INTERNATIONAL COMPUTER SCIENCE INSTITUTE

1947 Center St. • Suite 600 • Berkeley, California 94704-1198 • (510) 643-9153 • FAX (510) 643-7684

# Dynamic Pronunciation Models for Automatic Speech Recognition

**John Eric Fosler-Lussier**

**TR-99-015**

**September 1999**

## Abstract

As of this writing, the automatic recognition of spontaneous speech by computer is fraught with errors; many systems transcribe one out of every three to five words incorrectly, whereas humans can transcribe spontaneous speech with one error in twenty words or better. This high error rate is due in part to the poor modeling of pronunciations within spontaneous speech. This dissertation examines how pronunciations vary in this speaking style, and how speaking rate and word predictability can be used to predict when greater pronunciation variation can be expected. It includes an investigation of the relationship between speaking rate, word predictability, pronunciations, and errors made by speech recognition systems. The results of these studies suggest that for spontaneous speech, it may be appropriate to build models for syllables and words that can dynamically change the pronunciations used in the speech recognizer based on the extended context (including surrounding words, phones, speaking rate, etc.). Implementation of new pronunciation models automatically derived from data within the ICSI speech recognition system has shown a 4-5% relative improvement on the Broadcast News recognition task. Roughly two thirds of these gains can be attributed to static baseform improvements; adding the ability to dynamically adjust pronunciations within the recognizer provides the other third of the improvement. The Broadcast News task also allows for comparison of performance on different styles of speech: the new pronunciation models do not help for pre-planned speech, but they provide a significant gain for spontaneous speech. Not only do the automatically learned pronunciation models capture some of the linguistic variation due to the speaking style, but they also represent variation in the acoustic model due to channel effects. The largest improvement was seen in the telephone speech condition, in which 12% of the errors produced by the baseline system were corrected.

This technical report is a reprint of the dissertation of John Eric Fosler-Lussier filed with the University of California, Berkeley in Fall 1999.

Committee in charge:

Professor Nelson Morgan, Chair
Professor Jerry Feldman
Dr. Steven Greenberg
Professor John Ohala

# Contents

# List of Figures

# List of Tables

## Acknowledgments

In any engineering discipline, and particularly when working in the field of large-vocabulary speech recognition, finishing a dissertation would be nearly impossible without help from a great number of people.

First and foremost, I would like to thank my advisor, Nelson Morgan. His hard work has built a top-notch research group at ICSI of which it is a pleasure to be a part. Morgan has all of the qualities found in any good advisor: he knows when to say "That's interesting" and when to say "It's a bug," he promotes his students, and he teaches them the fine points of being good researchers. Most of all, he accomplishes all of this with a wonderful sense of humor. I am particularly glad that he chose not to trade me for a fancy cappuccino machine several years ago when the funding opportunity presented itself—so much so that unlike all of his other students to date, I will fail to mention his Rolaids habit in the acknowledgments to my dissertation.

The other members of my dissertation committee have also shaped my education at Berkeley, in addition to providing good feedback in the course of writing this document. Steve Greenberg turned me on to the relationship between information content and phonetic realization, as well as encouraging me to look beyond the phone for modeling pronunciations. His work in phonetically transcribing the Switchboard corpus has been invaluable to my research. Jerry Feldman, my first professorial contact at Berkeley when I visited in the spring of 1993, has taught me much about neural networks and language. John Ohala's view of phonetically-based phonology, particularly in using experimental data to find phonological processes, has made its mark in these pages. Jitendra Malik, in addition to serving on my qualifying exam committee, opened my eyes to the parallels between computer vision and computer speech recognition.

Several people were important in getting me started in graduate school. Mitch Marcus, my undergraduate advisor at the University of Pennsylvania, deserves credit on two counts: (a) for encouraging me to get a Ph.D. instead of a master's degree, and (b) for telling me to go to Berkeley at precisely the right moment (not to mention his involvement in getting me interested in computational linguistics in general). Mark Liberman, my linguistics advisor at U. Penn, showed me that there was more to computational linguistics than natural language processing through his phonetics courses. Shortly after I arrived at Berkeley, Gary Tajchman and Dan Jurafsky, both postdocs at ICSI, took me under their collective wing until they both "flew the coop." I continue to appreciate collaborating with Dan and his students in linguistic investigations.

I was very fortunate to have an excellent set of colleagues both in the Realization Group and in other groups at ICSI. Brian Kingsbury lifted a lot of weight off my shoulders in so many ways— as a friend, as a workout partner, and through his contributions to the ICSI speech recognition system. Nikki Mirghafori instilled in me the importance of stretching, as well as being a collaborator in early speaking rate studies at ICSI. The syllabic orientation of this thesis was inspired in part by the thesis of Su-Lin Wu, who fortunately shares my appreciation for ice cream. Dan Ellis, herder of the ICSI speech software and provider of useful comments on this thesis, did yeoman's work along with Adam Janin in training the ICSI recognition system on the Broadcast News corpus that is used in this thesis. I also thank David Johnson for insisting on good software engineering in both my work and

the group's work at large, Nancy Chang and Dan Gildea for helpful comments on drafts of this thesis, and Toshihiko Abe, Krste Asanović, David Bailey, Collin Baker, Jim Beck, Jeff Bilmes, Barry Chen, Philipp Faerber, Susanne Gahl, Ben Gomes, Alfred Hauenstein, Joy Hollenback, Rainer Klisch, Katrin Kirchhoff, Yochai Koenig, Hiroaki Ogawa, Srini Narayanan, Florian Schiel, Holger Schwenk, Mike Shire, Rosaria Silipo, Boris Weissman, Warner Warren, and Geoff Zweig for six years of wonderful interactions. Jane Edwards deserves a special thanks for pointing me to many of the references in chapter 1.

This work has also benefited extensively from connections outside the Realization Group. In the summer of 1996, I was fortunate enough to attend the Johns Hopkins Large Vocabulary Continuous Speech Recognition Summer Research Workshop; I thank in particular Mitch Weintraub and the rest of the automatic pronunciation learning group (Charlie Galles, Yu-Hung Kao, Sanjeev Khudanpur, Murat Saraclar, and Steve Wegmann) for providing the foundation of this work, as well as Fred Jelinek and all of the people at the 1996 Workshop for providing such a formative experience. The Realization Group worked with Gary Cook and Tony Robinson of Cambridge University, and Steve Renals, Yoshi Gotoh, and Gethin Williams of Sheffield University through the European SPRACH project to build the Broadcast News recognizer used in this thesis. I am particularly indebted to Gethin Williams for his collaboration on integrating confidence measures into pronunciation modeling. Hervé Bourlard and Hynek Hermansky, as external fellows to ICSI, also provided stimulating ideas in their many visits here; I have also benefited from the advice and wisdom of Liz Shriberg.

Thanks to Kathryn Crabtree for help navigating the bureaucratic minefield that is U.C. Berkeley, and to Renee Reynolds, Nancy Shaw, Lila Finhill, Devra Pollack, and Elizabeth Weinstein for making ICSI a great place to work.

Finishing a dissertation takes a certain amount of perseverance, a trait inherited, in my case, from my parents. I thank them for always working to expand my horizons, and delicately pushing me to achieve more. (Dad would always jokingly ask where the bonus points on my exams were—well, Dad, these pages *are* the bonus points.) I also thank my siblings, Pete, Beth, and Jess, for dealing with a brother who sometimes seemed just a bit too harried.

Finally, thanks to my spouse, Danielle, probably one of the most supportive people in the world. On so many levels, this dissertation could not have been completed without her aid. In particular, the English in this thesis would have suffered greatly without her editing skills.

Funding for my work has been provided by a number of sources, including a grant from the Center for Speech and Language Processing at Johns Hopkins University, NSF SGER Grant IRI-9713346, NSF Grant IRI-9712579, a European Community Basic Research Grant (Project SPRACH), and the International Computer Science Institute.

*For Danielle.*

# Chapter 1

# Introduction

> I would often notice that (that) in reporters' stories, the interviewees sounded way more interesting than the reporters. And it wasn't that the (rep-) reporters were bad writers or anything. It's because the interviewees were talking the way that people normally talk, whereas the reporters were simply reading from a script.... The thing that we respond to on the radio most readily is the sound of a real person talking, actually, really.
> — Ira Glass, interviewed by Terry Gross on National Public Radio's "Fresh Air," May 27, 1999

This thesis addresses issues surrounding the modeling of word pronunciations in both spontaneous and non-spontaneous speaking modes, particularly for use within an Automatic Speech Recognition (ASR) system. In particular, I investigate ways to automatically derive word pronunciations to capture pronunciation variability arising from differences in speaking style. I also examine the correlation of word pronunciations to various factors described in the next sections; incorporating these factors into an ASR pronunciation model may improve system performance. This chapter begins with a discussion of speaking mode; the latter part of the chapter is devoted to pronunciations and their modeling within ASR systems.

## 1.1 A matter of style

People use spoken language in a large number of ways: to give speeches, to tell stories, to talk to a friend on the phone, or to whisper good-night to a sleepy child. It is telling that English, like most languages, has many different verbs for the act of communication, including *spout off, gabble, harangue,* and *spill the beans.* The quotation from Ira

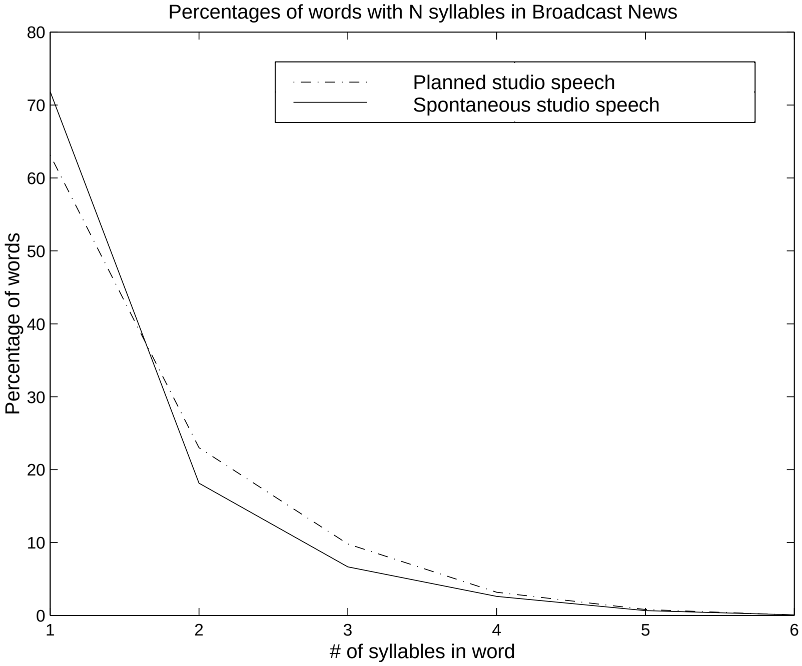


Figure 1.1: Percentage of words in the Broadcast News corpus with $n$ syllables for spontaneous and planned speaking styles.

Glass that starts this chapter suggests that humans react differently to the myriad types of speech found in the world. The different types of speech are often referred to as *speaking styles*, *speaking modes*, or *registers.*

Examining Glass's statement, we see that he makes a distinction between reading pre-written words and so-called normal speech. Using this dichotomy, Glass identifies two speaking styles: scripted and unscripted speech. The use of the word "normally" indicates that the unscripted mode of speech sounds more natural to him — a sentiment that is shared by most people.[1] Levin *et al.* [1982] found that subjects who were asked to distinguish between speakers telling stories and reading stories aloud were able to correctly do so 84% of the time (chance=50%). These two speaking modes are therefore perceptually distinct.

What is it that makes speaking styles different, particularly when spontaneous speech is compared with planned speech? Zwicky [1972b] warns that it is difficult to determine exactly what factors are involved:

> It is much easier to give clear examples of casual speech . . . than to say precisely what distinguishes casual speech from careful speech.

[1] What makes this quote even more interesting is the context in which it appears. At this point in the interview, Terry Gross asks why Glass's show sounds so extemporaneous, yet "every word is there for a reason, every sentence has an economy." Glass's response indicates that a third speaking style is really being used on his show that combines elements from the scripted and unscripted speaking modes.

Ayers [1994] also points out that "no single acoustic aspect of the signal conveys the spontaneity reliably." Despite these admonitions, there are differences in language usage that distinguish spontaneous and read speech. One factor is the choice of words. Glass, for instance, uses the word *way* in an informal manner ("interviewees sounded *way* more interesting..."), where a more formal speaker would probably use the word *far*. Spontaneous speakers also tend to reuse lexical items in their speech (*e.g.*, the four instances of *reporters* in the Glass quote) and choose monosyllabic words more often. Figure 1.1 compares monosyllabic word usage in the Broadcast News corpus [NIST, 1996, 1997]; words with one syllable account for almost 9% more word tokens (instances) in the spontaneous portion of the corpus.

Spontaneous speech also makes use of slightly different syntactic patterns than planned speech. For instance, at the end of the quotation above, Glass amplifies what he is saying by using the phrase "the sound of a real person talking, actually, really." If one were to reiterate an idea in this type of situation in planned speech, a full phrase would probably be used. Syntactic structures in read speech are often hierarchical, where spontaneous speech utilizes a more linear structure.

Another difference of the unscripted speaking style is the introduction of *dysfluencies* — extra repetitions, false starts, or hesitations that are edited out by the listener. Glass's statement contains two dysfluencies, indicated by the parenthetically marked words. While dysfluencies can also occur in planned speech, they are less prevalent in this domain. Spontaneous speech has a cyclical pattern of dysfluent followed by fluent speech that is not seen in read speech [Henderson *et al.*, 1966], presumably because of the online cognitive processing necessary in planning a spontaneous utterance. The periodicity of the cycle is usually on the order of 10 to 15 seconds; during the dysfluent phase of the cycle, there is a marked increase in the number of filled pauses[2] and hesitations at non-grammatical junctures (*i.e.*, in the middle of a grammatical constituent).

Hesitation and pausing are part of a larger set of linguistic variables called *prosodic variables*. Prosodics are an important cue to the speaking style of the utterance: Levin *et al.* [1982] removed most of the phonetic structure of read and spontaneous speech samples, retaining only intonation and pause structure of each utterance. This was accomplished by applying a low-pass filter with a cutoff of 312 Hz to the scripted and unscripted stories in their experiment. With this low a cutoff, the phonetic identities would be obscured, since the bulk of phonetic information is in the 300–4000 Hz frequency range. Human classification of the low-pass filtered stories into spontaneous and read speech was correct an average 72% of the time — lower than the 84% correct classification for full-bandwidth speech, but still significantly higher than chance. Thus, intonation and pauses are a cue for determining speaking mode. Laan [1997] extended this experiment by resynthesizing pairs of spontaneous and read utterances identical in word content. The resynthesis process involved taking measurements of the pitch, segmental duration, and relative energy of phonemes in both the spontaneous and read utterances, and copying one or more of these features from

[2] Filled pauses are voiced sounds made by a speaker in order to hold the floor in a conversation while the speaker is constructing the sentence, as in "*Um,* I don't know, *um* what to do in this situation." The actual sound is language specific: American English speakers often use "um" or "uh," while German speakers' filled pauses are closer to "er."

one utterance to the other. For instance, in one test condition the read-speech example was modified to have the intonational structure of the spontaneous sample. Laan showed that subjects' perceptions of unmodified speech was correct 71% of the time, while when all three features were swapped (pitch, duration, and energy), the speaking types of the utterances were identified as identical to the original speaking type only 43% of the time — less than chance. How these features correlate with speaking style is still debated in the literature; for instance, different studies have found that the mean fundamental frequency ($F_0$) for spontaneous speech can be higher [Daly and Zue, 1992], lower [Koopmans-van Beinum, 1990], or the same as read speech [Ayers, 1994]. However, the relative frequency of occurrence of some intonational patterns (such as rising contours found in questions that determine if the listener is comprehending) is agreed to be different in the two speaking styles [Ayers, 1994].

Differences in speaking style can also correlate with differences in the pronunciation of words. Casual speaking style is often accompanied by a number of phonological phenomena, as described by Zwicky [1972b]. These processes include assimilation (in which one sound becomes more like another in manner or place of articulation, *e.g.*, *tin pan* ⇒ *tim pan*), deletion of phonetic segments (*e.g.*, *don't know* ⇒ *don'know*), and monophthongization (*e.g.*, *I don't know* ⇒ *ah don't know*). Kaisse [1985] warns that casual speech is not the same as fast speech. This is exemplified by the phonologically reduced phrase *lea'me alone* [Levelt, 1989, p. 369], which can be spoken at slow speaking rates; the deletion indicates a casual register, rather than fast speaking rate.[3] Phonological phenomena also depend on higher-level linguistic structures; some phonologically short words are allowed to adjoin to other words (*cliticization*), as in *I have* ⇒ *I've*. This can happen only in a few syntactic contexts. Syllable structure also plays a part in determining reductions — syllable-final consonants are less likely to be fully realized in spontaneous speech [Fosler-Lussier *et al.*, 1999; Greenberg, 1997a; Cooper and Danly, 1981].

The implication of these speaking style differences for the field of Automatic Speech Recognition (ASR) is that system performance will depend on the ability of system components to model the variability introduced by speaking style. In fact, current state-of-the-art speech recognizers do not perform well when speaking style changes. For the last few years, there has been an annual assessment of how well speech recognition systems decode radio and TV news programs. Figure 1.2 shows the word error rates for all of the systems that participated in the 1998 DARPA Hub4E Broadcast News Evaluation [Pallett *et al.*, 1999]; in this figure, the error rates for spontaneous and planned utterances in studio recording conditions are shown. The number of errors produced by each of these systems was 60–100% more for the spontaneous condition.

[3] Levelt actually assumes that phonological phenomena are only indirectly a consequence of casual register, and that the speaking rate actually is the direct cause of this reduction, in contrast to Kaisse's view. However, I disagree with Levelt's notion that causal register is secondary, since at certain speaking rates, both *leave me* and *lea'me* are possible, the only difference being the speaking mode. Rather, it is the speaking rate and register in conjunction that determine the likelihood of this type of reduction.

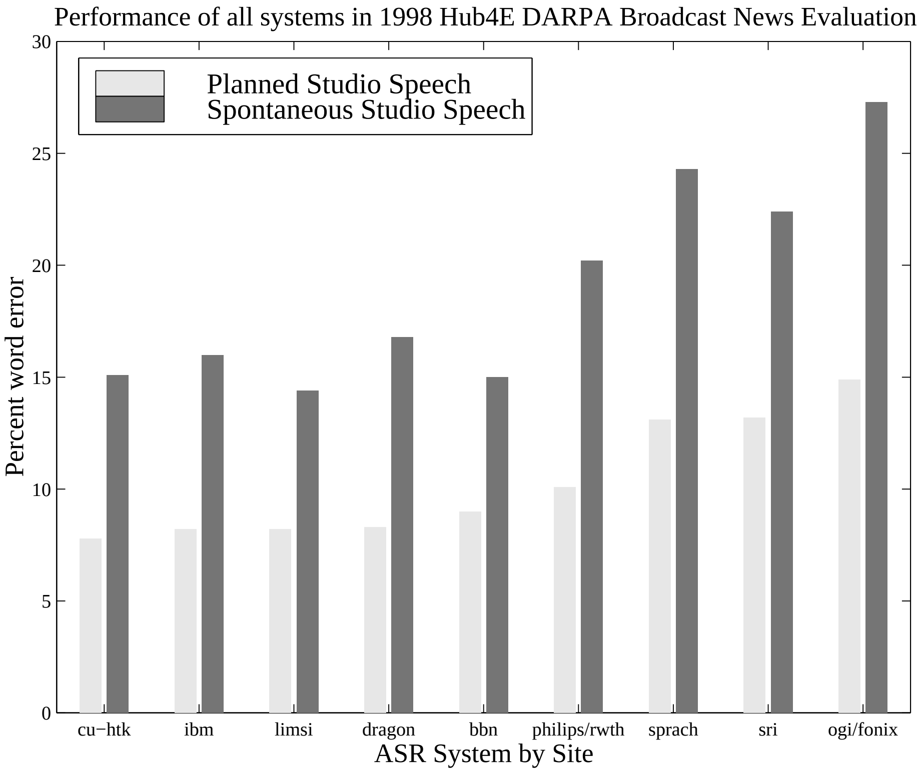


Figure 1.2: Recognizer word error rates for planned versus spontaneous speech in the Broadcast News corpus (after Pallett *et al.* [1999]).

## 1.2 What is pronunciation?

Since the goal of this dissertation is to improve models of pronunciation for ASR systems, a definition of the concept of *pronunciation* is in order. As speakers and listeners, humans have an intuitive feel for pronunciation — people chuckle when words are mispronounced and notice when a foreign accent colors a speaker's pronunciations. Most introductory linguistics texts, however, do not even define this basic concept. In essence, pronunciation refers to two related constructs: the act of producing an acoustic message (as in "Eric's pronunciation is horrible,") or the actual acoustic message itself in an abstract sense ("The pronunciation of *cat* is `[k ae t]`"). As described in the next chapter, linguistics devotes (at least) two fields to detailing the production of pronunciation: *phonetics*, which studies the range of vocal sounds made during spoken language generation, and *phonology*, which models the variation in phonetics by finding a smaller set of underlying categories and determining how these categories relate to phonetic phenomena.[4]

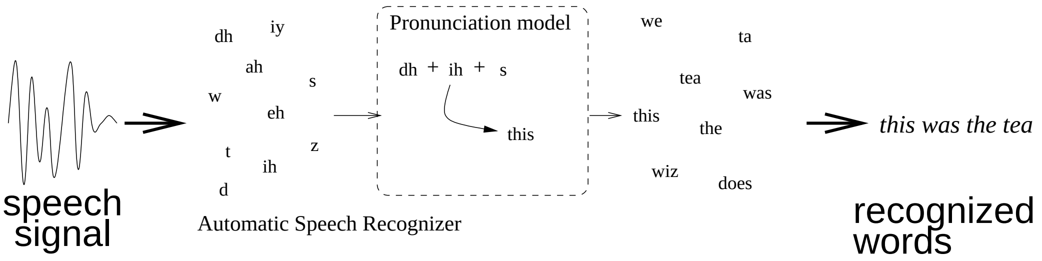


Figure 1.3: The pronunciation model within an automatic speech recognizer.

