Bacchiani. *Speech Recognition System Design based on Automatically Derived Units.* PhD thesis, Boston University, MA, USA, 1999.

[3] J. K. Baker. The Dragon system – an overview. *IEEE Trans. on Acoust., Speech, and Signal Proc.*, ASSP-23(1):24–29, 1975.

[4] D. Baron, E. Shriberg, and A. Stolcke. Automatic punctuation and disfluency detection in multi-party meetings using prosodic and lexical cues. In *Proceedings of ICSLP*, pages 949–952, 2002.

[5] R. Bates and M. Ostendorf. Modeling pronunciation variation in conversational speech using syntax and discourse. In *Proceedings of the Workshop on Prosody in Speech Recognition and Understanding*, pages 17–22, 2001.

[6] L. Breiman, J. Freidman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth and Brooks, CA, USA, 1984.

[7] W. Byrne *et al.* Pronunciation modelling at WS97. Technical Report 30, The Johns Hopkins University (Center for Language and Speech Processing) Summer Research Workshop, 1997.

[8] N. Chomsky and M. Halle. *The Sound Pattern of English.* MIT Acoustics Laboratory, Cambridge, MA, 1968.

[9] W. Chou. Discriminant-function-based minimum recognition error rate pattern-recognition approach to speech recognition. *Proc. of the IEEE*, 88(8):1201–23, 2000.

[10] J. Clark and C. Yàllop. *An Introduction to Phonetics and Phonology.* Blackwell, London, UK, 1990.

[11] M. Cohen. *Phonological Structures for Speech Recognition.* PhD thesis, Computer Science Division, Department of Electrical Engineering and Computer Science, University of California, CA, USA, 1989.

[12] N. Cremelie and J.-P. Martens. Automatic rule-based generation of word pronunciation networks. In *Proceedings of Eurospeech*, pages 2459–2462, 1997.

[13] A. Cutler, D. Dahan, and W. van Donselaar. Prosody in the comprehension of spoken language, a literature review. *Language and Speech*, 40:141–201, 1997.

[14] L. Deng. A dynamic, feature-based approach to the interface between phonology and phonetics for speech modeling and recognition. *Speech Communication*, 24:299–323, 1998.

[15] L. Deng and K. Erler. Structural design of a hidden Markov model based speech recognizer using multi-valued phonetic features: Comparison with segmental speech units. *Journal of the Acoustical Society of America*, 92(92):3058–3067, 1992.

[16] L. Deng and D. Sun. Phonetic classification using HMM representation of overlapping articulatory features for all classes of english sounds. In *Proceedings of ICASSP*, pages 45–48, 1994.

[17] E. Eide. Automatic modeling of pronunciation variations. In *Proceedings of Eurospeech*, pages 451–454, 1999.

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

[19] E. Eide, J. R. Robin Rohlicek, H. Gish, and S. Mitter. A linguistic feature representation of the speech waveform. In *Proceedings of ICASSP*, pages 483–486, 1993.

[20] M. Finke, J. Fritsch, D. Koll, and A. Waibel. Modeling and efficient decoding of large vocabulary conversational speech. In *Proceedings of Eurospeech*, pages 467–470, 1999.

[21] M. Finke and A. Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *Proceedings of Eurospeech*, pages 2379–2382, 1997.

[22] J. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *Proceedings of IEEE Workshop ASRU*, pages 347–352, 1997.

[23] J. Fiscus, W. Fisher, A. Martin, M. Przybocki, and Pallett D. 2000 NIST evaluation of conversational speech recognition over the telephone: English and Mandarin performance results. In *Notebook Proceedings of the Speech Transcription Workshop*, 2000. `http://www.nist.gov/speech/publications/tw00/html/cts10/cts10.htm`.

[24] J. Fiscus, J. Garofolo, M. Przybocki, A. Martin, G. Sanders, D. Pallett, and A. Le. Rich transcription 2002 evaluation. In *Notebook Proceedings of the Rich Transcription Workshop*, 2002. `http://www.nist.gov/speech/tests/rt/rt2002`.

[25] E. Fosler-Lussier. Multi-level decision trees for static and dynamic pronunciation models. In *Proceedings of Eurospeech*, pages 463–466, 1999.

[26] E. Fosler-Lussier and N. Morgan. Effects of speaking rate and word frequency on conversational pronunciations. In *Proceedings of ESCA Pronunciation Modelling Workshop*, pages 35–40, Kerkrade, The Netherlands, 1998.

[27] J. E. Fosler-Lussier. *Dynamic Pronunciation Models for Automatic Speech Recognition.* PhD thesis, University of California, Berkeley, CA, USA, 1999.

[28] T. Fukada and Y. Sagisaka. Automatic generation of a pronunciation dictionary based on a pronunciation network. In *Proceedings of Eurospeech*, pages 2471–2474, 1997.

[29] J. Glass. Lattice-based models for recognition. In *Proceedings of 1999 IMA Mathematical Methods for Speech Recognition and Langauge Processing*, to appear.

[30] J. Godfrey, E. Holliman, and J. McDaniel. Switchboard: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, pages 517–520, 1992.

[31] S. Greenberg. The Switchboard transcription project. Technical report, The Johns Hopkins University (Center for Language and Speech Processing) Summer Research Workshop, 1996. `http://www.icsi.berkeley.edu/~stp`.

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

[33] S. Greenberg, H. Carvey, and L. Hitchcock. The relation of stress accent to pronunciation variation in spontaneous american english discourse. In *Proceedings of the ISCA Workshop on Prosody and Speech Processing*, 2002.

[34] S. Greenberg, S. Chang, and L. Hitchcock. The relation between stress accent and vocalic identity in spontaneous American English discourse. In *Proceedings of the Workshop on Prosody in Speech Recognition and Understanding*, pages 51–56, 2001.

[35] B. Grosz and C. Sidner. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12:175–204, 1986.

[36] T. Hain. *Hidden Model Sequence Modelling in Automatic Speech Recognition.* PhD thesis, Cambridge University Engineering Dept, Cambridge, UK, 2001.

[37] T. Hain and P. Woodland. Dynamic HMM selection for continuous speech recognition. In *Proceedings of Eurospeech*, pages 532–535, 1999.

[38] T. Holter and T. Svendsen. Combined optimization of baseforms and model parameters in speech recognition based on acoustic sub-word units. In *Proceedings of IEEE Workshop ASRU*, pages 199–206, 1997.

[39] T. Holter and T. Svendsen. Maximum likelihood modelling of pronunciation variation. In *ESCA Workshop on Modelling Pronunciation Variation for Automatic Speech Recognition*, pages 63–66, 1998.

[40] X. Huang, A. Acero, and H.-W. Hon. *Spoken Language Processing.* Prentice Hall PTR, 2001.

[41] R. Jakobson, G. Fant, and M. Halle. *Preliminaries to Speech Analysis.* MIT Acoustics Laboratory, Cambridge, MA, 1952.

[42] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. Automatic detection of discourse structure for speech recognition and understanding. In *Proceedings of IEEE Workshop on Speech Recognition and Understanding*, pages 88–95, 1997.

[43] D. Jurafsky, R. Bates, N. Coccaro, R. Martin, M. Meteer, K. Ries, E. Shriberg, A. Stolcke, P. Taylor, and C. Van Ess-Dykema. SWBD discourse language modeling project final project report. Technical report, Center for Language and Speech Processing, Johns Hopkins University, 1997.