In ASR, the concept of pronunciation is embodied in the *pronunciation model*. Speech recognizers take speech as input and produce a word transcript. At a very coarse level, this is accomplished by classifying the speech signal into small sound units, or *phones* (Figure 1.3). The pronunciation model determines how these phones can be combined to make words. It is the job of the pronunciation model to determine what phone sequences constitute valid words, *e.g.*, to allow the phonetic sequence [d ow n n ow] to represent *don't know*. System builders often look to phonology for tools to represent this variation. A bit of caution is necessary, though: in the same way that airplanes do not fly by flapping their wings, it is not clear that machines perceive patterns of phones as humans do. Some phonological models may not carry over to the ASR domain completely.

## 1.3 Modeling pronunciations in an ASR system

As the next chapter describes, the pronunciation models within an ASR system can be constructed in several ways: for example, by writing down pronunciations of words

[4] In fact, the boundaries between phonetics and phonology are often blurred; some authors suggest that there is no separation between these two fields (see Ohala [1990] for further discussion).

by hand, by deriving them with phonological rules, or by finding frequent pronunciations in a hand-transcribed corpus. The models of pronunciation in this thesis are *automatically* constructed using machine learning techniques. The pronunciation modeling paradigm in this work allows for discovery of pronunciation patterns within particular corpora, without presupposing the types of patterns that will be seen.

The implication of automatic discovery of pronunciation patterns in machine-based phonetic models is that the resulting phonological patterns may not represent actual processes found in human phonological systems. On the one hand, the baseline models of the speech recognizer are typically linguistically seeded — that is, acoustic phone models and pronunciations are determined by a phonetic transcription.[5] Therefore, automatically constructed models will likely capture linguistically plausible variation patterns. On the other hand, because the actual phone models used are statistical pattern recognizers, rather than psychoacoustic models that would suppress channel effects, additional non-linguistic variations may be induced by, *e.g.*, background or channel noise in the acoustic input. Even if acoustic models are behaving in a non-intuitive manner, as long as the patterns of variation are regular, we may be able to correctly recognize words with them. A pronunciation model based on statistical learning techniques rather than on pre-defined linguistic constructs may represent this non-linguistic variability within phone models, since it is difficult to tell *a priori* what the effects of non-linguistic variability are on phone classification. This thesis shows that with a corpus that is marked for different speaking modes and acoustic conditions, one can determine the extent to which linguistic and non-linguistic acoustic variation is modeled by the pronunciation dictionary.

The development strategies of automatically learned ASR pronunciation dictionaries can be divided into two basic camps. In one general class of techniques, pronunciations are learned on a word-by-word basis from some corpus of phonetic transcriptions, either hand-transcribed or automatically generated. Most *dynamic* pronunciation models, on the other hand, model the variation in pronunciations on a phone-by-phone basis by choosing appropriate phonetic representations depending on some representation of context. Both of these paradigms have various advantages and disadvantages;[6] this thesis examines the idea of integrating context-based (dynamic) modeling of pronunciation variations into the automatic learning of pronunciations for syllables and words.

To accomplish this goal, one must first define the meaning of *context*. As indicated in the previous paragraph, the juxtaposition of two phonetic elements can cause a different acoustic realization than when each of the two phones is spoken in isolation. Syllabic structure also plays an important part in defining pronunciation context: phones at the ends of syllables are more likely to be pronounced in a variety of ways than phones at the starts of syllables. In Chapter 3, I present studies of the effects of two other factors on word pronunciation within spontaneous speech — the rate of speech, and the predictability of the word being spoken.

In this thesis, I expand the traditional meaning of context within ASR pronunciation models. Inclusion of speaking rate, word predictability, syllable structure, and other

[5] Sometimes transcriptions are from hand-annotated sources, but more often, the transcriptions are automatically generated using a phonetic dictionary.

[6] The range of models found in current ASR systems is further described in Chapter 2.

factors into pronunciation models may improve prediction of pronunciation variation. With decision tree models [Breiman *et al.*, 1984], analysis of the utility of various factors can be studied. One can ask questions such as "is factor X more important than factor Y?" I use analyses of this type to determine a rough hierarchy of feature relevance. Furthermore, these contextual models improve the recognition accuracy produced by ASR systems.

## 1.4 Thesis outline

Chapter 2 provides a broader introduction to speech recognition and pronunciation modeling, including an overview of formalisms for understanding pronunciation variation in linguistics and ASR systems. I also present a short survey of factors that affect word pronunciations. In Chapter 3, I investigate the effects of speaking rate and word predictability on pronunciations of phones, syllables, and words. The variation in pronunciations due to these factors can affect the performance of ASR systems; Chapter 4 presents evidence of this phenomenon in two recognizers. The following chapter describes baseline dictionary enhancements designed to improve recognition accuracy on the Broadcast News corpus. Chapter 6 expands on the lessons learned in Chapters 3 and 5 by dynamically determining the appropriate pronunciation models for a word or syllable based on an extended concept of context, including the speaking rate and predictability of the word as conditioning factors in the pronunciation model. Finally, conclusions and future directions are presented in Chapter 7.

# Chapter 2

# Background and Motivations

Several types of pronunciation models have been described in both the linguistics and ASR literature. These various approaches have included different ways of capturing the statistical variation of pronunciations. This chapter describes several formalisms that have been used for this purpose and argues that a *dynamic* model, in which the probabilities of pronunciations change due to different contexts, is more appropriate for modeling spontaneous speech in ASR than a *static* dictionary in which pronunciations and their probabilities are pre-set at the run-time of the recognizer. I also discuss some factors that have been shown to affect pronunciations in the linguistics literature and argue for their inclusion in a dynamic model.

## 2.1 Overview of a speech recognizer

Speech transcription, as generally defined in the literature, is the problem of writing down what someone has said. This definition makes no distinction regarding the agent that processes the utterance — whether human or machine. For speech transcription to be accomplished automatically, models of the utterances in a language are required, so that the best model can be selected as the transcription. This section discusses the various components of the typical automatic speech recognition system and how pronunciation models fit into the system.

Utterances that are the input to an ASR system are recorded as acoustic signals and digitally quantized into some representational vector $X = x_1, x_2, \ldots, x_t$, where $t$ depends on the length of the utterance. This representation is usually based on what speech scientists know about the psychological representation of sound in the human auditory system.

If the range of possible utterances (word sequences for speech transcription) in the universe is $\mathcal{M}$, the speech recognition problem can be stated formally as:

$$M^* = \operatorname*{argmax}_{M \in \mathcal{M}} P(M|X) \tag{2.1}$$

In other words, what is the string of words $M^*$ that has the highest probability given the acoustic waveform that was input into the computer? This probability is, in general, intractable to compute; Bayes' rule, however, is applied to break up this probability into components:

$$\begin{aligned} \operatorname*{argmax}_{M \in \mathcal{M}} P(M|X) &= \operatorname*{argmax}_{M} \frac{P(X|M)P(M)}{P(X)} && (2.2) \\ &= \operatorname*{argmax}_{M} P(X|M)P(M) && (2.3) \end{aligned}$$

During recognition, the prior probability of the acoustics $P(X)$ in the denominator of Equation 2.2 may be removed from consideration because the argmax operator does not depend on $X$ at all, that is, $P(X)$ is constant over all hypothesized utterances $(M)$. Thus, the ASR system must model two probability distributions: (1) the probability of the acoustics matching the particular hypothesis $P(X|M)$, and (2) the prior probability of the hypothesis $P(M)$.

Of course, it is difficult to model the likelihood $P(X|M)$ directly — this would involve modeling the relationship between all acoustic sequences and all possible utterances. In order to makes the models more tractable, in large vocabulary speech recognition we invoke conditional independence assumptions to subdivide the models. Words in utterances are represented by subword units called *states.* Models of utterances are deconstructed into a state sequence $Q$, representing the total joint probability of the acoustics $X$ and model $M$ with three separate models, each with its own linguistic correlate. As noted above, $X$ is an acoustic representation that might correspond to a model of human acoustic perception. The $M$ vector is usually representative of a sequence of words, and $Q$ is a sequence of subword units, usually phones. Three different terms that comprise the complete probability distribution map onto different subfields of linguistics:

$P_A(X|Q)$: The probability of acoustics given phone states (known as the *acoustic model*) is similar to psychological models of categorical perception.

$P_P(Q|M)$: This is the probability of phone states given the words, encompassing the *pronunciation model* and *duration model*, which maps onto the fields of phonetics, phonology, and to some extent, morphology. We will shortly see how the pronunciation and duration models are further decomposed into separate models.

$P_L(M)$: The prior probability of word sequences (the *language model*) has a correlate in the linguistic areas of syntax and semantics.

These three models, $P_A$, $P_P$, and $P_L$, are related to $P(X, M)$ as follows:

$$\begin{aligned}
\operatorname*{argmax}_M P(X,M) &= \operatorname*{argmax}_M P(X|M)P(M) && (2.4)\\
&= \operatorname*{argmax}_M \sum_Q P(X|Q,M)P(Q,M) && (2.5)\\
&\approx \operatorname*{argmax}_M \sum_Q P_A(X|Q)P_P(Q|M)P_L(M) && (2.6)\\
&\approx \operatorname*{argmax}_M \max_Q P_A(X|Q)P_P(Q|M)P_L(M) && (2.7)
\end{aligned}$$

Equation 2.5 follows directly from probability theory; the subsequent equation makes the assumption that the acoustic likelihood is independent of the word models given the state sequence. This represents a representational savings since the lexicon numbers in the tens of thousands, but the number of different states in a system is only on the order of 40 to 1000, depending on the type of sub-word unit. In order to restrict the search space over the models, a *Viterbi approximation* is often employed, where the summation is replaced with a maximum over the state sequences (Equation 2.7).

The typical ASR system has different components that estimate each part of the model (Figure 2.1). Acoustic features of the acoustic signal are produced by the auditory front end. MFCCs (Mel Frequency Cepstral Coefficients) are a popular choice of representation, although PLP (Perceptual Linear Prediction) [Hermansky, 1990] has recently gained in popularity for large-vocabulary tasks. At ICSI, the Modulation SpectroGram Filtered Features (MSG) have recently been developed for robustness to noise and reverberation [Kingsbury, 1998].

Acoustic likelihoods ($P_A(X|Q)$) can be calculated in one of several ways. In traditional Hidden Markov Model (HMM) systems, a Gaussian distribution over the phones can be determined for individual acoustics and phones ($P_A(x_t|q_j)$); these estimates are then multiplied together to give an overall estimate of the probability $P_A(X|Q)$.[1] An extension of this model is the *triphone*, which gives estimates of $P_A(x_t|q_{j-1}^{j+1})$: the probability of an acoustic vector given a phone and its immediate neighbors.

[1] This assumes that, given the state sequence, one acoustic vector is independent of the next — an assumption often attacked in the literature, although in practice this assumption is passable.

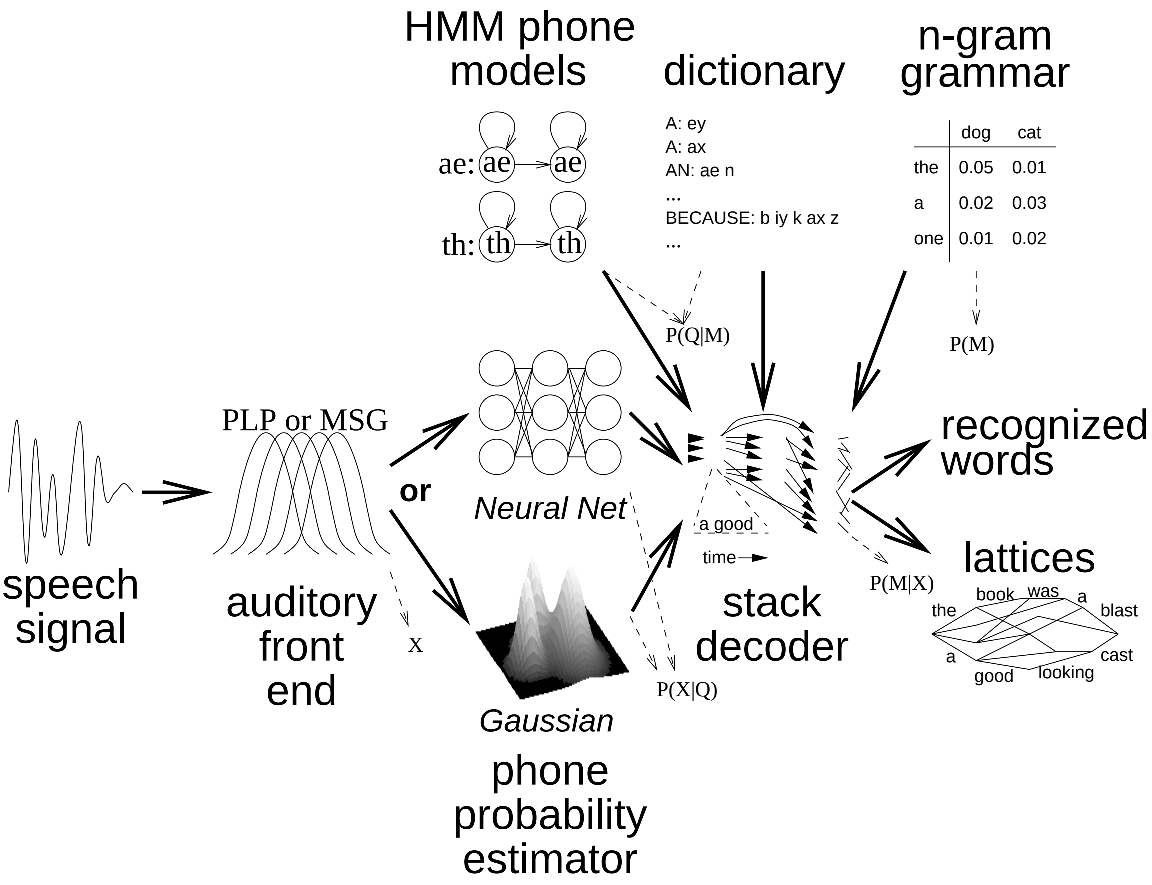


Figure 2.1: Block diagram of a typical ASR recognizer.

In a hybrid Hidden Markov Model-Artificial Neural Net (HMM-ANN) system [Bourlard and Morgan, 1993], a neural network is used to calculate posterior estimates of the states given the acoustics ($P(Q|X)$). With Bayes' Rule, this probability can be converted into the traditional HMM acoustic likelihood:

$$P_A(X|Q) = \frac{P(Q|X)P(X)}{P(Q)} \tag{2.8}$$

although the prior on the acoustics ($P(X)$) is not computed (since it is constant during recognition), giving scaled likelihoods. For the most part, in this thesis the source of likelihoods is not critical; I treat the acoustic model as a black box.[2]

Likewise, the language model (LM) that provides an estimate of $P_L(M)$ is not a concern of this thesis. Typical large-vocabulary decoders use $n$-gram grammars for the LM. In general, the probability of a word model sequence $M$ can be decomposed as follows:

$$P_L(m_1 \dots m_t) = P(m_t|m_{t-1}, m_{t-2}, \dots, m_1)P(m_{t-1}|m_{t-2}, \dots, m_1) \dots P(m_1) \tag{2.9}$$

[2] The only case where this is not true is in the use of acoustic confidence scores, which depend on the posterior probability $P(Q|X)$, as described in Chapter 5.

$$= \prod_{i=1}^{t} P(m_i|m_{i-1},\ldots,m_1) \tag{2.10}$$

An $n$-gram grammar makes the assumption that word histories more than $n-1$ words before the current word do not affect the probability:

$$P_L(m_1 \ldots m_t) \approx \prod_{i=1}^{t} P(m_i|m_{i-1},\ldots,m_{i-(n-1)}) \tag{2.11}$$

Trigrams are the most common LMs nowadays, although many systems have been moving toward incorporation of 4-grams for very large tasks.

The concern of this thesis is how to determine the pronunciation model, $P_P(Q|M)$. The pronunciation model serves an important role: it acts as the interface between acoustic models and words, creating mappings between these two models. As an interface, a pronunciation model must deal with variation from both sides: linguistic variation in pronunciations caused by such factors as predictable word sequences or increased speaking rate, and acoustic model variation due to noisy environments and the like.

In the system used at ICSI, the model $P_P(Q|M)$ is broken into two parts, as illustrated by Figure 2.2. First, a set of HMM phone models give durational constraints for individual phones. How the phones are concatenated into words is determined by the pronunciation dictionary. In most systems, this dictionary is a lookup table, giving phonemic representations (or *baseforms*) of each word. When words have more than one representation, the table is referred to as a *multiple pronunciation dictionary.* This dictionary provides a model of baseform sequences, $P_B(B|M)$, as part of the overall pronunciation model:

$$P_P(Q|M) = P_D(Q|B)P_B(B|M). \tag{2.12}$$

In general, the baseform pronunciations of a word are assumed to be independent of the word context; that is,

$$P_B(B|M) \approx \prod_{i=1}^{n} P_P(b_i|m_i), \tag{2.13}$$

where $b_i$ is the pronunciation given to the $i$th word in the pronunciation sequence B. In a *static dictionary* (left path of Figure 2.2), the model that provides HMM state sequences from the baseforms, $P_D(Q|B)$ contains only duration information, as the phones used to make up the word are given by $B$.[3] The duration model in our system indicates the number of HMM states to allocate to each baseform phone, and the transition probabilities between states of a phone.[4]

[3] This is true for the ICSI monophone recognizer; for triphone-based systems, $P_D(Q|B)$ also includes the mapping from baseforms to triphone models.

[4] In general, the duration model determines the *topology* of the HMM for a phone. This can be defined by a state-to-state transition matrix; absolute constraints on duration, such as a minimum path length through the HMM, can be implemented by inserting zeros in the transition matrix. Some systems choose to allow only ones and zeros in the transition matrix and explicitly incorporate a probabilistic term for the duration of phones or words.

In Equation 2.12 it is assumed that the choice of the state sequence $Q$ depends solely on the baseform sequence $B$. Moreover, in a static dictionary with monophone models, the choice of a particular $q_i$ usually only depends on a single baseform phone $b_j$.[5] In a *dynamic dictionary* (illustrated in the right path of Figure 2.2), $q_i$ can depend on several baseform phones, the model sequence $M$, and other features. This thesis examines the possibility of adding other factors to this probability distribution, replacing the static $P_D(Q|B)$ with the dynamic pronunciation model $P_{DP}(Q|B, M, \ldots)$. The model $P_{DP}$ introduces an intermediate *surface* pronunciation form $S$, distinguished from the underlying baseform sequence $B$, that represents the dynamic variation of pronunciations. Thus, $P_{DP}$ is further decomposed into a model of the dynamics $P(S|B, M, \ldots)$ and the duration model $P_D(Q|S)$.

The remainder of this chapter describes how the field of phonology in linguistics has

[5] The indices do not match here because several states can correspond to a baseform phone. In contrast to the monophone system, in a triphone system $q_i$ depends on the baseform phones $b_{j-1}$,$b_j$, and $b_{j+1}$.

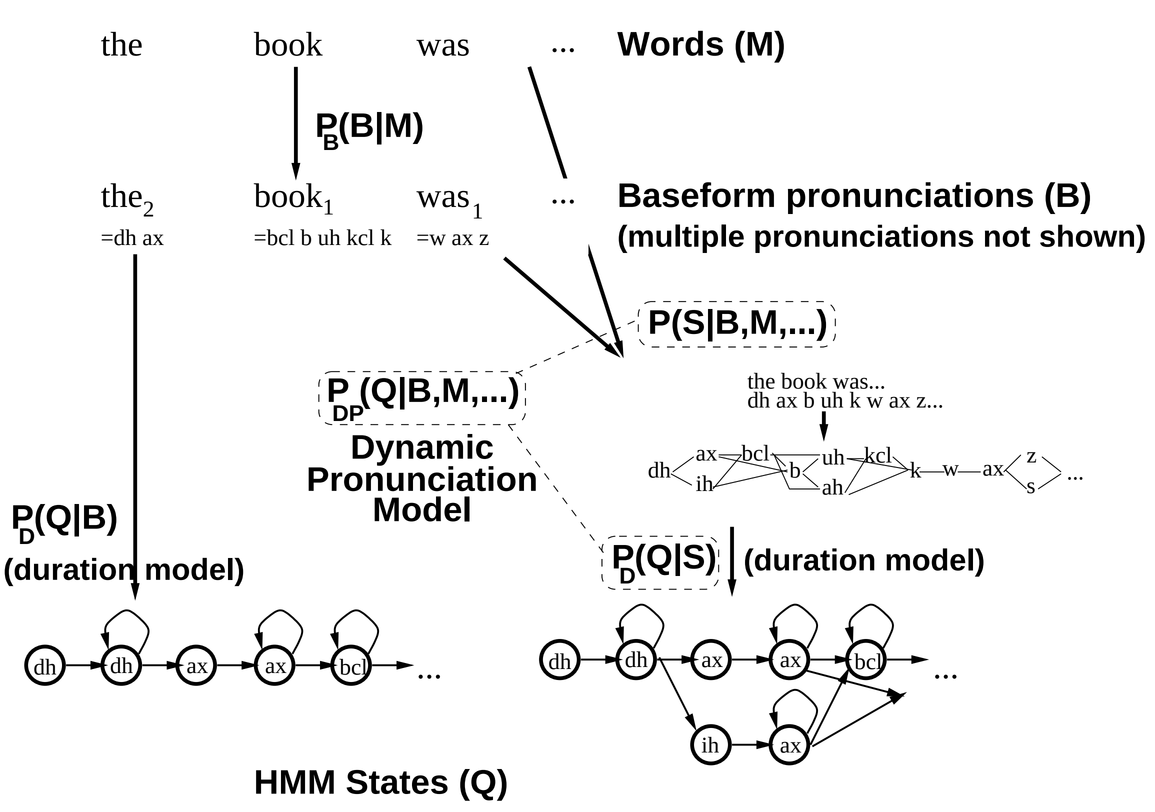


Figure 2.2: The decomposition of $P_P(Q|M)$ for static and dynamic dictionaries. For static dictionaries, the left path is followed; after words are converted into baseform pronunciations, duration models are added to obtain HMM state sequences. For dynamic dictionaries, an additional conversion from baseform pronunciations (B) to surface pronunciations (S) occurs (in $P(S|B, M, \ldots)$) that allows pronunciations to vary based on the surrounding phone and word context.

modeled patterns of sound and uses these formalisms to draw connections to how ASR pronunciation modelers construct the pronunciation models $P_B(B|M)$ and $P_{DP}(Q|B, M, \ldots)$. In the final portion of the chapter, I describe some factors that may be appropriate in the context of the dynamic pronunciation model.