[44] D. Jurafsky, A. Bell, E. Fosler-Lussier, C. Girand, and W. Raymond. Reduction of English function words in Switchboard. In *Proceedings of ICSLP*, pages VII–3111–3114, 1998.

[45] D. Jurafsky, E. Shriberg, and D. Biasca. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. Technical Report 97-02, University of Colorado Institute of Cognitive Science, 1996.

[46] S. King, T. Stephenson, S. Isard, P. Taylor, and A. Strachan. Speech recognition via phonetically featured syllables. In *Proceedings of ICSLP*, pages 1031–1034, 1998.

[47] K. Kirchhoff. *Robust Speech Recognition Using Articulatory Information.* PhD thesis, University of Bielefeld, Germany, 1999.

[48] K. Kirchhoff and S. Parandekar. Multi-stream statistical N-gram modeling with application to automatic language identification. In *Proceedings of Eurospeech*, pages 803–806, 2001.

[49] Entropic Research Laboratory. *ESPS Version 5.0 Programs Manual.* StatSci, 1993.

[50] D.R. Ladd. *Intonational Phonology.* Cambridge University Press, 1996.

[51] P. Ladefoged. *A Course in Phonetics.* Harcourt Brace, 4th edition, 2001.

[52] A. Lahiri. Speech recognition with phonological features. In *Proceedings of Int. Congress on Phonetic Sciences*, pages 715–718, 1999.

[53] M. Liberman and C. Cieri. The creation, distribution and use of linguistic data. In *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, Spain, 1998.

[54] B. Lindblom. *Speech Production and Speech Modelling*, chapter Explaining Phonetic Variation: A Sketch of the H&H Theory, pages 403–439. Kluwer Academic Publishers, 1990.

[55] K. Livescu. Hidden feature models for speech recognition using dynamic Bayesian networks. In *Proceedings of Eurospeech*, pages 2529–2532, 2003.

[56] Y. Lobacheva. *Discourse Mixture Language Modeling.* Master's thesis, Boston University, 2000.

[57] K. Ma, G. Zavaliagkos, and R. Iyer. Pronunciation modeling for large vocabulary conversational speech recognition. In *Proceedings of ICSLP*, pages VI–2455–2458, 1998.

[58] S. Manuel. Recovery of "deleted" schwa. In *PERILUS XIV*, pages 115–118. 1991.

[59] A. Martin and M. Przybocki. The 2001 NIST evaluation for recognition of conversational speech over the telephone. In *Proceedings of the LVCSR Workshop*, 2001.

[60] D. McAllaster, L. Gillick, F. Scattone, and M. Newman. Fabricating conversational speech data with acoustic models: A program to examine model-data mismatch. In *Proceedings of ICSLP*, pages 1847–1850, 1998.

[61] M. Mohri, F. Pereira, and M. Riley. The design principles of a weighted finite state transducer library. *Theoretical Computer Science*, 231:17–32, 2000.

[62] M. Mohri, F. Pereira, and M. Riley. Weighted finite-state transducers in speech recognition. *Computer Speech and Language*, 16:69–88, 2002.

[63] N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proceedings of ICASSP*, pages II–729–732, 1998.

[64] NIST. Automatic meeting transcription project. `http://www.nist.gov/speech/test_beds/mr_proj/`, 2001.

[65] H. Nock. *Techniques for Modelling Phonological Processes in Automatic Speech Recognition.* PhD thesis, Cambridge University Engineering Dept, Cambridge, UK, 2001.

[66] H. Nock and S. Young. Detecting and correcting poor pronunciations for multi-word units. In *ESCA Workshop on Modelling Pronunciation Variation for Automatic Speech Recognition*, pages 85–90, 1998.

[67] H. Nock and S. Young. Loosely-coupled HMMs for ASR. In *Proceedings of ICSLP*, pages III:143–146, 2000.

[68] J. Odell. *The Use of Context in Large Vocabulary Speech Recognition.* PhD thesis, Cambridge University Engineering Dept, Cambridge, UK, 1995.

[69] 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. Technical Report ECE-97-0002, Boston University, 1997.

[70] M. Ostendorf, C. Wightman, and N. Veilleux. Parse scoring with prosodic information: an analysis/synthesis approach. *Computer Speech and Language*, 7:193–210, 1993.

[71] D. Pallet, J. Fiscus, W. Fisher, J. Garofolo, B. Lund, A. Martin, and M. Przybocki. 1994 benchmark tests for the ARPA spoken language program. In *Proceedings of ARPA Workshop on Spoken Language Technology*, pages 5–36, 1995.

[72] D. Pallett, J. Fiscus, G. Garofolo, A. Martin, and M. Przybocki. Broadcast news benchmark test results. In *Proceedings of DARPA Broadcast News Workshop*, 1999.

[73] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. *Journal of the Acoustical Society of America*, 90(6):2956–2970, 1991.

[74] E. Prince. Toward a taxonomy of given-new information. In P. Cole, editor, *Radical Pragmatics.* Academic Press, 1981.

[75] L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of IEEE*, 77(2):257–285, 1989.

[76] A. Ratnaparkhi. A maximum entropy part-of-speech tagger. In *Proceedings of the Empirical Methods in Natural Language Processing Conference*, pages 133–141, 1996.

[77] H. Reetz. Converting speech signals to phonological features. In *Proceedings of Int. Congress on Phonetic Sciences*, volume 3, pages S. 1733–1736, 1999.

[78] M. Riley, W. Byrne, M. Finke, S. Khudanpur, A. Ljolje, J. McDonough, H. Nock, M. Saraclar, C. Wooters, and G. Zavaliagkos. Stochastic pronunciation modelling from hand-labelled phonetic corpora. *Speech Communication*, 29:209–224, 1999.

[79] M. Riley and A. Ljolje. Recognizing phonemes vs. recognizing phones: A comparison. In *Proceedings of ICSLP*, pages 285–288, 1992.

[80] M. Riley and A. Ljolje. Automatic generation of detailed pronunciation lexicons. In C.-H. Lee, F. Soong, and K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, chapter 1, pages 1–17. Kluwer Academic Press, 1996.

[81] K. Ross and M. Ostendorf. Prediction of abstract prosodic labels for speech synthesis. *Computer Speech and Language*, 10:155–185, 1996.

[82] E. Sagey. *The Representation of Features in Non-Linear Phonology: The Articulator Node Hierarchy.* Taylor and Francis, 1991.

[83] M. Saraclar. *Pronunciation Modeling for Conversational Speech Recognition.* PhD thesis, The Johns Hopkins University, MD, USA, 2000.

[84] M. Saraclar, H. Nock, and S. Khudanpur. Pronunciation modeling by sharing Gaussian densities across phonetic models. *Computer Speech and Language*, 14(2):137–160, 2000.

[85] R. Schwartz, L. Nguyen, and J. Makhoul. Multiple-pass search strategies. In C.-H. Lee, F. Soong, and K. Paliwal, editors, *Automatic Speech and Speaker Recognition*, pages 429–456. Kluwer Academic Press, 1996.

[86] Z. Shafran. *Clustering Wide-Contexts and HMM Topologies for Spontaneous Speech Recognition.* PhD thesis, University of Washington, 2001.

[87] Z. Shafran and M. Ostendorf. Use of higher level linguistic structure in acoustic modeling for speech recognition. *Proceedings of ICASSP*, 2:1021–1024, 2000.

[88] S. Shattuck-Hufnagel, 2002. Personal Communication.

[89] S. Shattuck-Hufnagel and A. Turk. A prosody tutorial for investigators of auditory sentence processing. *J. Psycholinguistic Research*, 25(2):193–247, 1996.

[90] E. Shriberg, R. Bates, and A. Stolcke. A prosody-only decision-tree model for disfluency detection. In *Proceedings of Eurospeech*, pages 2383–2386, 1997.

[91] E. Shriberg and A. Stolcke. Prosody modeling for automatic speech understanding: An overview of recent research at SRI. In *Proceedings of the Workshop on Prosody in Speech Recognition and Understanding*, pages 13–16, 2001.

[92] E. Shriberg, A. Stolcke, and D. Baron. Can prosody aid the automatic processing of multi-party meetings? Evidence from predicting punctuation, disfluencies, and overlapping speech. In *Proceedings of the Workshop on Prosody in Speech Recognition and Understanding*, pages 139–146, 2001.

[93] H. Shu and I. L. Hetherington. EM training of finite-state transducers and its application to pronunciation modeling. In *Proceedings of ICSLP*, pages 1293–1296, 2002.

[94] K. Silverman, M. Beckman, J. Pitrelli, M. Ostendorf, *et al.* ToBI: A standard for labeling English prosody. In *Proceedings of ICSLP*, pages 867–870, 1992.

[95] K. Sönmez, M. Plauché, E. Shriberg, and H. Franco. Consonant discrimination in elicited and spontaneous speech: A case for signal-adaptive front ends in ASR. In *Proceedings of ICSLP*, pages 548–551, 2000.

[96] K. Sönmez, E. Shriberg, L. Heck, and M. Weintraub. Modeling dynamic prosodic variation for speaker verification. In *Proceedings of ICSLP*, pages 3189–3192, Sydney, Australia, 1998.

[97] *S-PLUS: Guide to Statistics and Mathematical Analysis, Version 3.2.* StatSci, 1995.

[98] K. Stevens. *Acoustic Phonetics.* The MIT Press, 1998.

[99] A. Stolcke. SRILM – The SRI language modeling toolkit. `http://www.speech.sri.com/projects/srilm/`.

[100] A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauché, G. Tür, and Y. Lu. Automatic detection of sentence boundaries and disfluencies based on recognized words. In *Proceedings of ICSLP*, pages VII–2247–2250, 1998.

[101] S. Stüker, F. Metze, T. Schultz, and A. Waibel. Integrating multilingual articulatory features into speech recognition. In *Proceedings of Eurospeech*, pages 1033–1036, 2003.

[102] G. Tajchman, E. Fosler, and D. Jurafsky. Building multiple pronunciation models for novel words using exploratory computational phonology. In *Proceedings of Eurospeech*, pages 2247–2250, 1995.

[103] D. Talkin. Pitch tracking. In W. Kleijn and K. Paliwal, editors, *Speech Coding and Synthesis.* Elsevier Science B.V., 1995.

[104] P. Taylor. The tilt intonation model. In *Proceedings of ICSLP*, 1998.

[105] P. Taylor. Analysis and synthesis of intonation using the tilt model. *JASA*, 107(3):1697–1714, 2000.

[106] J. t'Hart and A. Cohen. Intonation by rule: a perceptual quest. *J. Phonetics*, 1:309–327, 1973.

[107] N. Veilleux and S. Shattuck-Hufangel. Phonetic modification of the syllable /tu/ in two spontaneous American English dialogues. In *Proceedings of ICSLP*, 1998.

[108] Y. Wakita, H. Singer, and Y. Sagisaka. Multiple pronunciation dictionary using HMM-state confusion characteristics. *Computer Speech and Language*, 13:143–153, 1999.

[109] F. Weber, L. Manganaro, B. Peskin, and E. Shriberg. Using prosodic and lexical information for speaker identification. In *Proceedings of ICASSP*, pages I–141–144, 2002.

[110] M. Weintraub, E. Fosler, C. Galles, K. Yu-Hung, S. Khudanpur, M. Saraclar, and S. Wegmann. Automatic learning of word pronunciation from data. Technical Report 24, Center for Language and Speech Processing, Johns Hopkins University, 1997.

[111] M. Weintraub, K. Taussig, K. Hunicke-Smith, and A. Snodgrass. Effect of speaking style on LVCSR performance. In *Proceedings of ICSLP*, pages S16–S19 (addendum), 1996.

[112] C. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Trans. on Speech and Audio Processing*, 2(4):469–481, 1994.

[113] S. Witt. *Use of Speech Recognition in Computer-Assisted Language Learning.* PhD thesis, Cambridge University Engineering Department, Cambridge, UK, 1999.

[114] P. Woodland, G. Evermann, M. Gales, T. Hain, A. Liu, G. Moore, D. Povey, and L. Wang. CU-HTK April 2002 Switchboard system. In *Notebook Proceedings of the Rich Transcription Workshop*, 2002. `http://svr-www.eng.cam.ac.uk/reports/svr-ftp/woodland_rt02.pdf`.

[115] P. Woodland, J. Odell, V. Valtchev, and S. Young. Large vocabulary continuous speech recognition using HTK. In *Proceedings of ICASSP*, pages II–125–128, 1994.

[116] S.-L. Wu, M. Shire, S. Greenberg, and N. Morgan. Integrating syllable boundary information into speech recognition. In *Proceedings of ICASSP*, pages 987–990, 1997.

[117] S. Young, J. Odell, and P. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings of the ARPA Human Language Technology Conference*, pages 307–312, 1994.

[118] S. Young and P. Woodland. The use of state tying in continuous speech recognition. In *Proceedings European Conference on Speech Communication and Technology*, volume 3, pages 2203–2206, 1993.

Appendix A

PHONEME TABLES

This appendix contains five sections. The first is a mapping of the phone symbols used in this work to their corresponding International Phonetic Alphabet (IPA) symbols. The second gives the articulatory features that define the phones used in this work. The third is a representation of the phonetic feature distance used for alignments and as a measure of pronunciation quality. The fourth is a set of statistics describing baseform phone transformations with respect to the phonetic distance measure. The final section details the phone tranformation types used for intermediate prediction.

A.1 IPA/ARPABET Phoneme Maps

For reference, a mapping of the ARPABET symbols used in this work and the International Phonetic Alphabet (IPA) is provided in Table A.1.

A.2 Phoneme Feature Sets

The first four tables in this section, Tables A.2, A.3, A.4 and A.5, are the articulatory features bundles that describe the phones used in this work. The features are described in Section 3.2. "+" means that the feature is on. "–" means the feature is off. "X" signifies the feature is not applicable for the phone. "0" means that the feature could be on or off for the phone. It is not used for distinguishing between produced phones but it can take on a value based on its phonetic context. The features are adapted from [98]. In Chapter 5, the features *coronal, anterior, distributed, lateral,* and *rhotic* for the phones /hh/ and /q/ had been "X" for the feature distance but were changed to "0" for Chapter 7 because the features are supraglottal and therefore variable rather than not applicable for these phones primarily distinguished by the glottis. While the "X" feature values worked

for the purposes of a creating a distance measure, prediction of "0" values seemed to better represent what was actually possible during production. Similarly, for the flaps (/dx/ and /nx/), the features used in Chapter 5 were "–" for *strident*, *advanced* and *constricted tongue root* rather than "X" for *strident* and 0 for the tongue root features.