## 2.2 Linguistic formalisms and pronunciation variation

The work in ASR pronunciation modeling is derived almost completely from linguistic theory. This section describes two major phonological theories in practice today, in order to better illustrate how ASR technology has made use of these theories.

### 2.2.1 Phones and phonemes

In linguistic theory, sound units are divided into two basic types: phones and phonemes. *Phones* are the fundamental sound categories that describe the range of acoustic features found in languages of the world. While the actual set of phones used to describe sound patterns in a language may vary slightly from linguist to linguist, phoneticians in general do have a system for codifying these sounds: the International Phonetic Alphabet (IPA). One representation of the phones of English, along with the corresponding IPA symbols, can be found in Appendix A.

*Phonemes*, on the other hand, are more abstract, language-specific units that correspond to one or more phones. The field of phonology is dedicated to describing which phone one would expect to see in particular instances. For example, there are (at least) two types of p sounds in English: an aspirated p ([$\mathrm{p}^{\mathrm{h}}$]), as in the word *pit*, and an unaspirated p ([p]), as in the word *spit*. The difference between these two phones is the amount of air expelled at the release of the mouth closure. However, if one were to substitute an aspirated p into *spit*, the meaning of the word would not change. This means that these two p phones are *allophones* of the phoneme /p/; in other words, the phoneme /p/ can have two realizations, [$\mathrm{p}^{\mathrm{h}}$] and [p], depending on the context.[6] Compare this with [p] and [n] — substituting [n] for [p] changes the word to *snit*, so /p/ and /n/ are different phonemes. As the reader may have observed, to distinguish phones from phonemes in text, one uses different delimiters. Phones are set off using brackets (*e.g.*, [p]), whereas for phonemes we use slashes (*e.g.*, /p/).

What makes the difference between phones and phonemes sometimes confusing for speech recognition researchers is the fact that most systems use neither phones nor phonemes, but something in between. The most common representation of sound segments in ASR systems is very much like a set of phonemes (around 40 units for English), although some systems (like the ICSI recognizer) use separate representations for stop closures, bursts and flaps, as it facilitates discrimination between these acoustically disparate situations. Given this more phonetic orientation in the ICSI recognizer, in this thesis I will use the

[6] The situation is more complicated than as it first appears. In fact, if the /s/ were removed from *spit*, the resulting word would sound more like *bit* than *pit*. Thus, context plays a large role in how phones are realized. Thanks to John Ohala for this example.

bracket notation when describing sound units, but to distinguish them from phones, I use a bold typewriter font (*e.g.*, **`[p]`**).

### 2.2.2 Phonological rules

Linguists' efforts in the field of phonology are devoted to capturing the variation of surface forms of pronunciation. There are two basic parts to a phonological system: a hypothesis of the underlying (phonemic) pronunciations of words, and a system for deriving the surface (phonetic) representation from this underlying form. One historically popular system is the derivation of surface pronunciations by transformational phonological rules [Chomsky and Halle, 1968]. In this system, rewrite rules are used to express transformations from an underlying form to a surface form. In general, the form of a phonological rule is:

$$\mathrm{A} \rightarrow \mathrm{B} \quad / \quad \mathrm{C} \underline{\quad} \mathrm{D} \tag{2.14}$$

This transformational rule can be read as "change A to B when it follows C and precedes D." The context (C and D) can be specified as a class of phones instead of phone identities, for generality. As an example, the commonly known "flapping" rule that differentiates the /t/ in British *butter* ([b ah t ax]) from American *butter* ([b ah dx axr]) can be written as:

$$/\mathrm{t}/ \rightarrow [\mathrm{dx}] \quad / \quad \left[\ +\mathrm{vowel}\ \right] \underline{\quad} \left[\begin{array}{c} +\mathrm{vowel} \\ \text{-stress} \end{array}\right] \tag{2.15}$$

This rule reads: change the phoneme /t/ to a flap ([dx]) when preceded by a vowel and followed by an unstressed vowel. A good phonological representation describes phonological alternations with as concise, general rules as possible.

### 2.2.3 Finite state transducers

Phonological rules have a deep connection with finite state automata (FSA). Johnson [1972] showed that phonological rules of the form A→B/C _ D, while appearing to be general rewrite rules, were equivalent to the much smaller set of finite state languages under the assumption that the output of a rule was never fed back into the rule (*i.e.*, *recursive* phonological rules were disallowed). The equivalent type of automaton is a *finite state transducer* — an FSA that associates pairs of input and output symbols. In the case of phonological rules, the inputs are phonemes and the outputs are phones.

A simple transducer for the flapping rule is shown in Figure 2.3. The state path 0-1-2-0 shows the main part of the phonological rule: if there is a stressed vowel on the input and the following input phoneme is a /t/, then transform the /t/ to a [dx] if the following input phoneme is an unstressed vowel. State 3 is necessary for explaining what happens if an unstressed vowel does not follow: the realization of /t/ is left unspecified. State 2 is the only non-final state in this transducer because if /t/ occurs at the end of a word, a flap is not allowed.

Kaplan and Kay [1981] observed that finite state transducers are closed under serial composition. This means that if one has an ordered set of phonological rules

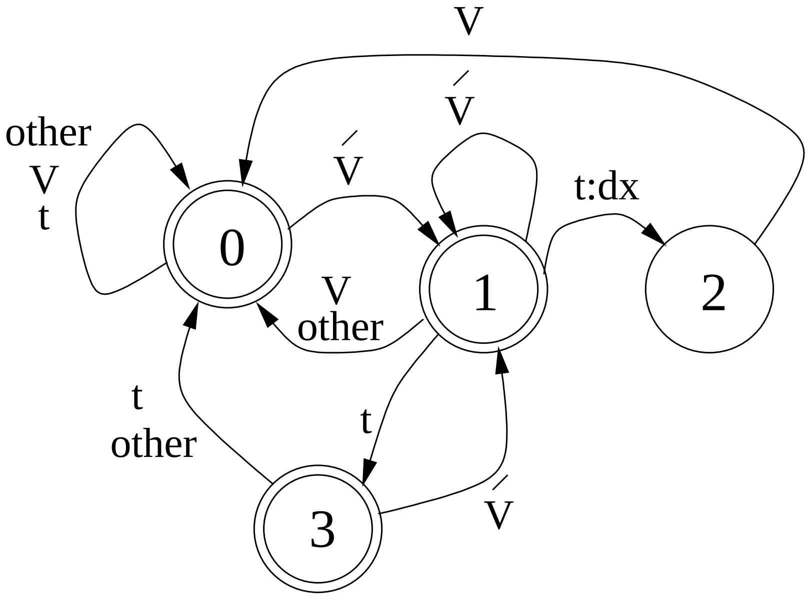


Figure 2.3: Finite-state transducer for the English flapping rule, after Jurafsky and Martin [2000]. Input-output pairs are indicated on each arc, separated by colons. V́ indicates a stressed vowel, V is an unstressed vowel, other represents any feasible pair of input-output relations. Arcs without a colon indicate only a constraint on the input, not the output. The initial state is state 0, and final states are indicated by double circles.

$\{R_1, R_2, \ldots, R_n\}$, corresponding to transducers $\{T_1, T_2, \ldots, T_n\}$, then when a string is given to $T_1$, and the output of $T_1$ is fed into $T_2$, the output from $T_2$ into $T_3$, and so on down to $T_n$, this series of operations will produce an output equivalent to the output from the input applied to the single transducer $T_1 \circ T_2 \circ \ldots \circ T_n$, where $\circ$ is the composition operator. Using this technique, all of the phonological rules of a language can be compiled into one large transducer. Koskenniemi [1983] developed a similar approach to transduction in his thesis, called *two-level morphology.*[7] In his paradigm, the transducers are used as parallel constraints, instead of applied serially as in Kaplan and Kay; rules are specified slightly differently to accommodate this parallelism. The advantage of this specification is that there is no rule ordering within the system. Finite state transducers are *invertible*, which means that they can be used not only for generation of pronunciations, but also for phonological parsing by interchanging the input and output of the system. Transducers may be induced directly from data [Oncina *et al.*, 1993], but Gildea and Jurafsky [1996] showed that for phonological-rule learning, seeding the transducers with linguistic knowledge will allow induction of more compact models. For a further introduction to finite state transducers and their use in phonology and morphology, see Karttunen [1993] and Jurafsky and Martin [2000].

The relevance of finite state transducers to ASR is that finite state grammars form the backbone of most pronunciation models; the state sequence of a Hidden Markov Model is a finite state automaton. When derivational models such as phonological rules or decision trees are used to construct new pronunciation models, underlying this transformation are the implicit grammar operations described above. Some researchers (*e.g.*, [Sproat and Riley, 1996; Mohri *et al.*, 1998; Riley *et al.*, 1998]) have chosen to make this representation more explicit in ASR by operating directly on the transducers; this often makes murky transformational operations in phonological models clearer. This thesis makes use of some of this technology in converting automatically learned rules into finite state grammars that are usable by the speech recognizer.

### 2.2.4 Constraint-based phonology

In contrast to the derivational paradigm of phonological rules, the other major class of phonological theories is constraint-based phonology. Instead of having rules that specify how to derive surface pronunciations from underlying forms, in constraint-based phonology surface pronunciations are generated by filtering a large set of candidate forms with a set of linguistic constraints. The prime exemplar of this paradigm is *Optimality Theory* (OT) [Prince and Smolensky, 1993], which, in the words of Eisner [1997], has taken the field of phonology by storm:

> Phonology has recently undergone a paradigm shift. Since the seminal work of Prince and Smolensky [1993], phonologists have published literally hundreds of analyses in the new constraint-based framework of Optimality Theory, or OT. Old-style derivational analyses have all but vanished from the linguistics conferences.

[7] Although Koskenniemi's thesis dealt primarily with morphology, this system applies to phonology as well.

The first step in determining the phonetic realization of an underlying phonological representation is to generate a possible set of phone sequences. In the most naïve approach, the generator provides an infinite set of candidates.[8] All possible surface forms are ranked against each other in a series of "competitions," corresponding to an ordered set of violable constraints on the correspondence between the phoneme input and phonetic output. The first constraint is applied (*e.g.*, minimize the number of coda consonants in the word), and the candidate surface forms that are tied for the highest score (or, put another way, the surface forms that violate the constraint the least) are kept in for the next round. As subsequent constraints are applied, the list of potential surface forms becomes shorter and shorter until there is only one form left, which is output. The ordering of constraints is language-specific; much of the effort in current phonological investigations is to find constraints (and the ordering thereof) that explain phenomena in different languages. Work has also been undertaken toward a computational implementation of the theory [Ellison, 1994; Bird and Ellison, 1994; Tesar, 1995; Eisner, 1997]; among the issues in this area are how to represent constraints, what types of constraints to allow in the system, and how to generate a finite set of potential candidate baseforms.

There is an interesting parallel in pronunciation modeling for speech recognition to this comparison of derivational versus constraint-based phonology. Some pronunciation learning systems generate a set of baseforms by transforming a base dictionary with a set of phonological rules. Other systems generate a candidate set of baseforms using a phone recognizer, and then constrain them with various filtering techniques, much like constraint-based phonology. In the next section, we will see how both derivational and constraint-based techniques are used in ASR modeling.

## 2.3 ASR models: predicting variation in pronunciations

One primary source of pronunciations is a hand-built lexicon, in which experts carefully craft pronunciation models for each word [Lamel and Adda, 1996]. When care is taken to minimize dictionary confusions and ensure consistency of pronunciation across similar words, the resulting lexicon is often excellent. However, the work, being manual in nature, is very time consuming; incorporating a new vocabulary into a recognizer can often be prohibitively expensive.

Researchers have been examining ways to build dictionaries in a more automatic fashion. Two components are needed for a pronunciation modeling system: a source of data, and a method of capturing variation seen in the data. Pronunciation data can be provided from either hand-annotated speech, or (more frequently) automatic phonetic transcriptions provided by a speech recognizer. In most of the systems below, a hand-crafted dictionary is used as a type of "phonological" representation. The linguistic formalisms discussed in the previous section have given pronunciation modelers many tools to use to model variation in phonetic transcripts given the baseline dictionary. In this section, I examine several different

[8] To make this computationally feasible in more advanced approaches, restrictions are placed on the pronunciation generator [Eisner, 1997]. However, all of these approaches require presupposing (or often post-supposing!) an underlying baseform pronunciation for generating candidate surface forms, but the theory of underlying forms is often vague in these analyses.

techniques used in building ASR pronunciation models, culminating in the transformational model used in this thesis.

### 2.3.1 Automatic baseform learning

The simplest method of learning pronunciation variants is to learn each word's various pronunciations on a word-by-word basis. Typically, a phone recognizer is utilized to determine possible alternatives for each word by finding a best-fit alignment between the phone recognition string and canonical pronunciations provided by a baseline dictionary, although hand-transcribed data can also be used for this task [Cohen, 1989].

Wooters [1993] and Westendorf and Jelitto [1996] used alignments between baseform pronunciations and frequent pronunciations derived from a phone recognizer to create a set of word-recognition baseforms. Others have used phone recognition output constrained by orthography-to-phone mappings [Lucassen and Mercer, 1984; Bahl *et al.*, 1981], or by generalizing phone recognizer output to broad phonetic categories (such as stops and fricatives) [Schmid *et al.*, 1987].

One can also generalize over the set of pronunciations learned by these techniques, using techniques such as HMM Generalization [Wooters and Stolcke, 1994], which allows induction of new baseforms not seen in the training data by finding common variations among pronunciations seen during training. Eide [1999] used a similar technique in their Broadcast News speech recognizer, finding improvements in word error rate for spontaneous and fast speech.

One problem with word-based techniques is that they do not model coarticulation between words well. For example, the fact that the second /d/ in "Did you eat?" often gets realized as /jh/ (as in "Didja eat?") is conditionally dependent on the fact that the following word is "you"; a word-by-word technique would not have access to this dependence. Sloboda and Waibel [1996] address this problem by adding common phrases (*tuples*) to their dictionary, allowing for coarticulation modeling across word boundaries in German.[9] They found that adding tuples reduced word error rate about 3% relative (increasing word accuracy from 65.4% to 67.5% on the Verbmobil[10] task), and learning multiple pronunciations brought relative improvement to 4% (68.4% word accuracy). Compounding words is a technique that has been used for language modeling for some time (*e.g.*, McCandless and Glass [1993]; Ries *et al.* [1995]), but the use of this technique within pronunciation modeling is only recently becoming commonplace [Adda *et al.*, 1997; Placeway *et al.*, 1997; Finke and Waibel, 1997b; Nock and Young, 1998; Ma *et al.*, 1998; Beulen *et al.*, 1998].

Another problem in word-by-word learning is that generalizations cannot be learned across words. For instance, the phoneme /t/ in *butter*, *flutter*, and *mutter* shares the same phonemic environment in each word; however, in a word-by-word setting, the probability for flapping the /t/ must be learned independently for each word; such probability estimations may run into sparse data problems. One solution to this problem may be to

[9] This technique is generally called *multi-word* modeling in the literature.

[10] A speech recognition task that operates as a first stage for a human-to-human translation system for appointment-making.

smooth pronunciations with the probabilistic phonological rules described above. Another possibility is to learn phonological rules automatically across words, in procedures described below.

### 2.3.2 Phonological rules

Phonological rules have been used extensively in ASR systems in order to model phonetic variations due to coarticulation and fast speech by expanding the pronunciation possibilities within the lexicon. The first studies describing the need for capturing phonological regularities appeared in a special issue on speech recognition in the IEEE Transactions on Acoustics, Speech, and Signal processing; two papers cited the inclusion of phonological rules in their systems [Oshika *et al.*, 1975; Friedman, 1975], although neither paper reported recognition results.[11] However, phonological rules were a part of early speech recognition systems, including a system at IBM [Bahl *et al.*, 1978], the BBN HWIM system [Wolf and Woods, 1980], CMU HARPY system [Lowerre and Reddy, 1980], and CMU SPHINX system [Lee, 1989].

With the development of relatively large, phonologically hand-transcribed corpora such as TIMIT [Garofolo *et al.*, 1993], investigations into phonological phenomena became feasible. Cohen [1989] provided one of the first comprehensive phonological analyses of TIMIT data and used the results to build pronunciation networks for the SRI Decipher recognizer. These insights were used to build probabilistic pronunciation networks for the Resource Management task [Price *et al.*, 1988]; performance on an early version of the DECIPHER system went from 63.1% word accuracy to 65.5% accuracy with the new pronunciation models. In a later version of the system with better acoustic models, the networks were pruned to realize an increase in performance (from 92.6% accuracy with unpruned models to 93.7% with pruned models).[12] In a similar vein, probabilistic rules have been used to describe pronunciation variation in English in the Wall Street Journal and Switchboard domains [Tajchman *et al.*, 1995a,b; Finke, 1996; Finke and Waibel, 1997b], as well as in German in the Verbmobil domain [Schiel *et al.*, 1998], and within a Dutch rail service system [Wester *et al.*, 1998]. Schiel [1993] used probabilistic phonological rules to describe speaker differences in a German system that adapted to individual speakers; the use of phonological rules as an speaker-adaptive technique has also been implemented by Imai *et al.* [1995] and De Mori *et al.* [1995].

One of the problems that can occur with phonological rule generation of pronunciations, however, is that rules can over-generalize, leading to an explosion of possible surface forms. Various methods can be used to reduce the number of generated forms; in one study [Tajchman *et al.*, 1995b], my colleagues and I developed a method for calculating the probability of rules so that the pronunciations produced by the model could be ranked according to their probabilities, thus allowing for pruning. The rules we used in this experiment, derived from [Zwicky, 1970, 1972a,b; Kaisse, 1985], are shown in Table 2.1, along with the probability estimates that we derived for some of the rules.

[11] Cohen and Mercer [1975] described a contemporaneous phonological rule system within the IBM recognizer; a description of their work can be found in [Humphries, 1997].

[12] See also Weintraub *et al.* [1988] for a discussion of these results.

| Name | Rule | Example | Prob |
|---|---|---|---|
| Syllabic Rules* | | | |
| Syllabic n | [ax ix] n → en | butt**on** | .35 |
| Syllabic m | [ax ix] m → em | bott**om** | .32 |
| Syllabic l | [ax ix] l → el | bott**le** | .72 |
| Syllabic r | [ax ix] r → axr | butt**er** | .77 |
| Flapping | [tcl dcl] [t d]→ dx /V — [ax ix axr] | bu**tt**on | .87 |
| Flapping-r | [tcl dcl] [t d]→ dx /V r — [ax ix axr] | bar**t**er | .92 |
| H-voicing | hh → hv / [+voice] — [+voice] | a**h**ead | .92 |
| L-deletion | l → Ø/ — y [ax ix axr] | mi**ll**ion | *n/a* |
| Gliding | iy → y / — [ax ix axr] | colon**i**al | *n/a* |
| Nasal-deletion | [n m ng] → Ø/ — [-voice -consonant] | ra**n**t | *n/a* |
| Function words | | | |
| h-deletion | h → Ø/ # — | **h**e, **h**im | *n/a* |
| w-deletion | w → Ø/ # — | **w**ill, **w**ould | *n/a* |
| dh-deletion | dh → Ø/ # — | **th**is, **th**ose | *n/a* |
| Dental-deletion | [tcl dcl] [t d] → Ø/ [+vowel] — [th dh] | brea**d**th | *n/a* |
| Final dental-deletion | ([tcl dcl]) [t d] → Ø/ [+cons +continuant] — # | sof**t** (as) | *n/a* |
| Slur | ax → Ø/ [+consonant] — [r l n] [+vowel] | cam**e**ra | *n/a* |
| Stressed slur | [+vowel +stress] r → er | w**ar**ts | *n/a* |
| Pre-stress contraction | ax → Ø/ [+cons] — [+cons] [+vowel +stress] | s**e**nility | *n/a* |
| Ruh-reduction | r ax → er / [-word bdry] — [-word bdry] | sep**ar**able | *n/a* |
| Transitional stops | | | |
| t-introduction | Ø→ tcl / [+dental +nasal] — [+fricative] | prin[**t**]ce | *n/a* |
| t-deletion | [tcl] → Ø/ [+dental +nasal] — [+fricative] | prin**t**s | *n/a* |

* Syllabic rules can also be implemented as hyperarticulation rules, depending on the baseform representation. For instance, if the baseform for *button* were `[b ah t en]`, then a hypterarticulation rule could be written as `[en]`→`[ax n]`.

Table 2.1: Phonological rules with probabilities from Tajchman *et al.* [1995b]. Probabilities marked with *n/a* were considered for later use, but not calculated in this study.