This work uses the ICSI hand-labeled portion of the Switchboard corpus (described in Chapter 4). All phones used in the hand-labeling can be mapped to a particular set of articulatory features. However, the hand-labeling also includes a set of diacritics indicating changes in the phones that can be described as feature changes. The transformation rules for most diacritics are in Table A.6 and the distribution of diacritics in the data (Table A.7). Table A.8 shows the effects on the feature sets when phones (typically stop consonants) are labeled with the diacritic "ap" meaning "approximate articulation". The occurences of these phones become more like flaps than the original consonants. The rules for the transformations are:

consonantal:	+ to −
sonorant:	− to +
continuant:	− to +
delayed release:	− to +
voicing:	− to +
advanced tongue root:	X to 0
constricted tongue root:	X to 0
round:	− to 0

Table A.1: *ARPABET and IPA symbols with examples. Phones used in the speech recognition system are in bold. Examples marked with * would occur during fast speech.*

ARPABET	IPA	Example	ARPABET	IPA	Example
iy	i	s**ee**n	**ih**	ɪ	s**i**n
ey	e	h**ay**	**eh**	ɛ	r**e**d
ae	æ	b**a**t	**aa**	ɑ	d**o**p
ao	ɔ	c**o**ffee	**ow**	o	m**oa**n
ah	ʌ	m**u**d	**uw**	u	**oo**ze
uh	ʊ	w**oo**d	ux	ʉ	y**ou***
ax	ə	b**a**n**a**	**axr**	ɚ	nutt**er**
ix	ɨ	p**o**tato*	**aw**	ɑw	c**ou**ch
ay	ɑj	**I**	**oy**	ɔj	t**oy**
h	h	**h**am	**w**	w	**w**eave
y	y	**y**amaguchi	**r**	ɹ	**r**um
dx	ɾ	bu**tt**er*	q	ʔ	bu**tt**on
nx	ɾ^{n}	ya k**n**ow*	hw	ʍ	**wh**ether
lg	ɬ	re**s**ult*			
l	l	**l**utz	**el**	l̩	pood**le**
m	m	**m**a**m**a	em	m̩	the**m***
n	n	**n**ose	**en**	n̩	garde**n**
ng	ŋ	si**ng**	eng	ŋ̩	buyi**ng***
v	v	**v**ine	**f**	f	**f**ine
dh	ð	ei**th**er	**th**	θ	e**th**er
z	z	**z**oo	**s**	s	**s**ue
zh	ʒ	a**z**ure	**sh**	ʃ	**sh**ut
b	b	**b**oy	**p**	p	**p**ine
d	d	**d**aisy	**t**	t	**t**onic
g	g	**g**irl	**k**	k	**k**ite
jh	ʤ	**g**in	**ch**	ʧ	**ch**in

Table A.2: *A: Standard feature sets for vowels.*

ARPABET symbol	iy	ih	ey	eh	ae	aa	ao	ow	ah	uw	ux	uh	axr	ax	ix
syllabic	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
consonantal	–	–	–	–	–	–	–	–	–	–	–	–	+	–	–
sonorant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
continuant	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
strident	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
delayed release	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
voicing	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
spread glottis	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
constricted glottis	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
advanced tongue root	+	–	+	–	–	–	–	+	–	+	+	–	–	–	–
constricted tongue root	–	–	–	–	–	+	+	–	–	–	–	–	–	–	–
nasal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
dorsal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
coronal	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0
labial	–	–	–	–	–	–	+	+	–	+	+	+	–	–	–
high	+	+	–	–	–	–	–	–	–	+	+	+	0	–	0
low	–	–	–	–	+	+	+	–	–	–	–	–	0	–	–
back	–	–	–	–	–	+	+	+	+	+	–	+	0	0	–
anterior	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0
distributed	0	0	0	0	0	0	0	0	0	0	0	0	–	0	0
lateral	0	0	0	0	0	0	0	0	0	0	0	0	–	0	0
rhotic	0	0	0	0	0	0	0	0	0	0	0	0	+	0	0
round	–	–	–	–	–	–	+	+	–	+	+	+	0	0	0

Table A.3: *B: Standard feature sets for diphthongs and glides.*

ARPABET symbol	aw		ay		oy		hh	w	y	r	dx	q	nx	hw	lg
syllabic	+	−	+	−	+	−	−	−	−	−	−	−	−	−	−
consonantal	−	−	−	−	−	−	−	−	−	−	−	−	−	−	−
sonorant	+	+	+	+	+	+	−	+	+	+	+	−	+	+	+
continuant	+	+	+	+	+	+	+	+	+	+	+	−	+	+	−
strident	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
delayed release	+	+	+	+	+	+	+	+	+	+	+	−	+	+	+
voicing	+	+	+	+	+	+	−	+	+	+	+	−	+	−	+
spread glottis	−	−	−	−	−	−	+	−	−	−	−	−	−	+	−
constricted glottis	−	−	−	−	−	−	−	−	−	−	−	+	−	−	−
advanced tongue root	−	−	−	−	−	−	0	+	+	+	0	0	0	+	X
constricted tongue root	+	−	+	−	+	−	0	−	−	−	0	0	0	−	X
nasal	0	0	0	0	0	0	−	−	−	−	−	−	+	−	−
dorsal	+	+	+	+	+	+	0	+	+	−	−	0	−	+	+
coronal	0	−	0	−	0	−	0	−	−	+	+	0	+	−	+
labial	−	+	−	−	+	−	0	+	−	−	−	0	−	+	−
high	−	+	−	+	−	+	0	+	+	X	X	0	X	+	+
low	+	−	+	−	+	−	0	−	−	X	X	0	X	−	−
back	+	+	+	−	+	−	0	+	−	X	X	0	X	+	+
anterior	0	X	0	X	0	X	0	X	X	+	+	0	+	X	+
distributed	0	X	0	X	0	X	0	X	X	−	−	0	−	X	−
lateral	0	X	0	X	0	X	0	X	X	−	−	0	−	X	+
rhotic	0	X	0	X	0	X	0	X	X	+	−	0	−	X	−
round	−	+	−	−	+	−	0	+	0	0	0	0	0	+	0

Table A.4: *C: Standard feature sets for consonants and corresponding syllabic consonants.*

ARPABET symbol	l	m	n	ng	el	em	en	eng
syllabic	–	–	–	–	+	+	+	+
consonantal	+	+	+	+	+	+	+	+
sonorant	+	+	+	+	+	+	+	+
continuant	–	–	–	–	+	+	+	+
strident	X	X	X	X	X	X	X	X
delayed release	+	–	–	–	+	–	–	–
voicing	+	+	+	+	+	+	+	+
spread glottis	–	–	–	–	–	–	–	–
constricted glottis	–	–	–	–	–	–	–	–
advanced tongue root	X	X	X	X	–	–	–	–
constricted tongue root	X	X	X	X	–	–	–	–
nasal	–	+	+	+	–	+	+	+
dorsal	–	–	–	+	–	–	–	+
coronal	+	–	+	–	+	–	+	–
labial	–	+	–	–	–	+	–	–
high	X	X	X	+	0	0	0	+
low	X	X	X	–	0	0	0	–
back	X	X	X	+	0	0	0	+
anterior	+	X	+	X	+	X	+	X
distributed	–	X	–	X	–	X	–	X
lateral	+	–	–	X	+	–	–	X
rhotic	–	–	–	X	–	–	–	X
round	0	–	0	0	0	–	0	0

Table A.5: *D: Standard feature sets for remaining consonants.*

ARPABET symbol	v	dh	z	zh	f	th	s	sh	b	d	g	p	t	k	jh	ch
syllabic	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
consonantal	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+	+
sonorant	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
continuant	+	+	+	+	+	+	+	+	–	–	–	–	–	–	–	–
strident	+	–	+	+	+	–	+	+	X	X	X	X	X	X	+	+
delayed release	+	+	+	+	+	+	+	+	–	–	–	–	–	–	+	+
voicing	+	+	+	+	–	–	–	–	+	+	+	–	–	–	+	–
spread glottis	–	–	–	–	+	+	+	+	–	–	–	+	+	+	–	+
constricted glottis	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
advanced tongue root	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
constricted tongue root	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
nasal	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
dorsal	–	–	–	–	–	–	–	–	–	–	+	–	–	+	–	–
coronal	–	+	+	+	–	+	+	+	–	+	–	–	+	–	+	+
labial	+	–	–	–	+	–	–	–	+	–	–	+	–	–	–	–
high	X	X	X	X	X	X	X	X	X	X	+	X	X	+	X	X
low	X	X	X	X	X	X	X	X	X	X	–	X	X	–	X	X
back	X	X	X	X	X	X	X	X	X	X	+	X	X	+	X	X
anterior	X	+	+	–	X	+	+	–	X	+	X	X	+	X	–	–
distributed	X	+	–	+	X	+	–	+	X	–	X	X	–	X	+	+
lateral	X	–	–	–	X	–	–	–	X	–	X	X	–	X	–	–
rhotic	X	–	–	–	X	–	–	–	X	–	X	X	–	X	–	–
round	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–

Table A.6: *Feature transformation rules for diacritics used in the ICSI hand-labeled set. (Note that the frication change for vowels was not applied here since less than 0.2% of all vowels had this diacritic.)*

Diacritic	Meaning	Transformation
n	nasalization of phone	*nasal* to +
vd	voicing of voiceless phone	*voicing* to +
vl	devoicing of voiced phone	*voicing* to –
on	trace of phone in syllable onset	segment ignored
co	trace of phone in syllable code	segment ignored
cr	creaky voice (marked only if contrastive)	no feature changed
epi	epenthetic stop	no feature changed
fr	frication of non-fricated phone	if /l/ or /r/: *voicing* to –
		if /b/, /d/, /g/, /p/, /t/, or /k/: *continuant* to +
		if vowel: *voicing* to –
ap	approximate articulation	phone becomes flap-like (see Table A.8)

Table A.7: *Counts of diacritic labels in ICSI training and held out sets.*

Diacritic	Train Set	Held Out
Phones	127,633	15,327
n	1951	840
vd	270	23
vl	1800	144
cr	1906	295
epi	151	8
fr	414	67
ap	475	5

Table A.8: *E: Feature sets for consonants marked with diacritic "ap" (i.e., approximate).*

ARPABET symbol	v	dh	z	th	b	d	g	p	t	k	jh
syllabic	–	–	–	–	–	–	–	–	–	–	–
consonantal	–	–	–	–	–	–	–	–	–	–	–
sonorant	+	+	+	+	+	+	+	+	+	+	+
continuant	+	+	+	+	+	+	+	+	+	+	+
strident	+	–	+	–	X	X	X	X	X	X	+
delayed release	+	+	+	+	+	+	+	+	+	+	+
voicing	+	+	+	+	+	+	+	+	+	+	+
spread glottis	–	–	–	+	–	–	–	+	+	+	–
constricted glottis	–	–	–	–	–	–	–	–	–	–	–
advanced tongue root	0	0	0	0	0	0	0	0	0	0	0
constricted tongue root	0	0	0	0	0	0	0	0	0	0	0
nasal	–	–	–	–	–	–	–	–	–	–	–
dorsal	–	–	–	–	–	–	+	–	–	+	–
coronal	–	+	+	+	–	+	–	–	+	–	+
labial	+	–	–	–	+	–	–	+	–	–	–
high	X	X	X	X	X	X	+	X	X	+	X
low	X	X	X	X	X	X	–	X	X	–	X
back	X	X	X	X	X	X	+	X	X	+	X
anterior	X	+	+	+	X	+	X	X	+	X	–
distributed	X	+	–	+	X	–	X	X	–	X	+
lateral	X	–	–	–	X	–	X	X	–	X	–
rhotic	X	–	–	–	X	–	X	X	–	X	–
round	0	0	0	0	0	0	0	0	0	0	0

A.3 Feature-based Distance Measure

Figure A.1 is a representation of the costs associated with substituting one phone for another between the hand-labeled ICSI data set and canonical pronunciations.

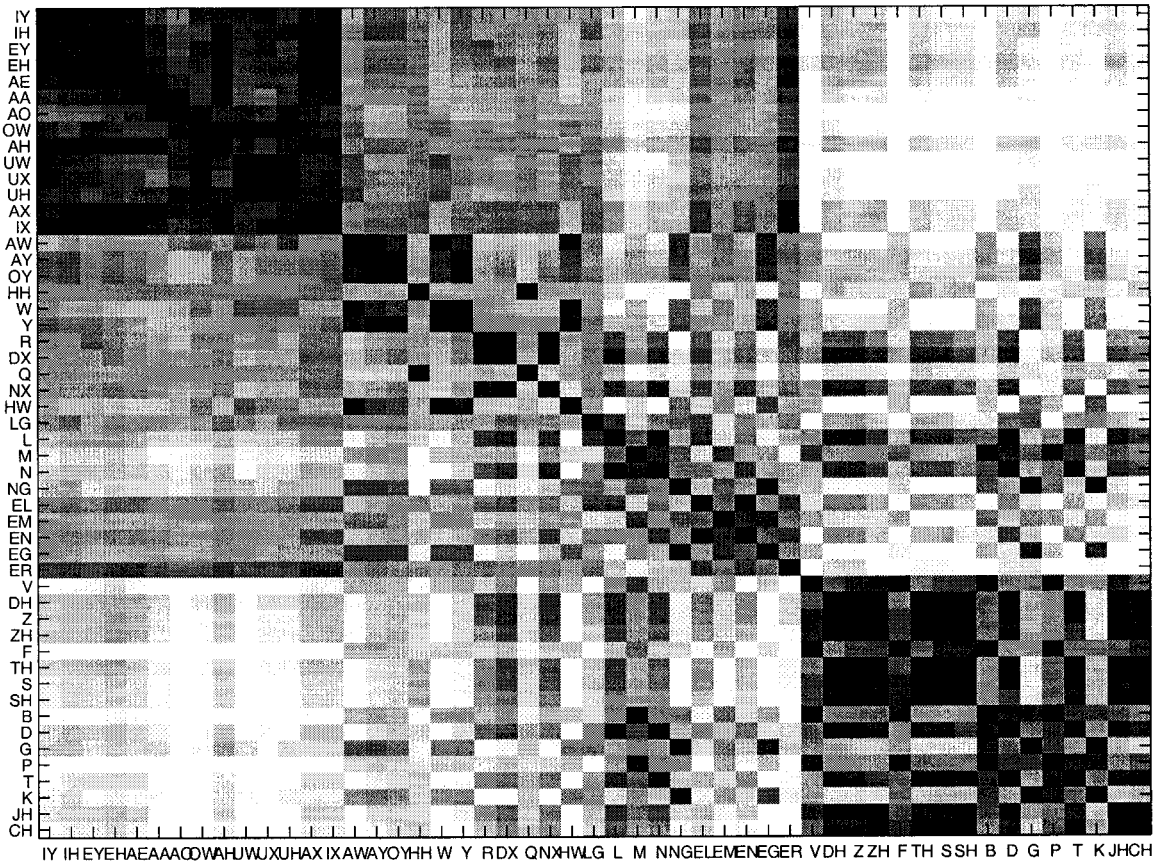


Figure A.1: Feature-based distance measure. Minimum value is 0 (black), maximum value is 76 (white).