The fast-speech pronunciation rules were used to expand the number of pronunciations in a baseline dictionary; the new dictionary was subsequently integrated into a recognition system for the Wall Street Journal database [Mirghafori *et al.*, 1995]. These rules provided an average of 2.41 pronunciations per word for the 5K WSJ test set lexicon. The results of running a recognition with this lexicon were insignificantly worse than the base system. When performing an error analysis on the results, we noted that the difference in error rate on a sentence-by-sentence basis between the two systems varied widely; for some sentences the base lexicon did much better, and for others, the new dictionary had up to 75% fewer errors. It has been reported by other researchers [Siegler and Stern, 1995] that modifying the word models by using pronunciation rules has not resulted in any improvements for fast speech in the Wall Street Journal read-speech database. One reason that these fast-speech rules were ineffectual may be that the phonetic reductions and deletions that they model are more often observed in conversational than read speech. Another possibility is that the rules must be applied judiciously to a subset of words (to *function words*, for example), instead of the whole lexicon. Finally, rules may need to be applied at more limited times, depending, for instance, on more local rate estimates, and on previous words or phones.

With probabilistic phonological rules, it is difficult to dynamically change the probabilities of individual rules at run-time. One can compute rule probabilities for several *classes* of inputs that one might see (such as dialect variation), but these must be known beforehand, and data for each condition are not shared easily across conditions. For example, in the New England dialect of American English, the phone /r/ is often deleted in the same contexts where in a Midwest accent the /r/ would remain, so a speech researcher might want to build two models — one for /r/-less and one for /r/-ful dialects. However, the flapping of /t/ occurs at about the same rate in both dialects, but if different dialect models are utilized, then the data for the estimation of probabilities are split across each class, possibly resulting in poorer probability estimation.

Moreover, when it comes to integrating continuous variables like speaking rate (measured in syllables per second, for instance), it is not clear how to build separate models for fast and slow speech — how does one decide where the cutoff for fast or slow lies in the speech rate domain? Ideally, one would like the data to indicate what the optimal cutoff point is. Building separate models also suffers from a data fragmentation problem: speaking rate tends have a roughly Gaussian distribution, so the number of very fast and very slow utterances for rule probability estimation may actually be quite low. Data sharing across models is imperative in this case.

The solution proposed by Ostendorf *et al.* [1997] was to incorporate all of the factors that could affect pronunciation into a single hidden variable, called a *hidden speaking mode*;[13] pronunciation probabilities were learned for each word dependent on the mode. The features used to determine mode were similar to some of the ones used in this thesis, including speaking rate, fundamental frequency, energy, and word duration. Finke and Waibel [1997b] extended this work by incorporating mode dependence in phonological rules, yielding impressive gains on the Switchboard corpus (roughly 2% absolute improvement).

[13] The hidden aspect refers to the fact that this mode is statistically inferred, rather than directly observed by an annotator.

Finke and Waibel's work is related to the work in this thesis, in that I am applying a similar mode-dependence to pronunciation rules automatically induced from data.

### 2.3.3 Induction of rules: stream transformations

Another alternative to pre-stating all of the phonological rules is to try to induce them from data. In general, this is done by learning a transformation model between underlying phonemes and surface phones.

The idea of learning a transformation model to predict phones originated with orthography-to-phone systems [Lucassen and Mercer, 1984; Bahl *et al.*, 1981] rather than phoneme-to-phone systems. The intuitive idea is to learn a probabilistic model $P(pronunciation|spelling)$. This is accomplished by predicting for every letter in the word the probability that a letter produces a particular phone. Bahl *et al.* give the output of their model on the word *humane*: A statistical model is constructed using decision tree techniques, taking into account the surrounding context of letters and previously emitted phonemes.

| **Letter** | h | u | m | a | n | e |
|---|---|---|---|---|---|---|
| **Predicted phoneme** | hh | y uw1 | m | ey1 | n | |

Figure 2.4: Letter-to-phoneme correspondences produced by Bahl *et al.* [1981].

A similar technique was used to predict surface phones from underlying baseform phonemes of words. The seminal experiments with this type of model were conducted by Chen [1990], Randolph [1990] and Riley [1991] using the (then newly constructed) TIMIT database [Garofolo *et al.*, 1993], which gave researchers phone-level hand transcriptions of the read sentences. Subsequent experiments have been conducted with the North American Business News corpus [Riley and Ljolje, 1995; Mohri *et al.*, 1998] and the Switchboard corpus [Weintraub *et al.*, 1997; Riley *et al.*, 1998].

Modeling by decision trees requires careful choices of the feature representation of the phoneme stream. In general, a decision tree is built by examining every feature in turn, choosing the optimal partition of the training data into two subsets based on the values of the feature.[14] However, for phoneme-to-phone mappings, some features (such as the identity of the previous phoneme) have as many as 40 values — requiring the evaluation of on the order of $2^{40}$ possible splits of data, just for that one feature. Chen [1990] used phone identity in her trees, but used phoneme clustering techniques proposed by Sagayama [1989] to reduce the partitioning space. They then annotated each phoneme with indications of syllabic position, stress, foot and word position, whether it was part of a cluster, whether the syllable was open, and whether the containing word was a function word. Riley [1991], on the other hand, used only phoneme context; he converted each phoneme to a 4-tuple of features (consonant-manner, consonant-place, vowel-manner, vowel-place), which reduced

[14] See Section 5.3.1 for a more detailed description of the decision tree algorithm.

the search space considerably. In our Switchboard system, my colleagues and I [Weintraub *et al.*, 1997] chose to represent each phoneme as a vector of binary features, based on features developed to cluster triphones in the HTK recognition system [Young *et al.*, 1994], and added features about stress and lexical syllable information. No direct comparison of these encoding schemes has been made.

The syllable has been utilized frequently in automatic stream transformation since it is an easy source of information to add to decision tree models and has substantial relevance in phonology. Chen [1990] and Weintraub *et al.* [1997] encoded each phoneme with information about lexical stress and syllabic position as additional features for training; we found that this information was used prominently in the decision trees we grew. Hieronymus *et al.* [1992] added both lexical stress and acoustic/phrase-level stress to their speaker-independent 1000-word vocabulary system. They found that the addition of lexical stress gave about a 65% improvement in their system (cutting word error from 2.86% to 0.97%); acoustic stress did not improve the system over just using lexical stress, possibly due to a lack of training data.

In the phoneme-to-phone transformation model, a stream of underlying phones is generated for each utterance by concatenating baseline dictionary pronunciations for each of the words in the utterance. This phoneme sequence is then aligned with the phone sequence that represents the surface pronunciation of the sentence.[15] Commonly, this alignment procedure uses a string-edit-distance algorithm, replacing the distance function with one based on phone confusion matrices, or difference in number of distinctive features. The output of one such alignment procedure is shown in Figure 2.5.

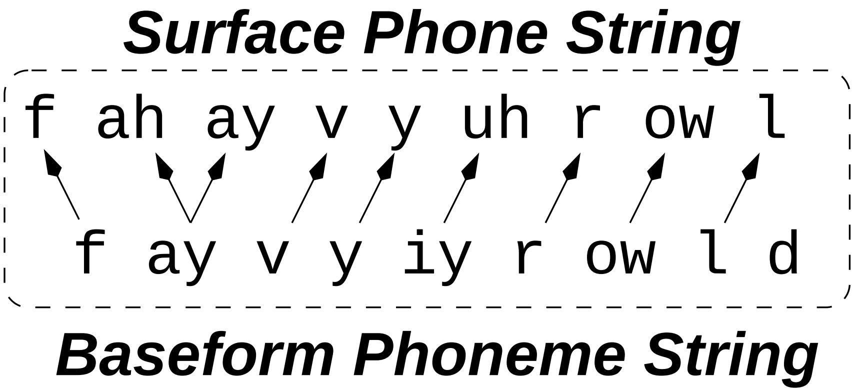


Figure 2.5: Example alignment of *five year old.*

The context of surrounding phonemes is used to predict the transformation from phoneme to surface phone, possibly using the previous output history of the model as additional conditioning information. In our "did you eat" example, the previous output history is important: in learning the transformation from underlying /d ih d y uw/ to surface [d ih jh uw], if the /d/ is transformed to [jh], this should increase the probability

[15] Surface pronunciations can be generated by hand or by automatic techniques like phone recognition.

that /y/ is deleted. Riley [1991] included a dependence on previous model output in his decision trees to accommodate co-occurrences of transformations.

Other learning techniques can be used to automatically model pronunciation variation. Neural networks can be used in the place of decision trees [Miller, 1998; Fukada *et al.*, 1999] to model phonetic variation;[16] Cremelie and Martens [1995] induced phonological rewrite rules directly from data. They automatically determined how phones are realized at word boundaries due to coarticulation effects within hand-generated phonetic transcripts; implementing these phonological rules showed a significant improvement on several Dutch databases. An extended version of their model was presented in [Cremelie and Martens, 1998], in which pronunciations were determined for both word-internal and word-boundary variations from a phone recognition transcript. They found that using both positive rules (declaring where variations may occur) and negative rules (declaring where variations may *not* occur), as well as extending the context window of the phonological rules, improved word recognition in the TIMIT database.

## 2.4 The WS96 transformational model

The model used in this thesis is based on a decision tree stream transformation model developed at the 1996 Johns Hopkins Large Vocabulary Conversational Speech Recognition Summer Research Workshop (abbreviated as WS96).[17] The WS96 "Automatic Learning of Word Pronunciation from Data" group constructed a transformational model to learn the mapping between a baseline pronunciation dictionary and a transcription generated by phone recognition. A key element of our approach was the model $P_{DP}(Q \mid B, M)$ described in Section 2.1, which estimated the probability of a surface phone sequence given a reference phoneme (dictionary) sequence. Employing this model required a dynamic pronunciation model construction algorithm for training, and pronunciation graph building algorithms for testing purposes, described briefly here.

In training, the string of baseform phonemes $B$ was derived from the word transcription by a look-up of the canonical pronunciation of each word in the static dictionary. Phone recognition automatically provided a transcription of the best acoustic models $Q$ that matched the acoustic signal. The surface phone string $Q$ was then aligned to the baseform string $B$ using dynamic programming; the result was that every phoneme in $B$ was mapped to zero or more phones in $Q$. The pronunciation model $P_{DP}(Q \mid B, M)$ was then statistically estimated from the entire training set using the $B \Rightarrow Q$ map. At recognition time, we used the mapping $P_{DP}(Q \mid B, M)$ to generate a graph of pronunciation alternatives $Q$ by expanding the hypothesized models $M$ into a baseform sequence $B$ and then applying the dynamic pronunciation dictionary to determine possible state sequences.

The rest of this section is devoted to a further explanation of the relevant pieces

[16] Miller's work was actually in the related realm of pronunciation modeling for speech synthesis, see his dissertation for an excellent discussion of this related field.

[17] The description of this model is documented further in [Weintraub *et al.*, 1997; Fosler *et al.*, 1996]. An earlier version of the material in this section was co-written with Murat Saraclar as part of an unpublished grant proposal.

of the WS96 model.

### 2.4.1 The baseform and surface symbol streams

In building the map $B \Rightarrow Q$, two symbol streams are needed. The dictionary of the recognizer provides the canonical phonemic representation, or baseform (reference) sequence $B$, given the word transcription of an utterance: the phoneme sequence of each word is obtained by looking up the pronunciation in the dictionary.

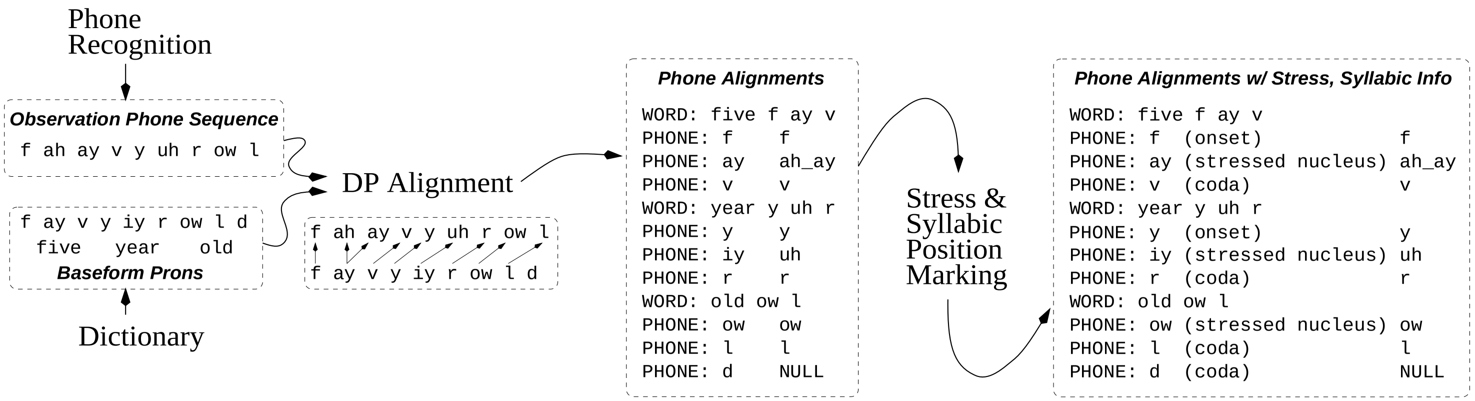
The surface phone sequence $Q$ should be unconstrained by the lexicon, reflecting the actual phone sequence of an utterance as perceived by humans or machine models. Hand-labeled phonetic data has been used to develop pronunciation models [Riley, 1991; Chen, 1990; Riley *et al.*, 1998] in tasks where such data is available. However, phonetically labeled speech data is not always available in adequate quantity. For WS96, only a small amount of hand-transcribed Switchboard data was available; for the Broadcast News corpus (used in the experiments in Chapters 5 and 6), there are no data available. At WS96, we obtained the surface stream automatically using phone recognition. This approach had the advantage that pronunciation models developed using automatic phonetic transcription could also compensate for non-linguistic variations in acoustic models that would not be present in hand-labeled data.

### 2.4.2 Alignment between streams

To determine the surface phone(s) corresponding to each reference phoneme, the streams are converted to aligned pairs of single reference phonemes to one or more surface phones, where a reference phoneme may map to the NULL symbol to indicate a deletion of the phoneme. To incorporate insertions, we allow one reference phone to pair with a contiguous sequence of surface phones. Figure 2.6 illustrates this procedure via an example alignment of the phrase "five year old."

This alignment was performed by a dynamic programming (DP) algorithm that uses a distance metric between phones. A reasonable distance measure was obtained by using a feature bit-vector representation for each phone (including the NULL symbol) and defining the distance between two phones as a weighted Hamming distance between these vectors. Timing information was also used to determine alignment boundaries. The output of the alignment was filtered to exclude the noise due to automatic phone recognition. Some general constraints related to factors such as the deletion rate and the maximum length of the surface phone string corresponding to each reference phoneme, were used to set parameters for filtering.

We also marked the string of reference phonemes with stress and syllabic information, as seen in Figure 2.6. This information, together with the phonemes to the left and right of a reference phoneme constituted the "context" used by the statistical classifier to map a phoneme to a surface phone.



*Phone Alignments*

```
WORD: five f ay v
PHONE: f    f
PHONE: ay   ah_ay
PHONE: v    v
WORD: year y uh r
PHONE: y    y
PHONE: iy   uh
PHONE: r    r
WORD: old ow l
PHONE: ow   ow
PHONE: l    l
PHONE: d    NULL
```

*Phone Alignments w/ Stress, Syllabic Info*

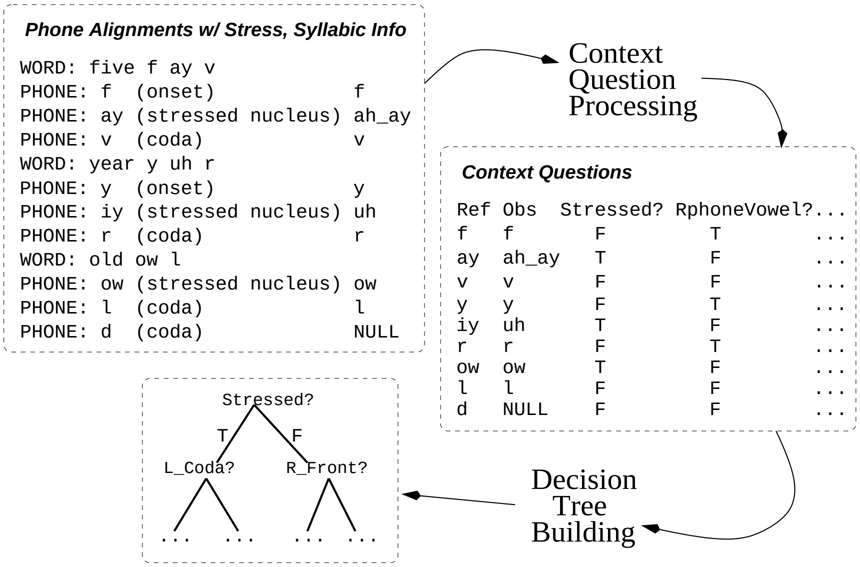
```
WORD: five f ay v
PHONE: f  (onset)            f
PHONE: ay (stressed nucleus) ah_ay
PHONE: v  (coda)             v
WORD: year y uh r
PHONE: y  (onset)            y
PHONE: iy (stressed nucleus) uh
PHONE: r  (coda)             r
WORD: old ow l
PHONE: ow (stressed nucleus) ow
PHONE: l  (coda)             l
PHONE: d  (coda)             NULL
```

Figure 2.6: Aligning reference baseform stream to surface phone recognition stream

### 2.4.3 The transformation

The pronunciation model $P_{DP}(Q|B,M)$ may be estimated by a probabilistic classifier that predicts the surface symbol based on the reference symbol in its context. Decision trees [Breiman *et al.*, 1984] presented themselves as a natural classification technique for this situation, treating the modeling problem as supervised regression for learning the transformation between these two strings.

As Section 5.3.1 describes, decision trees use a greedy top-down optimization procedure to successively partition the set of all contexts of a phoneme, using a set of predetermined "questions" about the context to assign the partitioning. The question that divides the set of surface phones to give the best split (by some criterion) is chosen to partition the set. This continues recursively on each partition of the contexts thus induced, until a stopping criterion is met. The goodness-of-split measure is an indicator of the purity of the set of surface phones. The WS96 model used entropy as the goodness-of-split metric, choosing questions that minimized the total conditional entropy of the surface realizations of a reference phoneme given its (phonemic) context. The decisions made while building the trees are based on fewer and fewer data as the tree-building process progresses; this causes unreliable splits toward the leaves. To remedy this, the trees were pruned using cross-validation on a held-out test set.



***Phone Alignments w/ Stress, Syllabic Info***

```
WORD: five f ay v
PHONE: f  (onset)           f
PHONE: ay (stressed nucleus) ah_ay
PHONE: v  (coda)            v
WORD: year y uh r
PHONE: y  (onset)           y
PHONE: iy (stressed nucleus) uh
PHONE: r  (coda)            r
WORD: old ow l
PHONE: ow (stressed nucleus) ow
PHONE: l  (coda)            l
PHONE: d  (coda)            NULL
```

***Context Questions***

```
Ref Obs   Stressed? RphoneVowel?...
f   f        F         T        ...
ay  ah_ay    T         F        ...
v   v        F         F        ...
y   y        F         T        ...
iy  uh       T         F        ...
r   r        F         T        ...
ow  ow       T         F        ...
l   l        F         F        ...
d   NULL     F         F        ...
```

Figure 2.7: Building decision trees using aligned data streams

For each reference phoneme, we built a decision tree that asked questions about the baseform and word context (Figure 2.7). These context questions included information about the reference stream itself (such as stress, syllabic position, or the classes of neighboring phones), or the past output of the tree (including the identities of surface phones to the left of the current phone). From this, the tree learned, for any given context, a probability distribution over the set of the surface phone(s) determined by the alignment step of Section 2.4.2.

In the WS96 model, the decision tree model estimated the surface stream probability $P_{DP}(Q|B, M)$ as $\prod_{i=1}^{N} P_{DP}(q_i \ |B)$, assuming that each surface phone was independent of both the previous surface phones and the word stream given the baseform stream. In the experiments in this thesis, the word models $M$ are re-introduced into the contextual conditioning; other factors, including word predictability and speaking rate, can affect the probability of $Q$ as well. The dependence on previous surface phones is also addressed by modeling phone pronunciation distributions jointly at the syllable and word levels.

### 2.4.4 Pronunciation networks and dynamic dictionaries

The incorporation of decision trees into recognition was relatively straightforward; a full description of the technique is found in Chapter 5. In short, each pronunciation distribution at the leaf of a decision tree can be thought of as a small pronunciation network. The network at each leaf has two nodes, and every surface phone at the leaf is represented by an arc between the nodes.[18] At test time, a baseform sequence for a word can be transformed into a pronunciation graph by filtering each baseform phone down through the decision trees and finding the sub-graph for the phoneme at the leaf. Concatenation of the sub-graphs gives a network for the entire word. This network may be pruned if needed and as easily provides the $n$-best pronunciations for a word in context.

Once we had a pronunciation network for a given word, we had the option of ignoring the contextual effects on the word, replacing the word's entry in the pronunciation dictionary of the recognizer with the frequent new pronunciations of the word (static baseform replacement). In other experiments, we required that the transformation be context-dependent even at word boundaries, making the dictionary entries dependent on the previous and next words. This dictionary was known as a *dynamic dictionary* to distinguish it from the static baseform replacements described above. The results of the WS96 experiments are described in Section 6.3.

## 2.5 Effects of linguistic variables on production and perception of speech

Within the framework of a dynamic pronunciation model, factors other than the baseform phoneme string may be useful for predicting how pronunciations vary. In this section, I review some of the linguistic and psychology literature that describes relevant linguistic features beyond the phoneme that affect pronunciation.