A.4 Baseform Phone Statistics

Table A.9 lists the phones from the dictionary and shows the number of times they occur in the ICSI hand-labeled data set (Count). Since /ux/, /ix/, /q/, /nx/, /hw/, /lg, /em/, /eng/, are not in the dictionary, their presence is represented only as phone substitutions for dictionary phones. While /el/, /en/, and /zh/ are in the dictionary, there were no instances in the ICSI set. /axr/ and /er/ are merged for these statistics. Each baseform phone is mapped to the hand-labeled surface form phone with a given phonetic distance. The maximum distance between the baseform and surface forms is given (Max) along with the percentage that phone is not transformed to another representation. The mean and standard deviation (Mean, SD) of the cost when the phone is transformed completes the table. The minimum cost distance for all phones is 0, indicating no phone change.

Table A.9: *Phone distance statistics for baseform phones.*

Phone	Count	Max	Percent	With Change	
			Not Changed	Mean	SD
iy	3017	60	14.2	5.07	4.01
ih	5315	58	21.6	5.77	5.52
ey	2158	32	9.5	6.38	4.50
eh	3845	52	23.5	5.77	5.60
ae	4805	52	28.2	5.95	3.68
aa	7091	60	13.0	9.58	6.78
ao	1751	44	26.7	7.90	5.63
ow	2921	60	13.9	11.02	5.68
ah	3679	54	20.6	6.70	5.63
uw	2601	56	50.0	8.84	5.13
uh	643	32	22.1	13.06	5.83
ax	1801	48	58.6	4.50	5.75
aw	733	64	2.5	31.89	16.04
ay	3202	50	3.0	23.73	10.12
oy	106	24	2.8	15.33	8.38
er	928	32	8.4	19.03	5.55

Table A.10: *Phone distance statistics for baseform phones (cont.).*

Phone	Count	Max	Percent	With Change	
			Not Changed	Mean	SD
hh	1391	52	1.0	36.71	14.53
w	3245	42	1.2	23.95	13.95
y	1905	60	5.4	34.43	17.39
r	2786	50	1.7	21.83	13.31
dx	301	52	19.9	31.27	11.59
l	2612	62	9.7	24.20	7.66
m	1904	44	1.6	27.23	9.66
n	2340	52	19.8	18.65	4.39
ng	7	0	100	NA	NA
v	846	40	1.7	17.73	14.53
dh	3752	58	10.2	17.58	7.80
z	289	16	3.5	9.60	4.45
f	1445	62	0.6	25.00	21.14
th	763	50	4.7	13.28	8.64
s	3235	28	1.1	9.50	4.04
sh	607	38	5.1	9.42	10.35
b	2134	44	0.9	20.32	12.30
d	2040	58	22.4	21.40	4.10
g	1286	38	0.4	19.20	15.37
p	1795	46	0.6	15.64	10.58
t	3476	66	16.9	23.37	8.62
k	2366	50	0.6	22.00	14.20
jh	563	54	14.4	11.31	9.50
ch	308	18	19.8	5.08	3.41
fp	1603	60	8.0	30.63	12.67

A.5 Phone Transformation Types

This work measures pronunciation variation in three different ways: word-level pronunciation distance, baseform to surface-form phone changes and phone transformation types. In the case of phone transformations, we group them based on the hypothesized mode of production into five categories. Phone changes are associated with either hyper-articulation or reduction phenomena. Phone insertions and substitutions of full vowels where a short or reduced vowel is expected suggest hyper-articulation. Phone deletions, substitutions of flaps, substitutions of reduced or short vowels for expected full vowels suggest reduction. In addition, there are other phone changes (e.g., from one full vowel to another) that could not be easily categorized as a reduction or hyper-articulation. Regional dialect, age or sociolect differences may be the source for many of these cases. The majority of phones match their dictionary baseform. The five categories are thus:

1. surface form matches baseform,
2. baseform phone deleted,
3. reduced version of baseform,
4. hyper-articulated version of baseform,
5. baseform changed but neither reduced nor hyper-articulated (other).

Baseform reductions occur when a full vowel is expected but a reduced one is said, when a flap (/nx/, /dx/, /q/) occurs, or when a voiced sound (e.g., /z/) is expected but an unvoiced phone (e.g., /s/) is said. We do not require that it is a voiced/unvoiced consonant pair for the transformation, but label all devoicing of phones as evidence of reduction. (See Table A.11 for sample pairs.) 94.1% are vowel reductions or flap substitutions. The remainder are devoiced consonants. Of this 5.9%, 64.6% are voiced/unvoiced phone pair distinctions. The phone categories used in determining the phone transformation types are shown in Table A.12. While a third class of vowels could have been used (i.e., classifying /ax/ and /ix/ as fully reduced), for simplicity, this work used only two categories. Hyper-articulated phones occur when a phone is inserted or when a reduced vowel such as /ax/ or /uh/ is in the dictionary but a full vowel like /ae/ is actually said. They are also labeled if a flap (/dx/) is indicated but a non-flap is said (/t/ or /d/). All other changes are categorized

Table A.11: *Voiced and unvoiced consonant pairs.*

Voiced Phone	Unvoiced Phone
/d/	/t/
/b/	/p/
/g/	/k/
/z/	/s/
/zh/	/sh/
/dh/	/th/
/v/	/f/
/jh/	/ch/

as a non-hyper-articulated, non-reduced change. We did not have a category representing assimilation, since our focus was on the hyper vs. hypo distinction. Unfortunately, by not explicitly representing this type of change, we may have had more phone changes labled with the "other" category than we would like.

Table A.12: *Phone categories used in determining phone transformation types.*

Category	Phone
Full Vowels and Diphtongs	/iy/, /ey/, /ae/, /aa/, /ow/, /uw/
Short Vowels	/ih/, /eh/, /ao/, /ah/, /uh/
Reduced Vowels	/ax/, /ix/, /ux/, /el/, /em/, /en/
Flaps	/dx/, /nx/, /q/
Voiced Consonants	/w/, /y/, /dx/, /r/, /nx/, /l/, /m/, /n/, /ng/, /v/, /dh/, /z/, /zh/, /jh/, /b/, /d/, /g/, /el/, /em/, /en/, /eng/, /lg/, /fp/
Unvoiced Consonants	/p/, /hh/, /q/, /f/, /th/, /s/, /sh/, /ch/, /t/, /k/, /hw/

Appendix B

SURFACE FORM PREDICTION TREES

Included here is a set of example trees for predicting surface form pronunciations from baseforms as developed in Chapter 5. For a given set of attributes, each baseform phone will have an individual tree predicting different surface form distributions conditioned on the questions asked in the tree. The trees presented here include questions about either text-based attributes alone or both prosodic and text-based features.

Included here are sample trees for /ae/ (Figure B.1), /ax/ (Figure B.2), /l/ (Figure B.3), and /t/ (Figure B.4). For ease of readability, the trees are pruned significantly to show the top nodes. Note that a wide range of attributes are used and that there are very different attributes used for the different phones. Prosodic attributes can change the entire structure of the tree or simply modify lower branches.