[18] Deletions were represented by a null transition between the two nodes; for insertions extra nodes were inserted in the graph between the start and end node.

### 2.5.1 Word predictability

Linguists have recognized that word frequency affects the perception and production of phones. In an extreme example of this, Ganong [1980] had subjects discriminate between /t/ and /d/ in the word pairs *dash–tash* and *dask–task*. For each word pair, a series of words were created with increasing voice onset times, so that the percept of the initial phoneme changed from /d/ to /t/ at some point in the series. Subjects listened to samples randomly chosen from the series, and were asked to classify the initial phoneme as /t/ or /d/. Ganong reported that the perceptual change point in each of these series was different; in the *dash–tash* case the voicing onset time at the perceptual shift was longer (*i.e.*, subjects preferred /d/ over /t/), showing that listeners used lexical knowledge by preferring an English word over a nonsense word.

Linguists have also found that phone deletions and reductions are more likely to occur in high-frequency words [Hooper, 1976; Labov, 1994; Bybee, in press].[19] The phonemes /t/ and /d/, for instance, are twice as likely to delete word-finally in a high-frequency word as in a low-frequency word, according to Bybee [in press]. Zwicky [1972b] also postulates that deletions of initial /h/,/w/, and /dh/ can also occur in function words.

Semantic context beyond the word frequency also affects the production and perception of speech. Lieberman [1963] compared examples of words excised from the speech signal in predictable and unpredictable contexts, finding that out of context, predictable words were more difficult for subjects to understand than unpredictable words. This difficulty is correlated with the fact that the examples of redundant words were on average shorter in length than unpredictable examples, and they often had a smaller signal amplitude. Pollack and Pickett [1964] demonstrated that word perception is influenced both by the syntactic context of following words and by the acoustic context of the word itself. Subjects were asked to guess the initial word of an extracted speech segment; the following $n$ words ($n = 1, 2, 4, 7$) were included as context in the speech signal. To eliminate the influence of syntactic context, they presented the full written transcript of seven words to subjects;[20] they found that identification of the initial word in the excerpt was still dependent on the number of words heard in the segment, signifying that the acoustic context could influence perception independent of the syntactic context. When the test was repeated with the written transcript removed (with different subjects), the average intelligibility was lower — thus, syntactic context was also an important factor in determining identification rates. In a demonstration of the long-term effects of semantic information, Fowler and Housum [1987] showed that when a word is spoken a second time within a monologue, its duration is generally shorter than than the initial occurrence of the word (*i.e.*, when the word constituted *new* information).

In the Switchboard corpus, the effect of predictability may be even more pronounced. In a recent survey of pronunciations of the ten most frequent words[21] within a transcribed portion of the Switchboard corpus, Greenberg [1997a] cites an average of over 60

[19] *Function words*, or words that perform mostly syntactic functions and carry little semantic information, are usually high-frequency words.

[20] In this way, the subjects knew the syntactic and semantic context of the following words, even though they did not have access to all of the acoustic context.

[21] This list consists of *I, and, the, you, that, a, to, know, of,* and *it*.

pronunciations per word. In some cases, the syntactic predictability of these frequent words is so strong that there is no phonetic (segmental) evidence for the words at all, particularly in the spectrograms of the utterance. Yet, transcribers can hear these non-phonetic words when listening to the entire phrase [Greenberg, 1997a; Fosler, 1997]. In Section 3.2.4, I will describe my investigations into this effect.

### 2.5.2 Rate of speech

An under-utilized factor in pronunciation modeling is speaking rate. In general, speaking rate is defined in terms of words, syllables, or phones uttered per unit time, although linguists tend to use the latter two because of the variability in the length of words. My colleagues and I [Mirghafori *et al.*, 1995, 1996], along with others [Siegler and Stern, 1995; Fisher, 1996a] have also shown that ASR word error rates track better with rate calculations based on phones than on words.

#### Durational constraints due to speaking rate variability

Obviously, one factor of increased speaking rate is that the durations of phones decrease. In a series of reports, Crystal and House [1988] provided the linguistics community with quantitative data on the duration of phones for slow and fast speakers in stories read aloud. The extreme speakers showed 5-8% variation from the mean phone duration. Since this was for read speech, it is probably a *lower* bound for the durational variation that one would see in conversational speech.

#### Phonetic differences due to speaking rate variability

More crucial to pronunciation modeling is the fact that rate of speech variation can also affect phone perception and production. In one study, Miller and Liberman [1979] experimented with changing the initial consonantal transition duration of a syllable that was perceptually ambiguous between /ba/ and /wa/. By lengthening the initial consonantal transition they could elicit a perceptual change from /ba/ to /wa/; moreover, the length of transition duration at which the perceptual shift occurred was dependent on the duration of the entire syllable.[22] In a further investigation of this phenomenon, Miller and Baer [1983] had speakers produce /ba/ and /wa/ syllables in time with metronome flashes, varying the speed of flashing to increase speaking rate. They found, for all speech rates, that the initial formant transitions of /w/ were longer than /b/; however, when the data were pooled across all speaking rates, the distribution of these onset transition durations of /b/ and /w/ overlap. From these two experiments, Miller and Baer concluded that humans take the speaking rate (or, conversely, syllable duration) into account when trying to disambiguate ambiguous /b/-/w/ onsets.

[22] From this study, they concluded that humans are normalizing for rate when they process speech. Pisoni *et al.* [1983] subsequently showed that rate normalization may be a general human perceptual property, rather than a linguistic property. The lesson is that rate normalization is occurring at some level in humans, and therefore rate information may be useful to machine-based models of speech.

In another set of experiments, Port [1979] and Miller and Grosjean [1981] examined the perception of /p/ and /b/ under different speaking rates. In the experiments of Miller and Grosjean, subjects were asked to distinguish between sentences of the following nature:

- The tiger that the man chased was rapid.
- The tiger that the man chased was rabid.

Sentences were read by a professional speaker with different articulatory rates and rates of pausing. Stimuli were prepared by taking the sentence and replacing the /p/ or /b/ segment in the carrier word with a variable length of silence. Miller and Grosjean found that for a particular silence length, as the articulation rate increased, speakers tended to judge the silent portion to be more /p/-like than /b/-like. These perceptual curves were dependent on the articulatory rate; the effect was not as marked for pause rates.

Cooper *et al.* [1983] studied the effect of stress and speech rate on palatalization. They looked at the palatalization of /d/ and /t/ to /jh/ and /ch/, respectively, as in the following sentences:

- Did you eat? → Di*j*ou eat?
- I bet you lost! → I be*ch*a lost!

They found that the most important factor conditioning this phonological process was sentence-level stress. Cooper et al. had their subjects read three versions of each sentence, with different words receiving prominence in each one:

1. (No stress) Did you eat?
2. (D-word stress) DID you eat?
3. (Y-word stress) Did YOU eat?

Palatalization was most likely to occur if neither the D-word (*i.e.*, did) or the Y-word (*i.e.*, you) received stress (case 1); the frequency of occurrence was reduced under D-word stress (case 2), and even further under Y-word stress (case 3). However, once sentence-level stress was accounted for, speaking rate affected the palatalization probability significantly; faster speakers tended to palatalize 10-20% more.

One important note is that all of these studies dealt with either isolated syllables or calculated speech rate over an entire sentence. However, Summerfield [1981] showed that a local calculation of rate is perceptually important; the rate of speech of words not adjacent to a tested word or phone is less influential in the perception of the stimulus. Many factors can affect local speaking rate, including syllable stress, syllabic complexity, pre-pausal lengthening (in English), and the part-of-speech of words. These factors are often used in prediction of syllable or phone durations (*e.g.*, for speech synthesis) [Fant *et al.*, 1992; Campbell, 1989, *inter alia*]. Klatt [1979] developed one of the earliest rule-based models of duration for English; Carlson and Granström [1989] applied this model to predicting duration of Swedish phones, and found a cyclical pattern of errors (sometimes overestimating duration and sometimes underestimating), corresponding to local changes in speaking rate.

### 2.5.3 Syllables in phonology

In this thesis, I model pronunciations of phones, syllables, and words. While models at the phone and word level have been common within the ASR community, not many systems have utilized syllable models for pronunciation. Nonetheless, there exist linguistic motivations for considering models of this type; in this section, I present a short review of the syllable literature.

Linguists have claimed since the early 1900s that the syllable is an organizational unit in speech. However, the concept of syllable is difficult to pin down exactly, particularly in English [Kenestowicz and Kisseberth, 1979]. Various descriptions have included syllables as peaks of sonority, pulses of sound energy, groupings of speech movements, and basic units of speech perception [Ohde and Sharf, 1992]. However, all of these definitions are problematic in one way or another — for example, the English syllable *spa* has two peaks of sonority [Kenestowicz and Kisseberth, 1979]. Humans, however, seem able to parse the speech stream into syllables without awareness of how they do it. Greenberg [1997a] claims that the syllable is important for temporal organization in speech perception. It remains to be seen how humans derive this information from the speech stream or utilize it in perception.

The syllable was prominent in phonological theories until the early 1960s, when the generative phonological theories of Chomsky and Halle [1968] relegated the syllable to the position of derived unit, subservient to the phonological features that constituted their theory. However, Kahn [1980] (among others) argued that the formulations of syllable phenomena predicted by generative theory were, at best, awkward. In his thesis, he provided an analysis of English syllabification, which was subsequently used in a publicly available computer syllabification program written by Fisher [1996b]. Kahn's position is that one of the reasons why syllabification of English is difficult is the presence of *ambisyllabicity* — the assignment of one phoneme to both the end of one syllable and the beginning of the next, as seen in the word *coming*, where it is unclear if the syllable boundary should precede, follow, or even divide the /m/.

Information about the position of phonemes in a syllable greatly simplifies descriptions of some phonological phenomena. Kenestowicz and Kisseberth [1979] point out that the phoneme /t/ syllable-initially (as in *top*) is realized as an aspirated [t], whereas after an /s/ in the onset of a syllable (e.g. *stop*), /t/ becomes unaspirated. Kahn [1980] also provides an example of /r/ in New York City dialects.[23] Consider the following two

[23] Kahn also analyzes the realization of /r/ in New England and British dialects and concludes that synchronically, there is no underlying /r/ in /r/-less words like *card* and *tuner*, essentially making them homophonous with *cod* and *tuna*. However, he cannot bring this analysis to bear on New York accents, since the distribution of /r/ pronunciations are different: *tuna* (when not followed by a vowel) and *cod* in the data he presents are never pronounced with /r/ (as opposed to New England accents) [Labov, 1972]. Native New York speakers have challenged this claim, though.

sentences, with canonical New York accent transcriptions:

- Park the *car* later.

  /p ao k dh ax k ao l ey dx ax/

- Park the *car* in here.

  /p ao k dh ax k ao r ih n hh iy ax/

Kahn postulates that the general rule of /r/-deletion in effect here is that /r/ can be dropped when it is exclusively in the coda of a syllable. The /r/ remains undropped in the second case because it becomes *resyllabified* — associated with the following syllable in fast speech contexts.[24]

### 2.5.4 Discussion

The evidence in the literature indicates that word and phone pronunciations can depend on various contextual factors. In this short survey alone, speaking rate, word predictability, and the syllabic structure of words have been shown to affect pronunciations. This is not an exhaustive list; for example, the dialect of the speaker can be an important determining factor for pronunciations.[25] A model of pronunciation for ASR systems may do well to take these variables into account and determine pronunciations dynamically. This thesis tries to integrate some of these contextual factors into the probabilistic model $P_{DP}(Q|B, M, \ldots)$.

It is an open question whether these factors, which have been shown to affect human phone production and perception, will have an effect when added to automatically learned models of pronunciations. Machine acoustic models, while linguistically seeded, are not the same as the human perceptual system. Some of the phonetic variability due to the factors of speaking rate and word predictability may already be accounted for in the ASR system. On the other hand, the acoustic models may be affected in a different way by such factors as increases in speaking rate. This work tries to address whether these factors can improve an automatic learning scheme.

## 2.6 Summary

The speech recognition problem is defined as selecting the best word sequence from the space of possible hypotheses, given an acoustic waveform. To find the best utterance,

[24] Kahn also describes a hyper-correction process, in which NYC speakers add /r/ spontaneously to words like Indiana /ih n d iy ae n ax r/. The allowed context of this phenomenon is only syllable-final, as in the /r/-deletion rule.

[25] While speaker accent is an important factor in determining pronunciations, accent-specific modeling is not a focus of this thesis, partly because it can be difficult to detect the accent of the speaker (although advances are being made in this area; see, *e.g.*, Lincoln *et al.* [1998] for a description of unsupervised detection of British versus American accents). Humphries [1997] provides one approach to adapting pronunciation models to a new accent.

ASR systems use three statistical models: an acoustic model that provides the probability of acoustics given an HMM state sequence, a pronunciation model that gives the mappings between state sequences and words, and the language model that furnishes a prior probability over all model sequences. The pronunciation model has an important role as the interface between the other two models; it must accommodate both linguistic variability and variability from the acoustic models.

Linguists, particularly in the field of phonology, has provided tools for ASR pronunciation modelers to use in accounting for this variability. In particular, phonological rules have been useful for generating a large set of potential pronunciations that models can select from; rule probabilities can be used to select appropriate pronunciations from a large corpus. These rules can also be automatically induced using techniques such as decision trees. The fully automatic pronunciation learning system developed at WS96 serves as the basis for the work in this thesis.

A review of the linguistics literature revealed that pronunciations are dependent on more than just the segmental context. Including the features of speaking rate and word predictability may allow ASR systems to judiciously choose the set of pronunciations to include while running — a dynamic pronunciation model. This may be important for modeling the linguistic coarticulation found in unconstrained-vocabulary continuous speech databases, particularly for conversational speaking styles.

# Chapter 3

# Speaking Rate, Predictability, and Pronunciations

The previous chapter introduced speaking rate and word predictability as recognized influences on word pronunciation. In this chapter[1] I investigate the relationships between speaking rate, word predictability, and pronunciations within the Switchboard corpus of human-to-human conversations. We can refer to these factors as extra-segmental factors, because they do not depend on the identities of phones, syllables, or words. In the first section, I describe how speaking rate and word predictability are estimated in this study, and I explain the metrics used to evaluate differences in pronunciation within the corpus that are due to these factors. The second portion of the chapter is devoted to a study of pronunciation variation in words, syllables, and phones that can be related to these factors.

[1] Some of the experiments in this chapter and the next have been reported previously in Fosler-Lussier and Morgan [1998], and will also appear in Fosler-Lussier and Morgan [in press].

## 3.1 Experimental Design

In order to determine how extra-segmental factors can affect word pronunciations, one must first determine a set of measurements for these factors. In this section I discuss several measures of the rate of speech and of the predictability of words that were used in this study.

### 3.1.1 Speaking rate

Speaking rate is generally measured as a number of linguistic units per second, although the choice of units has been subject to debate. In previous work at ICSI, we have shown that using units other than words per second (e.g., phones per second) as a metric allowed us to predict word error more reliably in ASR systems [Mirghafori *et al.*, 1995]. In contrast, Fisher [1996a] showed that, for the Hub3 North American Business News task, the difference between words and syllables per second was insignificant for prediction of recognizer error rate. However, Fisher preferred the syllabic measure because it was likely to be easier to calculate independent of any recognizer.

For this study, syllabic rate was chosen as the metric of speaking rate. Syllables are far less likely to be deleted than phones; in the Switchboard corpus, the phone deletion rate is roughly 13% [Weintraub *et al.*, 1997], whereas complete phonetic deletion of the syllable occurs only 2.5% of the time. Since phone deletions are one of the phenomena that should be detected by pronunciation models, syllabic rate is a more stable measure for this purpose.

Syllabic rate can be determined from speech data in several ways. In this study I use *transcribed syllable rate*, which is determined from syllabic boundaries notated by linguistic transcribers. More specifically, the *interpausal rate* is determined by counting the number of syllables between transcribed silences and dividing by the amount of time between pauses. Because of the required human transcription, this particular measure is generally not determinable at recognition time for interactive systems. Nonetheless, because speaking rate estimators can sometimes be unreliable (particularly for spontaneous speech), this metric was used as an irreproachable measure for determining the effects of speaking rate on pronunciations.

When syllabic annotations are not available, one can also determine a syllabic rate from the alignment of the word transcription to speech data (*aligned syllable rate*); since the syllable deletion rate is roughly 2.5% and the insertion rate is negligible, this corresponds closely to the transcribed syllable rate.

At the run-time of the recognizer other metrics must be used, because it is not feasible to have a human transcribing the speaking rate as the recognizer is running. Mirghafori *et al.* [1996] described the tactic of running the recognizer twice, using the first pass to hypothesize sound unit boundaries and hence the speaking rate, which would then be incorporated in a second pass, providing a measure called *recognition rate*. Aside from the additional computation, this method requires the assumption that the speaking rate determined by a potentially erroneous recognition hypothesis is sufficiently accurate. For difficult tasks such as conversational speech recognition this is often not the case, particularly for unusually fast or slow speech.

Alternatively, one can use signal processing or classification techniques to estimate speaking rate directly from the acoustic signal [Kitazawa *et al.* 1997; Verhasselt and Martens 1996 *inter alia*]. At ICSI, we have also derived such a measure, dubbed *mrate* for its **m**ultiple **rate** estimator components [Morgan and Fosler-Lussier, 1998]. The measure correlates moderately well with transcribed syllable rate ($\rho \sim .75$), although it tends to underestimate the rate for fast speech. Mrate is further described in Section 6.4.2.

Other methods of determining speaking rate are possible: Campbell [1992] calculated the distribution of durations for each type of phone in a Japanese corpus; from this each phone instance in the corpus could be assigned a z-score (standard deviation) based on its duration.[2] This metric was useful for predicting various linguistic effects, such as the shortening of /a/ both after unvoiced plosives and sentence-finally in Japanese, but z-scores were not applied to the task of localized speaking rate determination in this study. Given the constraints of the task in Campbell's study (a single speaker reading newspaper and magazine texts), it is not clear whether z-scores would be usable for detecting speaking rate in a multi-speaker spontaneous speech database.

### 3.1.2 Word predictability

The most obvious candidate for determining word predictability is the unconditional probability of the word (*i.e.,* $P(\text{word})$), determined from the number of instances of the word in the reference transcription of the entire corpus. This is known in ASR parlance as the *unigram probability* of the word. The results reported here use the base 10 logarithm of the unigram probability.[3]

However, the predictability of a word in its local context may also have an effect on its pronunciation. One simple measure of the localized predictability used by ASR systems is the *trigram probability* ($P(\text{word}_n|\text{word}_{n-2}, \text{word}_{n-1})$) — the probability of the word given the previous two words. In the case where the trigram probability was not available, a Good-Turing backoff strategy was employed [Good, 1953; Katz, 1987], in which the trigram probability is estimated from a bigram probability ($P(\text{word}_n|\text{word}_{n-1})$) and a weighting factor — a strategy employed by most ASR systems. While I could have chosen to examine bigram probabilities instead of trigram probabilities (one of the ASR systems evaluated in the next chapter uses a bigram grammar, the other a trigram grammar), I wanted to include as much contextual information as possible in order to distinguish this measure from the unigram score.

One can also imagine more elaborate models, such as semantic triggers, word collocations, or syntactic constraints, that could be used to predict when a word is more likely. However, for the sake of simplicity, only models conveniently available to most speech

[2] This assumes that the durations of each phone type in the corpus have a normal (Gaussian) distribution; as Campbell notes, the distributions are actually slightly skewed from normal, having more longer phones than shorter phones.

[3] Most ASR systems represent probabilities in the logarithmic domain, so that multiplication of probabilities becomes addition of log probabilities, an operation that is faster on standard computer processors.

recognizers were used.

The studies in this chapter will evaluate pronunciations of phones, syllables, and words. However, when unigram probabilities or trigram probabilities are given, they are computed on the word level. This means that unigram and trigram probabilities in the syllable and phone investigations are of the word that contains the phone or syllable, not the unit itself.

### 3.1.3 Pronunciation database generation

The speech data from the Switchboard corpus used for this study are a subset of the complete database, consisting of approximately four hours of phonetically hand-transcribed utterances provided by ICSI for the Johns Hopkins Summer Research Workshop series [Greenberg, 1997b]. About one half hour of this data was from the development test set, while the rest was from the training set. Starting from an automatic syllabic alignment generated by the Johns Hopkins HTK recognizer, linguists from ICSI realigned the syllable boundaries and identified syllables with their phonetic constituents.[4]

Using an automatically syllabified version of the Pronlex dictionary[5] (LDC 1996), I generated a mapping from dictionary baseforms to these hand transcriptions using a dynamic programming technique developed by Weintraub *et al.* [1997]. The procedure uses a string-edit-distance algorithm, where the distance metric between two phones $\phi$ and $\psi$ is given by:

$$d(\phi,\psi) = \sum_{f \in \text{Features}} g(f(\phi), f(\psi)) \tag{3.1}$$

where

$$g(f(\phi), f(\psi)) = \begin{cases} 0 & \text{if } f(\phi) = f(\psi) \\ 1 & \text{otherwise} \end{cases} \tag{3.2}$$

A phonetic feature is a binary question about the phone (*e.g.*, "Is this phone front?" or "Is this phone nasal?"); at ICSI, Tajchman [1994] defined a set of 24 binary phonetic features (*e.g.*, front, nasal, high) similar to the features used in *Sound Patterns of English* [Chomsky and Halle, 1968].[6] Under this scheme, the phones `[f]` and `[v]` differed in two phonetic features:

$$\texttt{[f]}: \begin{bmatrix} +tense \\ \text{-}voiced \end{bmatrix}, \texttt{[v]}: \begin{bmatrix} \text{-}tense \\ +voiced \end{bmatrix}$$

[4] See also **http://www.icsi.berkeley.edu/real/stp** for more information about the Switchboard Transcription Project.