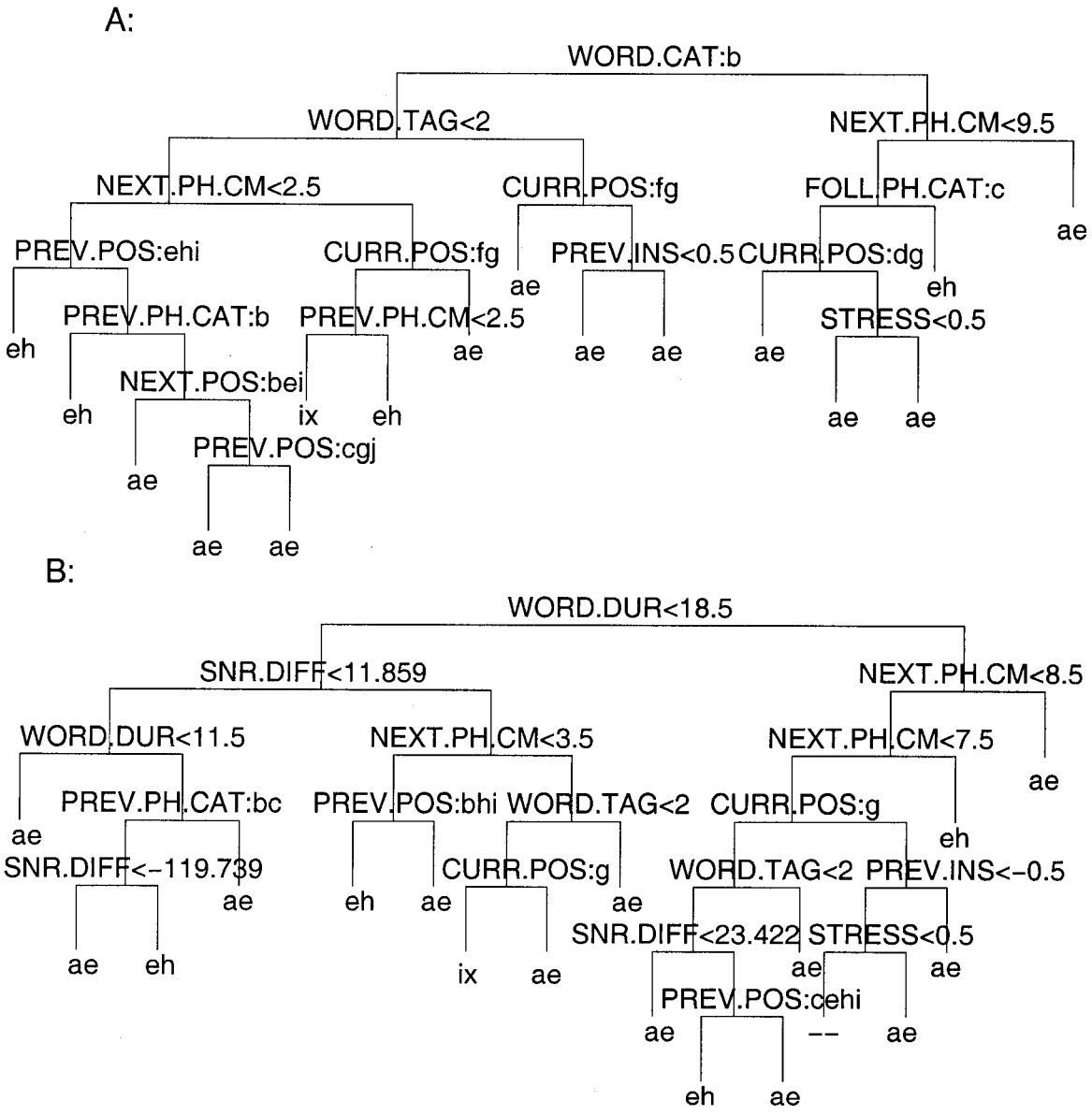


Figure B.1: Pruned (k=50) surface form prediction trees for /ae/. A: text based. B: prosody and text-based.

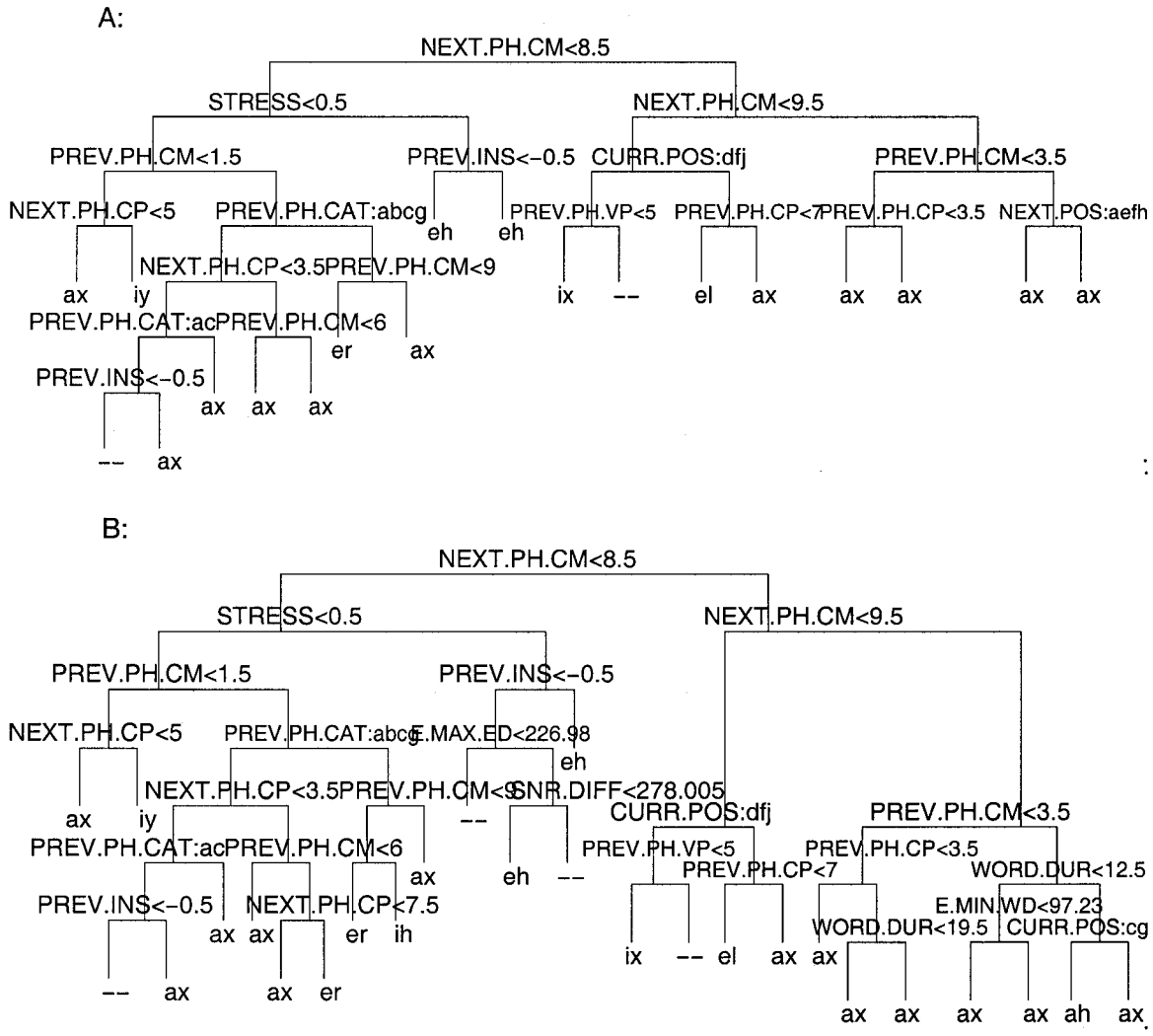


Figure B.2: Pruned (k=50) surface form prediction trees for /ax/. A: text based. B: prosody and text-based.