[5] The Pronlex dictionary was syllabified at the 1996 Johns Hopkins Summer Research Workshop (WS96) using Fisher's [1996b] automatic syllabification program, which is based on Kahn's [1980] thesis. This dictionary is used by several ASR systems (without syllabification), including the HTK recognizer analyzed in Section 4.1. Thanks to Barbara Wheatley and others at WS96 for help with this lexicon.

[6] These features have been incorrectly attributed to Withgott and Chen [1993] in some of the pronunciation modeling literature. In fact, Withgott and Chen do not insist on binary features; rather, they point out that binary features can lead to bizarre formulations of phonological phenomena [Withgott and Chen, 1993, p. 6].

The distance score $d(\texttt{[f]}, \texttt{[v]})$ was 2, whereas the distance score between [f] and the more dissimilar vowel [ae] was much higher (11).[7]

Every baseform (dictionary) phone was mapped to zero or more hand-transcribed phones; deletions caused the baseform phone to be mapped to zero phones, and insertions caused the dictionary phone to be mapped to multiple transcription phones. Where multiple pronunciations existed in the dictionary,[8] the closest baseform (in terms of the distance metric $d$) to the realization was used. The output of the alignment procedure was a map $\alpha$; each instance of a baseform phone $\phi$ in the database was mapped to an $n$-tuple of realized phones:

$$\alpha(\phi) = \langle \psi_1, \psi_2, \ldots, \psi_n \rangle \tag{3.3}$$

Typically, $n$ was between 0 and 2.

These pronunciation maps were created for every baseform phone in the transcribed database. I then annotated every word (and its syllabic and phonetic constituents) with measures of speaking rate and word predictability, namely transcribed syllable rate of the interpausal region, unigram frequency of the word, and trigram probability in the utterance context. The result was a database of pronunciation variation for every word, syllable, and phone in the transcribed portion of the Switchboard corpus.

### 3.1.4 Pronunciation difference metrics

It was initially difficult to determine the best method of characterizing pronunciation behavior as a function of the independent factors. I experimented with a number of metrics; each has some advantages and disadvantages. Each of these metrics was designed to describe pronunciation variations of individual phones, syllables, or words as factors such as speaking rate change, as well as describing the effects of these factors on the pronunciation of each of these units in the corpus at large.

**Probability of a single pronunciation**

The simplest measure of pronunciation variation is to determine the probability of canonical pronunciations as extra-segmental factors vary.

This metric is particularly useful for estimating how well the dictionary pronunciations in the baseline recognizer match the transcribed data. I define a pronunciation as *canonical* if it matches a listed pronunciation from the ASR dictionary. For spontaneous speech the canonical and most frequent pronunciations often differ, so the probability of the single most frequent pronunciation was noted as well, assuming that a system that performs automatic baseform learning would also have that particular pronunciation in its dictionary.

[7] Ohala [1985] points out that the use of *tense* as a phonetic feature is problematic, in the same way that *flat* is problematic: the term is informal and impressionistic, and it describes a class of phonetic phenomena that arise from disparate phonetic configurations. The choice of phonetic feature systems is not critical for this work, however. All that is required is some estimate of distance between two phones.

[8] The Pronlex dictionary has mostly one pronunciation per word; the average number of pronunciations per word for the 22K lexicon was 1.07, whereas for the 100 most frequent words in Switchboard, this average is 1.14.

Using this metric, I was able to determine when the probability of a particular pronunciation changed significantly[9] due to a change in one of the factors. A drawback to this metric is that analysis becomes more difficult when tracking more than just a few pronunciations.

**Entropy**

This is a traditional measure for pronunciation learning systems (*e.g.*, Riley 1991) and is a good measure of the number of pronunciations of a unit in a training set, as well as the relative frequency of the alternatives. For a set of pronunciations $X$ with a probability distribution estimate $p$, the entropy $H(X)$ in bits is defined as:

$$H(X) = - \sum_{x \in X} p(x) \log_2(p(x)) \tag{3.4}$$

Entropy, in this case, is a measure of how skewed a distribution is toward one pronunciation. When all pronunciation alternatives are equally likely, entropy is at its highest; it is lowest when one pronunciation has a probability of one and the rest are zero (*i.e.*, the unit is *definitely* pronounced this way). This measure becomes unwieldy, however, if one tries to use it to predict the relative entropy of a particular test set. Pronunciation models are typically pruned to some cutoff (assigning zero probability to some test events), which causes relative entropy to approach infinity. Simple measures of entropy also treat all pronunciations as distinct and unrelated. Thus, a pronunciation distribution for *this* $\langle p(\texttt{[dh ih s]}) = 0.8, p(\texttt{[d ih s]}) = 0.2\rangle$ and a second distribution $\langle p(\texttt{[dh ih s]}) = 0.2, p(\texttt{[dh]}) = 0.8\rangle$ have the same entropy, although the second distribution intuitively seems less canonical.

**Phonetic distance score**

I also developed a metric that was smoother than the hard binary decision of whether a pronunciation was canonical or not by using the phonetic feature distance ($d$) between the two pronunciations as described in Section 3.1.3. The formula for the distance score between two syllables, $\sigma_{\text{base}}$ and $\sigma_{\text{transcribed}}$, depended on each phone $\phi_{\text{base}}$ of the syllable $\sigma_{\text{base}}$, and the phonetic alignment $\alpha$ from equation 3.3:

$$D(\sigma_{\text{base}}, \sigma_{\text{transcribed}}) = \sum_{\phi_{base} \in \sigma_{base}} d(\phi_{\text{base}}, \alpha(\phi_{\text{base}})) \tag{3.5}$$

where $\alpha(\phi_{\text{base}})$ returns the aligned transcription phones.[10]

This distance can be interpreted as a measure of how far the realized pronunciation has deviated from the expected pronunciation. Rather than treating pronunciations as discrete entities, as is done in the entropy and single probability measures, this score integrates the distance between phonetic features associated with each string of phones. This procedure can also be extended to give an aggregate corpus score using

[9] When significance is reported here, I mean that two distributions are significantly different at $p \leq 0.05$ using a difference of proportions test.

[10] Technically, $\alpha$ returns an $n$-tuple of phones, but here the interpretation of the distance metric $d$ is extended to include the concept of insertions and deletions: for each insertion or deletion, the insertion/deletion penalty distance used in the alignment procedure is added to the score total.

a particular pronunciation model; the distance between each baseform pronunciation in the model and the target phone sequence is weighted by the probability of the baseform pronunciation. However, as this measure is not a probabilistic quantity, it is difficult to give it a statistical or information-theoretic interpretation.

## 3.2 Relationships between speaking rate, word predictability, and pronunciation differences

In this section I present statistical analyses that show the relationship between speaking rate, word predictability, and pronunciations in the Switchboard corpus [NIST, 1992], a collection of telephone conversations between two strangers in which speakers were asked to talk about one of hundreds of topics and were recorded for up to five minutes. I begin with an analysis of how pronunciations of individual words deviate from the canonical.

### 3.2.1 Pronunciation statistics for words

In a pilot experiment to show the effects of the features on a coarse level, I extracted the word-pronunciation pairs for the 117 most frequent words from a two-hour subset of the transcriptions from the training set. Each word had at least 40 occurrences in the set. For every selected word, the pronunciation population was divided into two halves: words above the median speaking rate and words below the median speaking rate, giving two pronunciation probability distributions. I compared the probability of both the most common transcribed pronunciation and the canonical pronunciation (as given in the Pronlex dictionary) between partitions. A sample comparison for the word "been" is shown in Table 3.1.

In this case, the probability of the canonical pronunciation `[b ih n]` drops significantly for the faster half of the examples. The distribution of alternate pronunciations changes as well: the reduced-vowel variant, `[b ix n]`, occurs only in the fast speech examples. A significant difference in probability for the canonical pronunciation between fast and slow speech was a common occurrence in the 117 most frequently occurring words; a significant ($p$ <0.05) difference in canonical pronunciation probability for 30% of the words was found due to rate differences. For speaking rate differences that were significant, a faster rate indicated fewer canonical pronunciations, without exception.

The partitioning was repeated, only this time separating words with high trigram probability (*i.e.*, more likely words) from low trigram probability. Table 3.2 displays the number of words with significant differences in pronunciation probability due to each factor. When the trigram probability was used as the splitting criterion, 18% of the words had a significant shift in canonical pronunciation probability. Similar results were seen with the most likely pronunciations.

As with speaking rate, a higher trigram probability (*i.e.,* if the word was more likely) also meant a decrease in canonical pronunciation probability. It is noteworthy that the words that showed a significant difference in canonical pronunciation probability were

| Pronunciation | Low Syllable Rate | High Syllable Rate |
|---|---|---|
| Canonical | 0.6087 b ih n | 0.3636 b ih n |
| Alternatives | 0.1304 b eh n | 0.1818 b ix n |
| | 0.0870 b ih nx | 0.1364 b ih nx |
| | 0.0435 b ih n n | 0.0909 b ih |
| | 0.0435 b eh n | 0.0909 b eh n |
| | 0.0435 b eh nx | 0.0455 b eh |
| | 0.0435 b ih | 0.0455 b ah n |
| | | 0.0455 v ih n |

Table 3.1: Distribution of the pronunciation probabilities for 45 realizations of the word "been."

| Number of words (out of 117) with significant pronunciation differences | | | |
|---|---|---|---|
| Pronunciation type | Dividing metric | $p$ <0.05 | $p$ <0.01 |
| Canonical | Syllabic rate | 35 (29.9%) | 12 (10.3%) |
| Canonical | Trigram probability | 21 (17.9%) | 5 (4.3%) |
| Most likely | Syllabic rate | 31 (26.5%) | 12 (10.3%) |
| Most likely | Trigram probability | 20 (17.1%) | 7 (6.0%) |

Table 3.2: The number of words in which significant pronunciation probability differences were seen based on syllable rate and language model probability for the most frequent words in Switchboard. Pronunciation changes were calculated for the canonical pronunciation and most likely surface pronunciation for each word.

often distinct between the rate and language model lists; at a significance level of $p < 0.05$, only nine words were affected by both trigram probability and speaking rate, meaning that 40% of the words had significant differences in canonical probability due to either rate or trigram score. I could find no clear-cut rationale for why these two lists were mostly distinct.

Probability shifts in the canonical pronunciation were more often due to speaking rate than to word predictability; this is understandable, since the distribution of the examined words is already skewed with respect to trigram scores. In order to obtain enough data for per-word scores, I chose to study the most frequent words, which are *a priori* more likely to have higher trigram scores — the range of scores for these words is smaller than that for the general population of words.

Thus, using a relatively gross measure of pronunciation variation, I was able to find interrelations between extra-segmental factors and the probability of word pronunciations. However, the 117 studied words covered only 68% of the corpus.

### 3.2.2 Pronunciation statistics for phones

In order to better characterize pronunciation variation in a wider cross-section of the corpus, we decided to look at pronunciation statistics for individual phones and syllables, for which we had more data. For each dictionary phone, I extracted the corresponding hand-transcribed phone(s), along with the applicable speaking rate. In this study, I examine the overall trends for the phone alternations within the corpus.

As seen in Figure 3.1, from very slow to very fast speech the phone deletion rate rises from 9.3% to 13.6%; the phone substitution rate also changes significantly ($p < 0.05$), rising from 16.9% to 24.2%. As speaking rate increases, the entropy of the distribution of phone pronunciations also increases (Figure 3.2). A further examination of the data partially explains the entropy increase: as speaking rate increases, phones are not just pronounced as a single alternate phone instead of the canonical. Rather, phones are realized in a greater variety of ways in fast speech. For the slowest speech, the average phone had 3.4 different corresponding realizations occurring at least 2% of the time, whereas in fast speech, phones had an average of 4.0 different realizations. When counting only variations that accounted for at least 10% of the pronunciation instances of a phone, phones occurring in fast speech still had more distinct realizations (1.9) than slow speech (1.5).

From Figure 3.2 one can also see that word frequency, rate, and phone pronunciation interact: when the 100 most common words[11] are separated out from the general population, both frequent and infrequent words have similar entropy curves as speaking rate changes. Frequent words as a whole demonstrate wider variation in phone pronunciation than words containing semantic content. One might hypothesize that the high frequency of these words contributes to this disparity. It is also possible that, since function words (and therefore frequent words) are shorter in general, the average pronunciation deviation per phone is higher, but per-word deviation is similar to that of infrequent words. I will revisit

[11] Function words, which are usually short and syntactically necessary words such as *to, and,* or *the*, are almost always among the most frequent words. Yet the assumption that all frequent words are function words is not completely correct; for instance, the verb *know* is very frequent in the Switchboard corpus. However, these two classes do overlap to a significant degree.

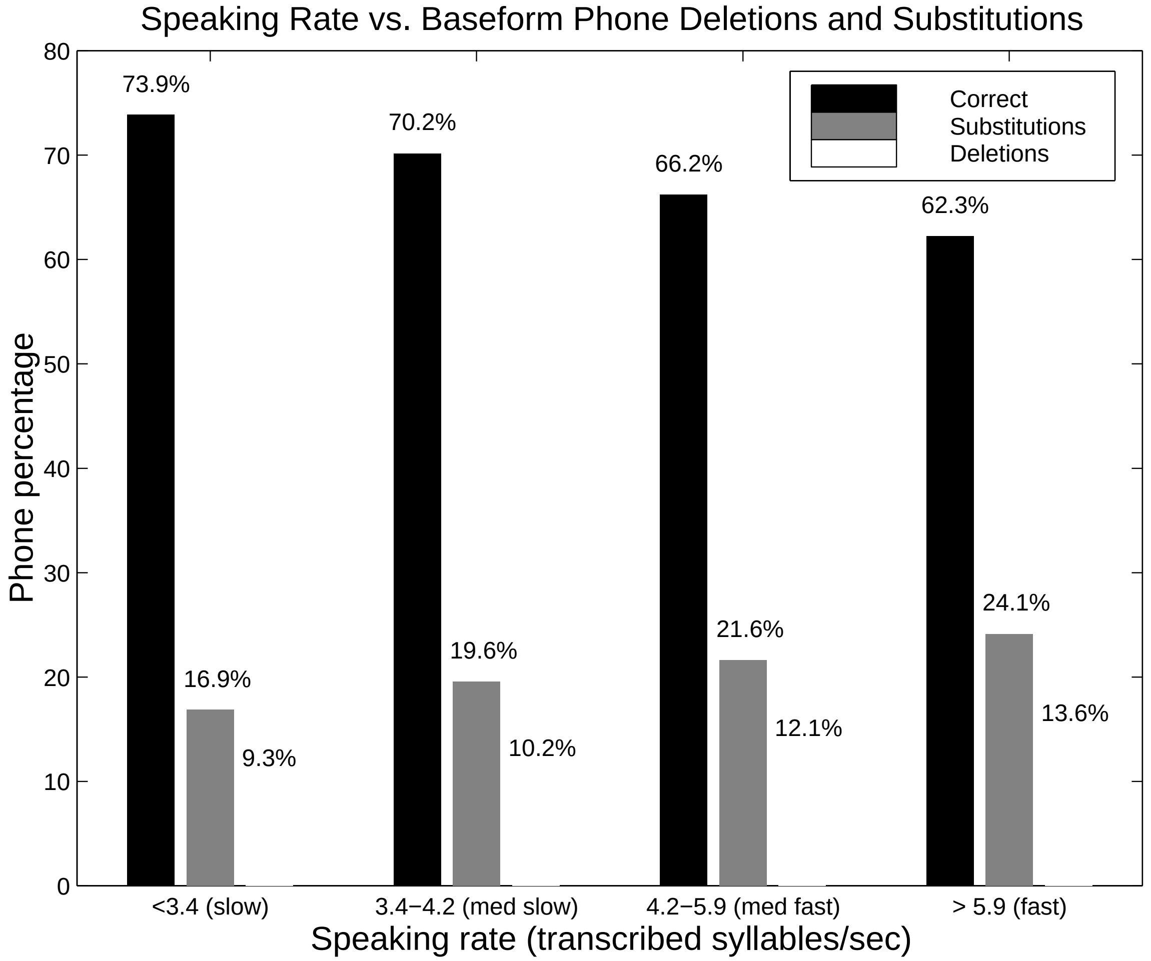


Figure 3.1: Phone-level statistics for effects of speaking rate on pronunciations.

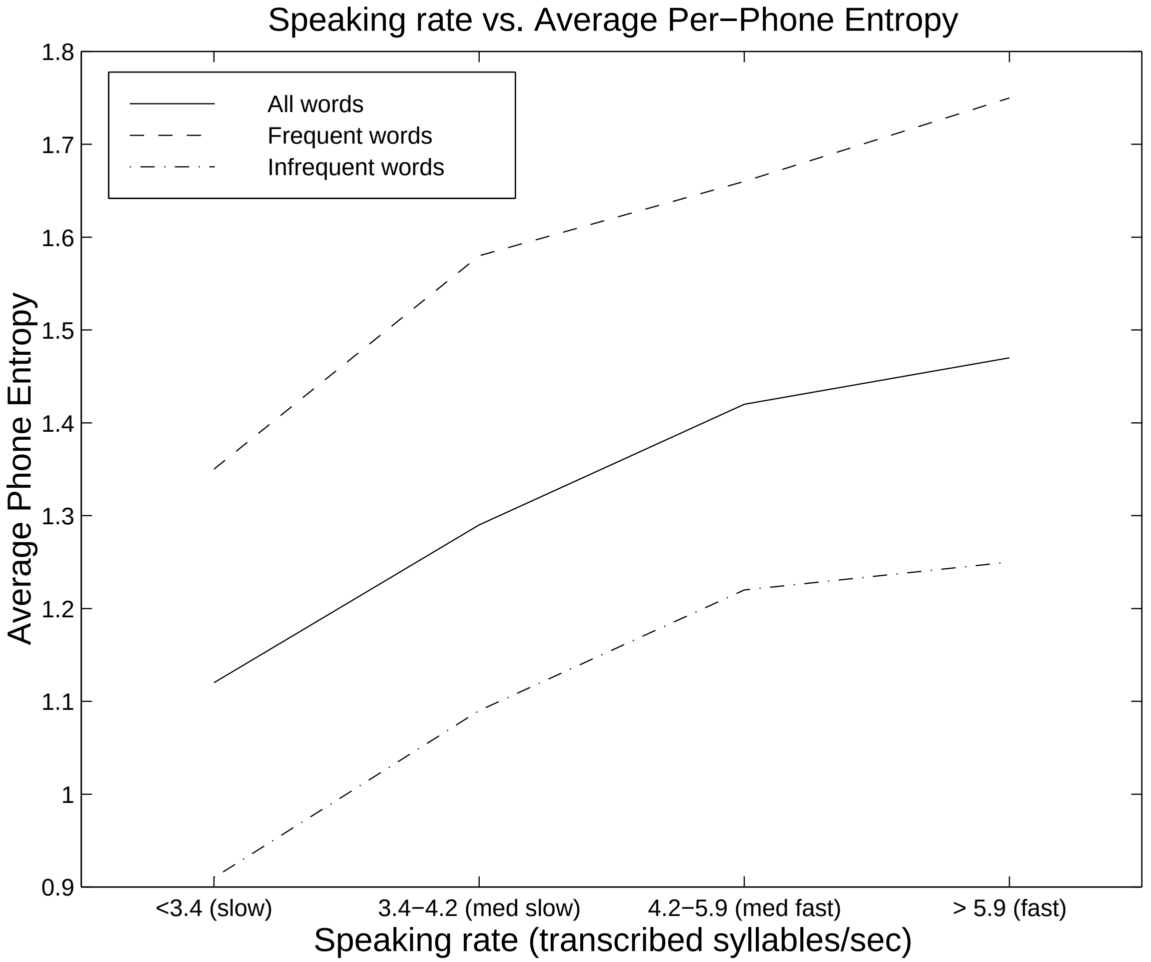


Figure 3.2: Phone-level entropy for different speaking rates. The 100 most common words in the corpus were labeled as frequent words.

this hypothesis in the next section.

### 3.2.3 Pronunciation statistics for syllables

In the final analysis of the Switchboard hand transcriptions, pronunciation statistics were examined both for groups of syllables and individual syllable types in the entire four-hour transcription set. This allowed me to cluster some of the data from the word-level experiments and to evaluate the effects on individual phones within their syllabic contexts. It has been suggested that pronunciation phenomena are more often affected by contextual factors that occur within the syllable rather than across syllable boundaries [Greenberg, 1998].

#### Statistics for groups of syllables

I computed the average syllabic distance (Equation 3.5 in Section 3.1.4) for all of the syllables in the set and plotted them against the speaking rate and the unigram frequency of the word. As can be seen from Figure 3.3, there is a non-linear relation among unigram probability, speaking rate, and the average distance for each syllable from the Pronlex baseforms: in less frequent words there is some increase in mean distance as speaking rate increases, but for syllable instances occurring in more frequent words the rate effect is more marked. This complex interdependency among these three variables makes sense from an information-theoretic viewpoint — since high-frequency words are more predictable, more variation is allowed in their production at various speaking rates, since the listener will be able to reconstruct what was said from context and few acoustic cues.

Above, I posed a question about the relationship between phone entropy and frequent words: is the increase in average phone entropy due to the shorter length of frequent words? Put another way, is the number of phonetic differences per syllable the same in frequent words as in other words, so that the higher phone entropy is just the result of frequent words having fewer phones on average? As the syllabic distance is not normalized with respect to the number of phones, one can see in Figure 3.3 that the length of frequent words is not a factor; if it were, then one would expect lower syllabic distance scores for syllables with a high unigram word probability.