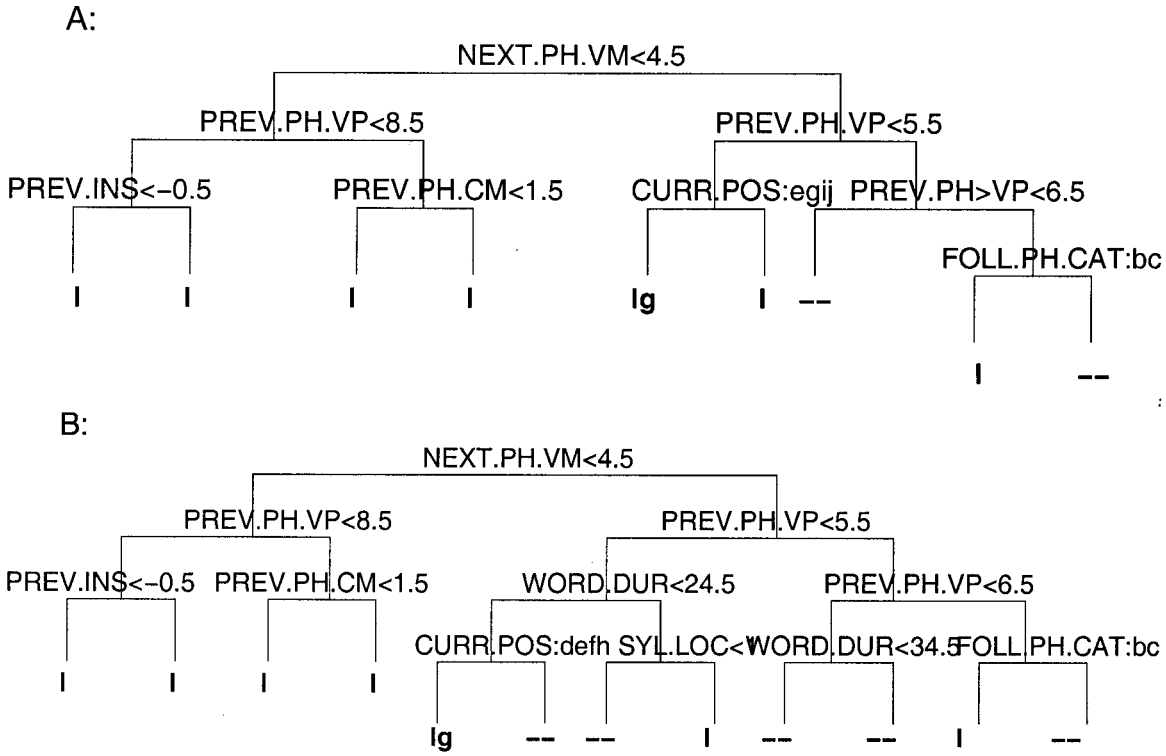


Figure B.3: Pruned (k=40) surface form prediction trees for /l/. A: text based. B: prosody and text-based.

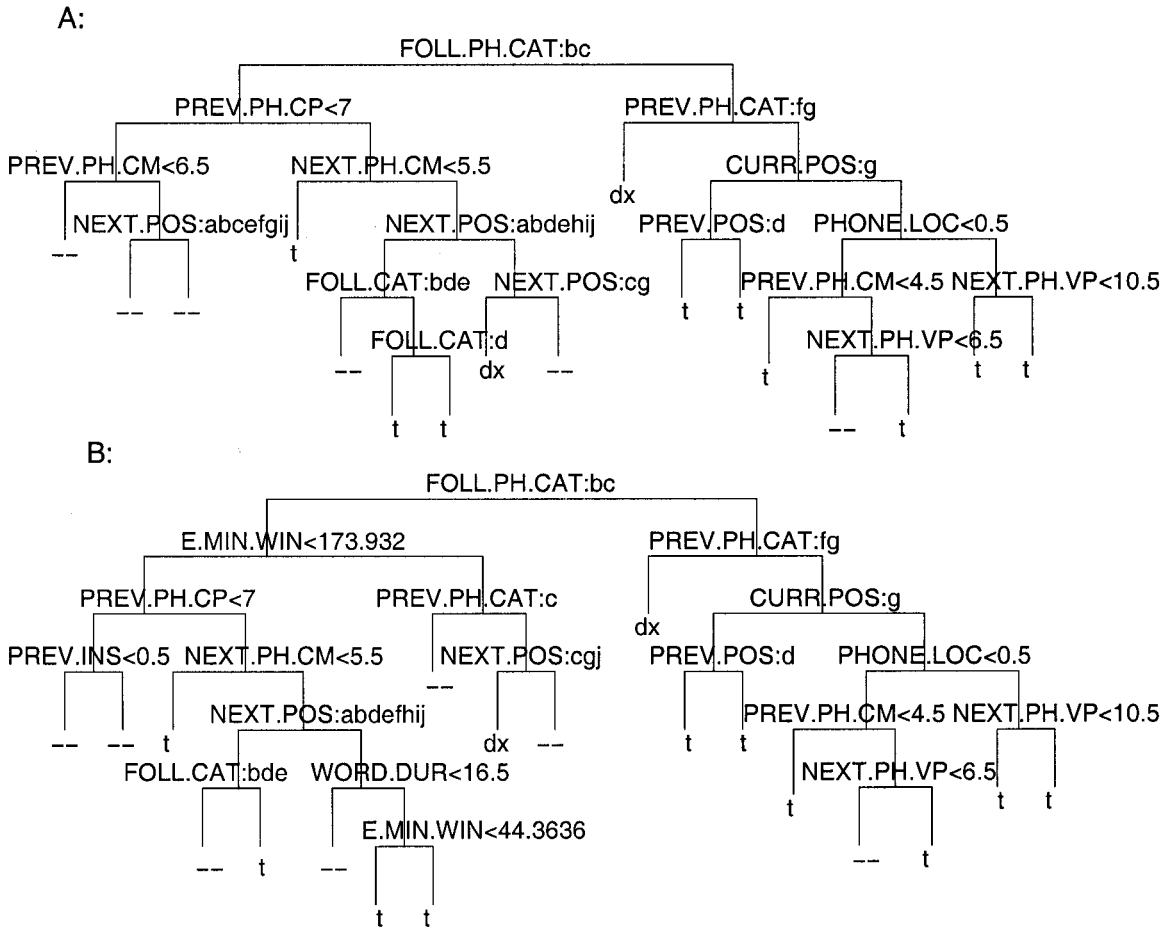


Figure B.4: Pruned (k=80) surface form prediction trees for /t/. A: text based. B: prosody and text-based.

VITA

Rebecca Anne Bates grew up in Great Falls, Montana, and received a B.S. in Biomedical Engineering from Boston University in 1990. She remained in the Boston area to study and received an M.T.S. (Theological Studies) from Harvard University in 1993, followed by an M.S. in Electrical Engineering from Boston University in 1996. She worked as a Research Engineer at SRI International, Menlo Park, CA for six months before returning to school for her Ph.D. and was a participant in the 1997 Summer Workshop on Speech Recognition at Johns Hopkins University. She is currently a professor of Computer and Information Sciences at Minnesota State University, Mankato. Current information about her publications can be found on the web.