When I replaced the probability of canonical pronunciations metric with the syllable distance metric for these same data, this general interaction among the metric, unigram frequency, and speaking rate was not observed. The key to understanding this initially puzzling result was to observe that the probability of canonical pronunciations *did* change as a function of rate when one took lexical stress (as marked in Pronlex) and syllabic structure into account. Syllables were annotated with `O` for onset consonants, `N` for nuclei, and `C` for codas, repeating symbols for clusters. Thus, `OONC` represents a syllable with a 2-phone onset cluster, a nucleus, and a single-phone coda (*e.g., step*). For each syllable type and stress type (primary, secondary, and none), I calculated the probability that syllables of that type were pronounced canonically, as a function of rate (Figure 3.4).

Most of the function (frequent) words were marked in the dictionary with sec-

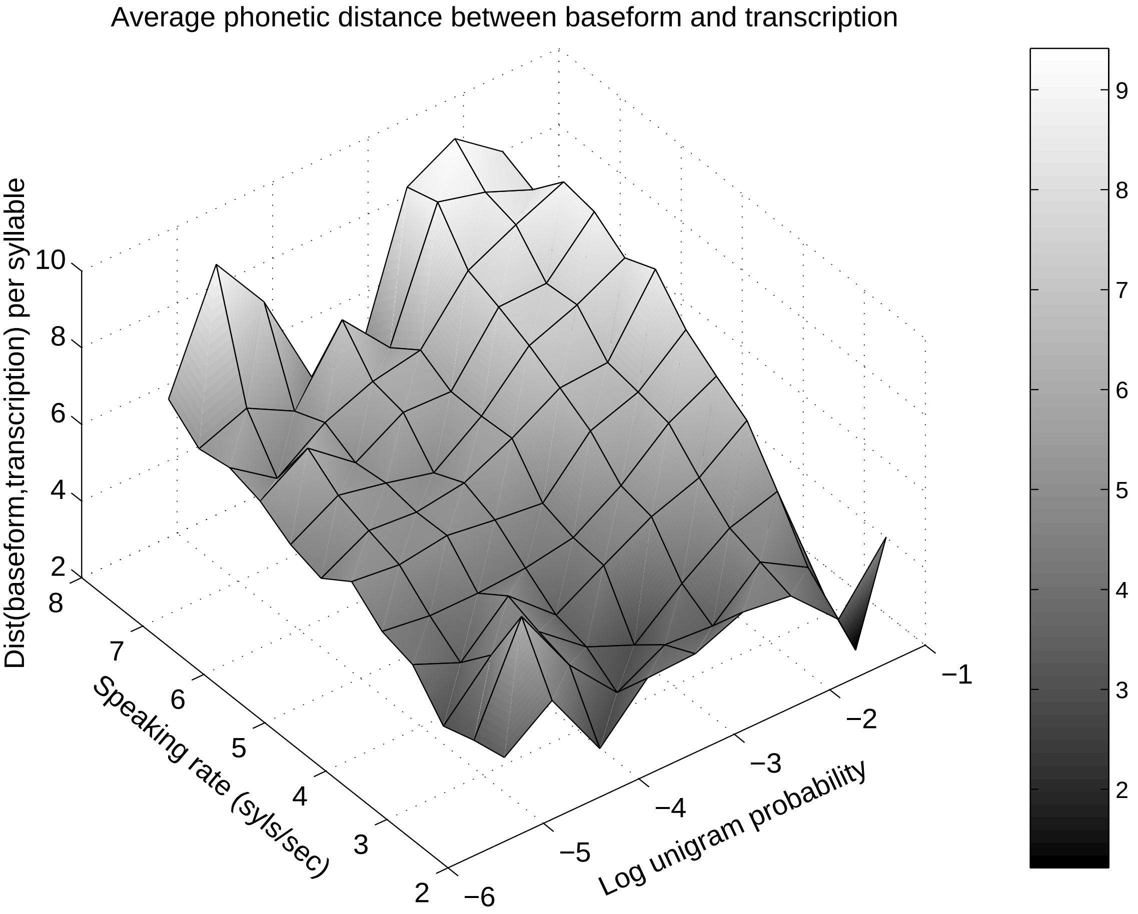


Figure 3.3: Average syllable distance from baseform as a function of speaking rate and unigram probability. For low unigram probability and very high or low speaking rates, the number of samples is low; the peaks in the graph for these extremes are unreliable due to statistical noise.

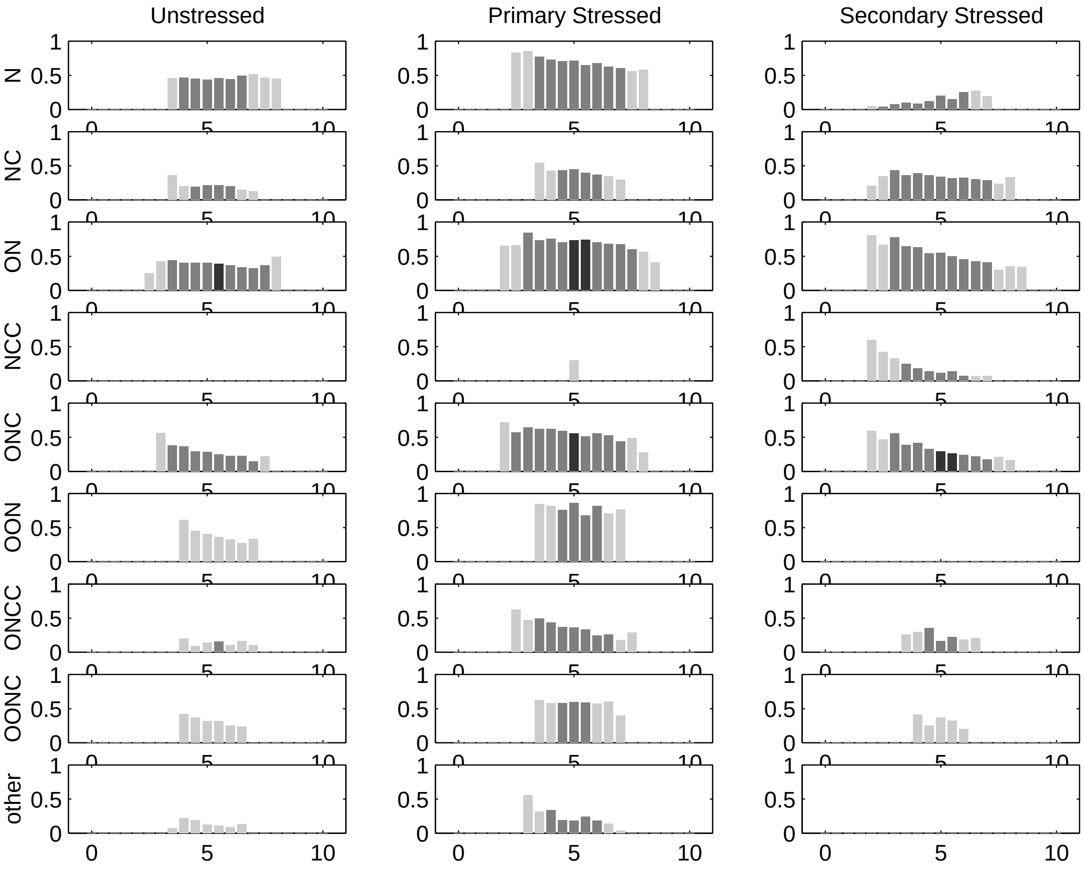


Figure 3.4: Probability of canonical pronunciations (y-axis of each subgraph) for different speaking rates (in syllables/second on x-axis), partitioned by lexical stress and syllabic structure. Light grey bars have 20-100 samples, medium grey 100-1000, dark grey >1000. O=onset, N=nucleus, C=coda; multiple occurences of O or C indicates consonant clusters. Empty boxes indicate syllable structures for which no histogram bin had 20 or more examples.

ondary stress rather than primary stress; therefore, the secondary stress category resembles a function word category somewhat in this analysis. This is supported by the fact that the average number of variants per secondary-stressed syllable in this database is 5.3, versus 1.8 alternatives for primary-stressed syllables and 3.2 for unstressed syllables. Looking across columns for each syllable type, these data also confirm that syllabic stress is an important factor in pronunciation models, as has also been observed by other researchers [Finke and Waibel, 1997a; Ostendorf *et al.*, 1997; Weintraub *et al.*, 1997]. It also seems that syllables without codas tend to be pronounced canonically more often than do syllables with codas, as can be seen by (for example) comparing `OONC` and `ONC` to `OON`. This accords well with the fact that coda consonants are more frequently pronounced non-canonically (usually because of phone deletion) than are onsets in this database, as reported by both Keating [1997] and Greenberg [1998].

The implication of these findings is that words may be identified most strongly by the syllable-initial portion of the word. Less variation is observed in onsets because they are used to discriminate between lexical items. Given the words in the transcribed portion of the Switchboard corpus, I located pairs of words that differed by one phone in the Pronlex dictionary (*e.g.*, *news* and *lose*) [Fosler-Lussier *et al.*, 1999]. These pairs were classified by whether the phone difference was in onset, nucleus, or coda position. Onset discrepancies outnumbered nucleus discrepancies by a factor of 1.5 to 1, and coda discrepancies by 1.8 to 1, indicating that at least for this crude measure, onsets appear to be more important for word discriminability.

For some syllable types, (*e.g.*, primary-stressed nucleus-only), rate had a strong effect on whether the syllable was pronounced canonically, but for others the effect is negligible. For one case (secondary-stressed nucleus-only),[12] a surprising reverse effect occurs — the probability of canonical pronunciation increases as rate increases. Thus, stress and syllabic structure do interact with speaking rate in terms of syllable pronunciations.

### Individual syllables

I then examined the 200 most frequent syllables in the Switchboard corpus; this is the equivalent of 77% syllable coverage of the four-hour transcription set, and 75% coverage of the corpus at large. For each syllable, the data were clustered into speaking rate histogram bins. The probability of the canonical and most likely pronunciations per syllable[13] were determined for each syllable as a function of the rate bin. I also reclustered the data in a similar fashion using trigram probability as the clustering criterion.

[12] There are only two words in the dictionary that fall into this category: *a* `[ey]` and *uh* `[ah]`.

[13] The canonical and most likely pronunciations differed for 55 of the 200 syllables; for example, *don't* (`[d ow n t]`) was most frequently transcribed as `[dx ow]` (*i.e.*, with a dental flap and deletion of the coda consonants).

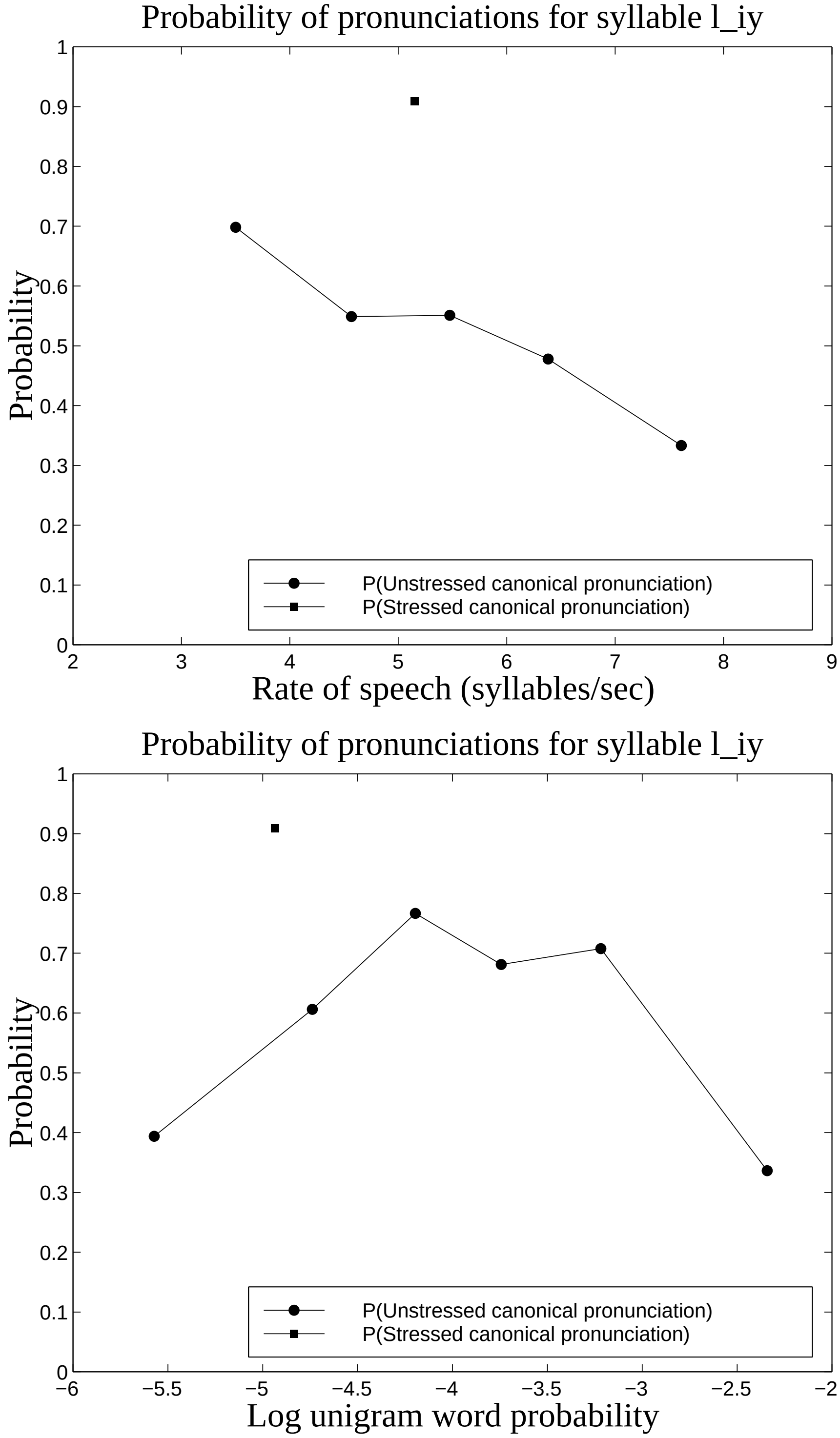


Figure 3.5: Probabilities of the canonical pronunciation for the syllable [l_iy] dependent on rate and unigram frequency.

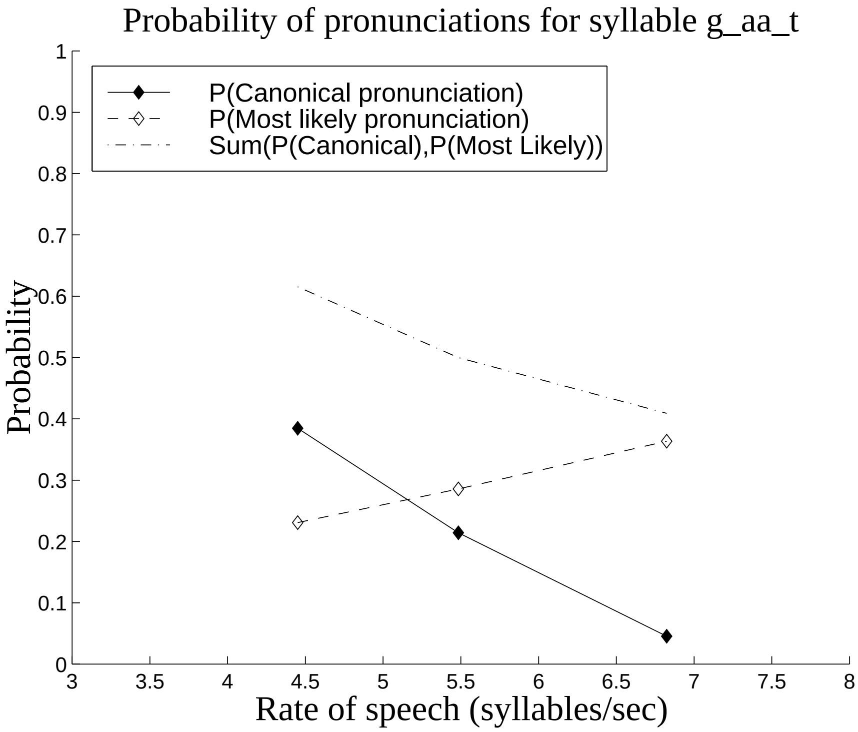


Figure 3.6: Canonical [g_aa_t] vs. most-likely [g_aa] for varying rates.

For every histogram bin, I determined the percentage of instances of the syllable that were pronounced according to either the canonical or most likely pronunciation. Figure 3.5 illustrates how the pronunciation probabilities change as a function of rate and word frequency for the syllable `[l_iy]`. There is a significant movement toward alternate pronunciations at faster speaking rates. Unigram frequency of the containing word also has a distinct effect on pronunciations of this syllable — they are least often canonical in the extremes of this metric. I am uncertain why this relationship is non-monotonic in this instance; while this is a rare occurrence, there are several other syllables in the Switchboard database that exhibit this behavior.[14] On a side note, one sees in these graphs the influence of stress on canonality: the stressed versions of `[l_iy]` appear unadulterated much more frequently than the unstressed versions.

For each of the 200 syllables, I ascertained whether the probability of the canonical or most likely pronunciation changed significantly between any two histogram bins as speaking rate or trigram probability varied. Table 3.3 shows that changes in the rate of speech significantly affected the probability of a syllable's canonical pronunciation in 85 of the syllables; when the most likely pronunciation is considered as well, the number of affected syllables increases to 95. The trigram probability of the word induces canonical pronunciations in fewer syllables, although roughly one-third of syllables are still affected.

High grammar probabilities is the major characteristic that describes the class of syllables that display significant differences in pronunciation with increased speaking rate. The mean unigram log (base 10) word probability for these syllables is -2.33; for unaffected syllables the mean unigram log probability is -3.03. The syllables that experience greater pronunciation differences as a function of rate are generally part of the more frequent words. This is consistent with the earlier syllable distance results in Figure 3.3, which showed a more marked effect of speaking rate on the pronunciation of syllables appearing in words with a high unigram frequency.

For some syllables (Figure 3.6), there is a tradeoff between the most likely and canonical pronunciations as a function of rate: that is, canonical probability will decrease for increased rate, whereas the most likely pronunciation will increase in probability. However, this tradeoff is not completely one-for-one: the sum of the canonical probability and most likely probability is lower for faster examples than for slower examples. In faster speech, other alternate pronunciations receive more of the probability mass.

Up to this point, unigram and trigram scores seem to be roughly equivalent in predictive power. For the vast majority of cases, it appears that using trigram scores provides little extra modeling power, since the trigram probability is often correlated with the unigram probability and the trends in pronunciations often match for the two features. Nevertheless, for a small number of frequent syllables it distinctly helps to have the trigram score. For example, in Figure 3.7, the syllable `[ih_f]`, which corresponds only to the word *if* in the training set (*i.e.,* all examples share the same unigram probability), is significantly reduced in very likely word sequences. In this case, the trigram score supplies extra information for forecasting reductions that the unigram does not provide, since the unigram is constant for all instances of *if.* Further evidence that the trigram probability is an effective

[14] Examples of this include **`[k_ah_z]`** and **`[m_ay]`**.

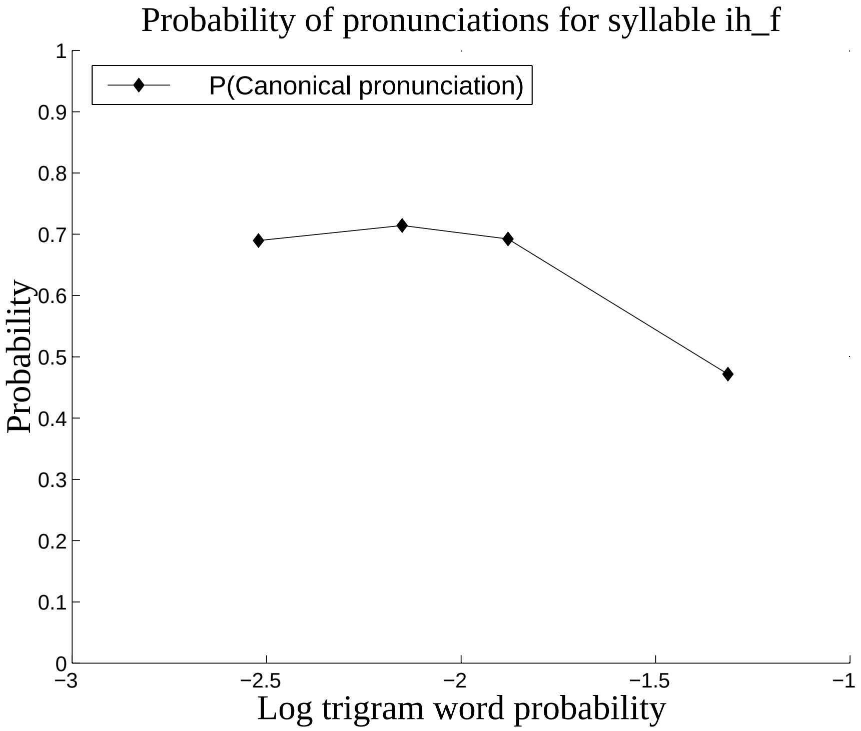


Figure 3.7: Probability of canonical `[ih_f]` pronunciation for various trigram probabilities.

| | # of syls w/ significant differences | | | |
|---|---|---|---|---|
| Clustering on: | Canon. | Most likely | Either | Both |
| Speaking rate | 85 | 81 | 95 | 71 |
| Trigram prob | 64 | 59 | 70 | 53 |

Table 3.3: Number of syllables (out of 200) with significant ($p$ <0.05) differences in pronunciation probabilities between histogram bins. The data are taken from the pronunciation histograms with varying speaking rate and trigram probability (*e.g.*, Figures 3.5 through 3.7).

tool for predicting reductions in high frequency words is presented by Jurafsky *et al.* [1998]; using regression models, they found that trigram probabilities were a significant factor in prediction of word length for six of the ten most frequent words. Trigram probabilities also helped in predicting differences in vowel quality (*i.e.,* whether the vowel was canonical, another full vowel, or reduced) for six of these ten words.[15] Therefore, this component, already present as the language model in many ASR systems, may be useful for predicting pronunciation change in frequent words.

### 3.2.4 The issue of "null" pronunciations

Occasionally, the phonetic transcribers in the Switchboard Transcription Project found that both they and the original word-level transcribers would hear words on the phrase level (*i.e.*, when the whole utterance was replayed), but when attempting a phonetic transcription, they could find no phones or syllables that could be associated with these words. I call words that show this effect *null pronunciation words.* This does not mean that there is no phonetic evidence for the word at all; rather, the phonetic features of the word are usually spread onto surrounding words, and timing cues often remain.[16]

Examples of this phenomenon from the corpus, initially described in [Greenberg, 1997a] are shown in Figure 3.8; the examples fall into six different classes based on part of speech. Words in parentheses are words transcribed by either the court reporter or the linguistic transcriber for which the linguistic transcriber could find no phones/syllables corresponding to the word.

These words are not phonetically realized, yet are heard on the phrase level. As discussed in Section 2.5.1, the predictability of the word, especially in terms of the grammar, plays an important part in determining the phonetic realization of words. Here it appears that the syntax and semantics embedded in the listeners' linguistic capabilities are "repairing" an acoustic gap left by low information-bearing words. All of the words that exhibit this behavior share several properties:

**Phonologically "short":** They are all one syllable long, which suggests that this process might best be regarded as extreme phonological reduction rather than deletion.

**Closed class restriction:** All of the null pronunciation words are members of closed syntactic classes[17] (preposition, complementizer, pronominal subjects, determiners, auxiliary verbs, conjunctions). Words in these classes are generally phonologically short, and occur in relatively predictable contexts.

[15] Only one of the ten words was unaffected in either the length or vowel quality categories.

[16] Several people have debated the authenticity of these non-segmental words; the reader is invited to see a demonstration of this phenomenon at `http://www.icsi.berkeley.edu/~dpwe/research/etc/phnless.html`. Regardless of the outcome of the debate, in these cases there is severe phonological reduction, so learning how to predict them will be advantageous to ASR modeling of these pronunciations.

[17] A *closed* syntactic class is a functional category that does not allow new words to be easily introduced into the language, as opposed to *open* syntactic classes, like nouns and verbs, in which new words can be generated.

1. more (of) that
2. decided to [pause] carve up that part (of) the world and call part of it [pause] persia
3. nice talking (to) you also
4. try (to) tell my kids
5. so we gave him to um (i) don't know if you've
6. since [breath] (you) know my parents had to force me to get my driver's license when i was young
7. are you (a) vietnam veteran dudley
8. you know we('d) all go camping my dad (and) my mom and and the kids

Figure 3.8: Examples of null pronunciation words (given in parentheses) in the Switchboard corpus

**Syntactic requirement:** The null pronunciation word is often syntactically required, although it may itself have little semantic meaning. Readers who are reading acoustic transcriptions usually find the sentences without the null pronunciation word ungrammatical; replacing the word renders the sentence grammatical.

**Likelihood requirement:** Null pronunciation words must be predictable within the syntactic context in which they occur.

**Timing and contextual cues:** non-segmental words are often accompanied by lengthening of surrounding syllables, which can act as a time filler [Greenberg, 1996; Cardinal *et al.*, 1997]. Furthermore, while there may be little to no segmental evidence for the word itself, it may leave traces by modifying the acoustic properties within surrounding segments.

If null pronunciations are extreme cases of phonological reduction, then it is likely that all of these factors would be useful for predicting when reductions can occur, which can lead in turn to better prediction for pronunciation models in ASR.

In particular, I would like to focus briefly on the likelihood requirement mentioned above. What does it mean to be syntactically likely? In automatic speech recognition systems, a *language model* is used to predict the likelihood of a sequence of words. Usually this model is simplified so that the probability of a word is conditioned on the previous one or two words (in general, called an *$n$-gram grammar*). Is an $n$-gram grammar enough to predict this type of variation? I will discuss some of the phenomena in terms of predictability by $n$-gram grammars, since this is the easiest information to incorporate into an ASR pronunciation model.

| | more | some | rest | part | couple |
|---|---|---|---|---|---|
| of | 54% | 98% | 94% | 99% | 100% |
| than | 21 | | | | |
| in | 7 | <2 | 3 | | |
| for | 6 | <2 | 3 | | |
| *others* | 12 | <2 | | <2 | |

Figure 3.9: Probability of prepositions following the nouns *more, some, rest, part,* and *couple.*

### more/rest/some/part/couple (of)

One of the simplest cases of syntactic predictability is the null pronunciation of the word *of.* In the chart displayed in Figure 3.9, we see the probability of a preposition in the syntactic sequence [`Noun Preposition NounPhrase`] for the nouns *more, some, rest, part,* and *couple*, taken from 124,759 acoustically segmented utterances within the Switchboard corpus. The probability of the word *of* following the nouns listed above is much higher than the probability of any other preposition, which is easily representable by $n$-gram statistics.

| | talk |
|---|---|
| about | 43% |
| to | 32 |
| with | 3 |
| *others* | 22 |

Figure 3.10: Probability of words following *talk.*

### talk (to) you

In looking at another preposition, *to*, we would have expected similarly that *to* would be the most likely co-occurring preposition with *talk.* However, this is not the case, as seen in the chart of P(*prep*|talk) (Figure 3.10).

| P⇓ NP⇒ | you | me | him | her | them | us | people | that | this |
|---|---|---|---|---|---|---|---|---|---|
| to | 32 | 1 | 4 | 4 | 4 | 2 | 2 | 1 | |
| about | 4 | | 2 | | 1 | | 3 | 14 | 9 |

Figure 3.11: Comparison of words following *talk to* and *talk about.*

How can the model predict that the null pronunciation word should be *to* rather than *about*? One could appeal to the "phonologically short" criterion — since the probabilities of *about* and *to* are very close, the deciding factor could be that *about* is (phonologically) a much larger word, having two syllables. Another possibility is to look at more than just subcategorization of the preposition, and take into account some of the right context of the word. For the *talk (to) you* example, I found all of the noun phrases that followed *talk to* and *talk about* in the corpus. Figure 3.11 shows the number of instances of *to* and *about* co-occurring with various noun phrases.[18]

When the referent of the following noun phrase is a person, the preposition is much more likely to be *to*. This suggests that an $n$-gram grammar may not be enough to capture this generalization — one may have to appeal to lexically based grammars, where the preposition choice can be determined by valence probabilities.[19] However, it seems that $n$-grams would not be a bad approximation for predicting preposition identity.

### Pronominal subjects

1. (i) think it was called credit union news
2. so we gave him to um (i) don't know if you've
3. since [breath] (you) know my parents had to force me to get my driver's license when i was young

Figure 3.12: Null pronunciation pronominal subjects

Some examples of null pronunciation pronominal subjects are given in Figure 3.12. In this instance, the words preceding the non-segmental word do not necessarily predict the pronominal subject well, but the main verb that follows does. While this is very strange from an $n$-gram grammar point of view, in lexically-based syntactic frameworks there is a similarity with the preposition cases above; we are just looking at the co-occurrence of a main verb with one of the items it can subcategorize for.[20]

In examining the data in Figure 3.12, one of the questions that comes to mind is why the filled-in subject of *know* is *you*, while the other filled-in subjects are *I*. Looking at the syntactic patterning of subjects of *know, don't know, think,* and *don't think* in the corpus, however, reveals the answer (Figure 3.13).[21] The data show that *you* is the most likely subject of *know,* while *I* is the most likely subject of the other three. One analysis

[18] Since the numbers are quite small, I do not present percentages here.

[19] Essentially, the valence properties of verbs are requirements that verbs have about their objects. In this case, the verb *talk* could specify that, with a following prepositional phrase containing a "person" noun phrase, the most likely preposition would be *to*.

[20] In Head-driven Phrase Structure Grammar, for instance, the subject of the sentence is a special object of the verb, subject to the subcategorization requirements of the verb.

[21] Again, I present the exact numbers of instances so that one can see the relative frequencies of the verbs in question.

| Verb⇒ | think | | know | |
|---|---|---|---|---|
| Subj⇓ | (do) | don't | (do) | don't |
| I | 2853 | 321 | 897 | 1034 |
| you | 279 | 10 | 7535 | 21 |
| we | 21 | 0 | 27 | 9 |
| they | 21 | 2 | 35 | 13 |

Figure 3.13: Counts of verb subjects for the verbs *think* and *know* in both positive and negative forms.

of these data is that *I* is the (probabilistic) default subject for *don't,* which explains why *don't know* favors *I* over *you.* One caveat: *you know* may be used as a filler phrase, so the syntactic patterns may not be the same as with other verbs. It may be the case that the filler *you know* is just one lexical item.

In essence, $n$-gram grammars are not going to be of use in the prediction of this reduction, but perhaps one can utilize inverse $n$-gram grammars, where previous words are predicted from following words. Of course, this is only possible where one has the entire hypothesis word string (or at least $n$ words ahead of the current word), so in a recognizer, this is only implementable as a rescoring pass over lattices or $n$-best hypothesis lists. Another possibility exists: since the effects described here are local and occur in frequent contexts, adding the identities of frequently occurring adjacent words may capture some of the alternate pronunciations described here.

## 3.3 Summary

Analysis of phonetic transcriptions of the Switchboard corpus has demonstrated significant effects of speaking rate and two measures of word predictability on pronunciations of words, syllables, and phones. One of the most significant findings is that not every linguistic unit is affected by changes in these factors. This suggests that modeling individual words and syllables may improve incorporation of these factors into pronunciation models. An increase in transcribed syllable rate is correlated with deviation from dictionary baseforms in roughly half of the syllables and a little less than a third of the examined words. High word predictability (both using the unigram and trigram metrics) also tends to accompany lower canonical pronunciation probabilities, although for some syllables lower canonical pronunciation probabilities can be found in words with low grammatical probabilities.

There is a significant interaction between the investigated features and pronunciations. In particular, Figure 3.3 shows that word frequency has a distinct influence on how much pronunciation variation is present with changes in transcribed syllable rate: syllables in high-frequency words are most affected by the rate of speech. Stress and syllable structure also play an important part in cooperation with these features; Figure 3.4 illustrates that variations due to rate are more visible when these factors are included.

In some situations, there can be severe phonological reduction in the Switchboard corpus when the information content of a word is very low. $N$-gram statistics may help to predict some of these instances; in other cases, it appears that different measures, such as the identity of the neighboring words, will be necessary to predict environments in which reductions can occur.

A short epilogue: the database produced for the investigations in this chapter has proved useful in further studies of pronunciation phenomena within Switchboard, performed primarily by colleagues at the University of Colorado, Boulder. Jurafsky *et al.* [1998] studied the ten most frequent words in the corpus, finding that planning problems, predictability, segmental context, and rate of speech are good independent factors for predicting when reductions can occur. This study was extended in [Bell *et al.*, 1999], where the effects of word position in the conversation turn and the speaker-specific variables of gender and age were included in the analysis. The speaker-specific variables correlated mostly with speaking rate and dysfluencies — only a small independent effect on pronunciations was attributable to gender after these latter features were taken into account. On the other hand, the pronunciation of function words was affected by the turn and utterance boundaries. Reduction was less likely at the start of turns or the end of utterances. Gregory *et al.* [1999] showed that phonological effects for word-final `[t]` and `[d]`, including flapping, deletion, and length increases, were correlated with various measures of $n$-gram frequency and contextual probability. These results suggest that many features exist that could be useful for predicting pronunciation variation in ASR models.

# Chapter 4

# Pronunciation Error Analyses in ASR Systems

## 4.1 Switchboard recognizer error analyses

It is clear from the analysis in the previous chapter that speaking rate and word predictability are both distinctly correlated with pronunciation change. In this chapter, I investigate the effects of mismatches between the ASR pronunciation model and the Switchboard hand transcriptions on ASR word error rates, as well as how these mismatches correlate with speaking rate and word predictability.

For these analyses of recognition performance on the Switchboard corpus, I used the HTK recognizer trained with the Pronlex dictionary developed at the 1996 Johns Hopkins Workshop (hereafter referred to as the WS96 recognizer) to provide recognition hypotheses for error analysis. This Hidden Markov Model (HMM) recognizer is a 12-mixture state-clustered cross-word triphone system, trained on 60 hours of mel cepstrum (MFCC) features (including first and second derivatives of the features). The recognizer used a bigram language model trained on 2.1 million words of Switchboard transcripts.

| | Overall | Canonical pronunciations | Alternative pronunciations |
|---|---|---|---|
| % correct | 57.4 | 65.0 | 53.9 |
| % deleted | 12.0 | 8.1 | 13.9 |
| % substituted | 30.5 | 26.1 | 32.2 |
| # of words | 4085 | 1337 | 2748 |

Table 4.1: Breakdown of word substitutions and deletions with WS96 Switchboard Recognizer for canonical and alternative pronunciations.

### 4.1.1 Mismatch in pronunciations as a correlate of word error

Previous studies [Weintraub *et al.*, 1997] have shown that when the hand transcriptions of the Switchboard corpus were compared to the Pronlex dictionary, only two-thirds of the dictionary phones matched the transcriptions. In an elaboration of this study, I have tried to characterize the effects of these phone-level statistics on word-level pronunciations. While 67% of the phones retained canonical form in spontaneous speech, only 33% of the word pronunciations found in the Switchboard development test set (using ICSI hand transcriptions) were found in the Pronlex dictionary.[1] Thus, the observed phone transformations are not concentrated in a few words, but rather are spread throughout the corpus.

What remains to be shown is that these pronunciation errors have an effect on ASR systems. Intuitively, one would believe that recognizers would fail miserably if 67% of hand-transcribed word pronunciations are not in the dictionary. However, it is not necessarily true that ASR acoustic models are modeling the same linguistic ideals given by the hand transcriptions. They are biased by their training set: performance tends to be better on words that occur many times in the training corpus. Acoustic models may also compensate for pronunciation variation to some degree by smoothing out the phonetic classes, accepting variations within the canonical phonetic class estimates. It is important, therefore, to ascertain whether ASR systems perform worse in cases where there is a mismatch between the hand transcriptions and dictionary pronunciations.

For the WS96 system, I compared recognizer results in conditions where linguists determined that pronunciations were canonical against results in conditions where alternative pronunciations were used by the speaker. In this study, 439 phonetically transcribed sentences were examined from the Switchboard development test set. Each word in the test set transcriptions was annotated with whether it was correctly recognized, substituted, or deleted by the WS96 system, and whether the transcribers observed a canonical or alternative pronunciation, as defined by the Pronlex dictionary (*i.e.*, the recognizer lexicon). Recognizer insertions were disregarded. Although pronunciations certainly have an effect on insertions, it is difficult to mark them as canonical or alternative pronunciations compared to the hand transcriptions, since the speakers did not actually utter the inserted words.

The WS96 system recognizes words correctly much more often when the linguists' transcription matches the dictionary pronunciation (Table 4.1). There is a large (70% relative) increase in the recognizer word deletion rate for words with alternative pronuncia-

[1] For the data set examined, the average word had 3.1 phones.

tions, as well as a significant increase in recognizer substitutions. The fact that the overall recognizer accuracy for alternatively pronounced words is not much worse than that for commonly pronounced words, however, does indicate that there is some compensation for pronunciation variation by the acoustic model.

It is difficult to separate the effects of different factors on word error rates; for instance, a mispronounced word can result in a substitution, causing a language model error for the following word. Hence, some of the words labeled as having a canonical pronunciation may be identified incorrectly by the recognizer due to surrounding pronunciation errors; the extent of this phenomenon is difficult to characterize. Nevertheless, these numbers suggest that there is a real effect of non-canonical pronunciations on word error. The numbers also suggest that solving "the pronunciation problem" will not solve the speech recognition problem, but will contribute toward the reduction of error rates.

### 4.1.2 Relationships between factors and recognizer error

Although there is a relation between the pronunciation model and recognizer errors, it is not clear how recognizer errors relate to factors such as speaking rate, unigram probability, and trigram probability. I labeled every word in the development test set with the syllable rate, unigram probability, and trigram probability of the word. I then partitioned the words into histogram bins and determined the recognizer accuracy for each bin. The following series of graphs show how recognizer scores (y-axis) change as a function of each extra-segmental factor (x-axis). Included on each graph is the percentage of words that had canonical pronunciations (represented by the line with the crosses) and average recognizer accuracies for all words (solid line with dots), for words with canonical pronunciations (dashed lines), and for words having alternative pronunciations (dot-dashed lines) as a function of the extra-segmental factor.

In Figure 4.1a, one sees that there is a 14% (absolute) drop in recognizer accuracy as the speaking rate moves from very slow to very fast speech. This is due mainly to the poorer performance on words pronounced non-canonically, which are more common in fast-speech conditions, as seen in Figure 3.1. Note that for this test set the percentage of utterances in the fastest (>6 syllables/sec) bin is non-trivial, containing 35% of the data; thus, there is a real and significant effect from fast speech for this set. One additional note: as in Section 4.1.1, these graphs do not include insertions. Since rate is calculated over an interpausal region, insertion rates can be calculated for each speaking-rate bin. Insertions decrease from 7.7% to 2.3% as the speaking rate increases from the slowest to the fastest bin; when this decrease in insertion rate is taken into account in the word error rate, the difference in errors between slow and fast speech is still roughly 9% absolute.

In the case of language model probabilities (Figures 4.1b and 4.1c), recognizer performance improves as words become more likely. This is not surprising, since both language models and acoustic models in the recognizer tend to favor more likely words during recognition. The trigram graph has a larger spread (from 30% to 69%) than the unigram (31% to 61%), probably because the recognizer (which utilizes a bigram grammar) takes into account more contextual information than is provided by unigram probabilities. What is interesting here is that, even though the recognition rate increases as words become

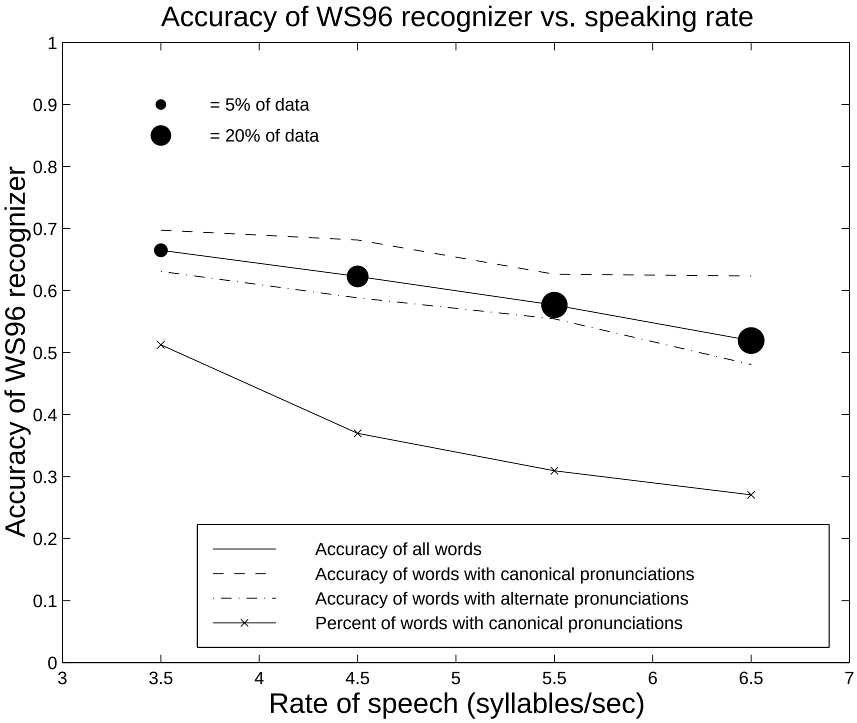


Figure 4.1a. Speaking rate

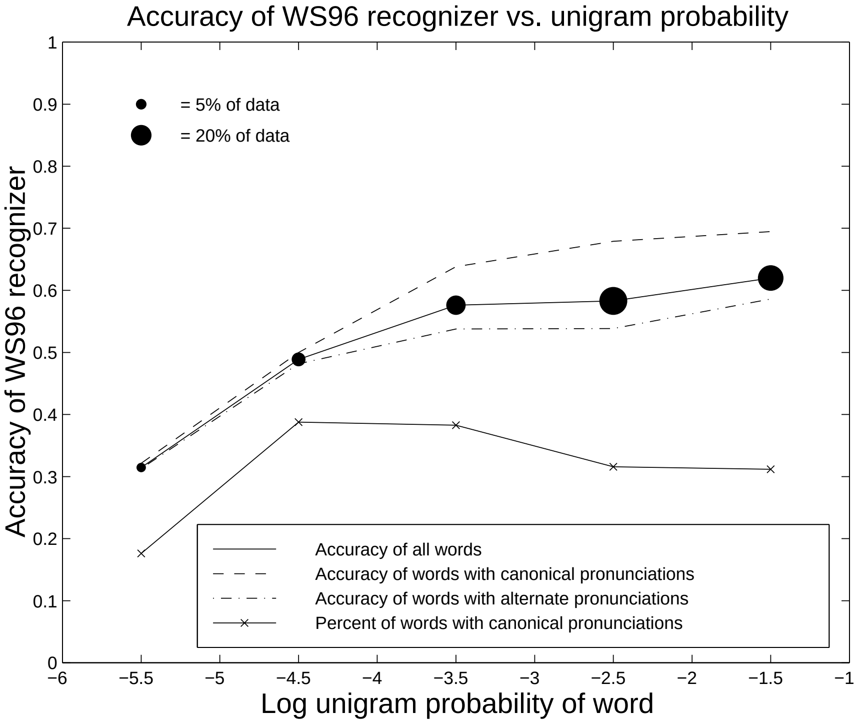


Figure 4.1b. Unigram probability

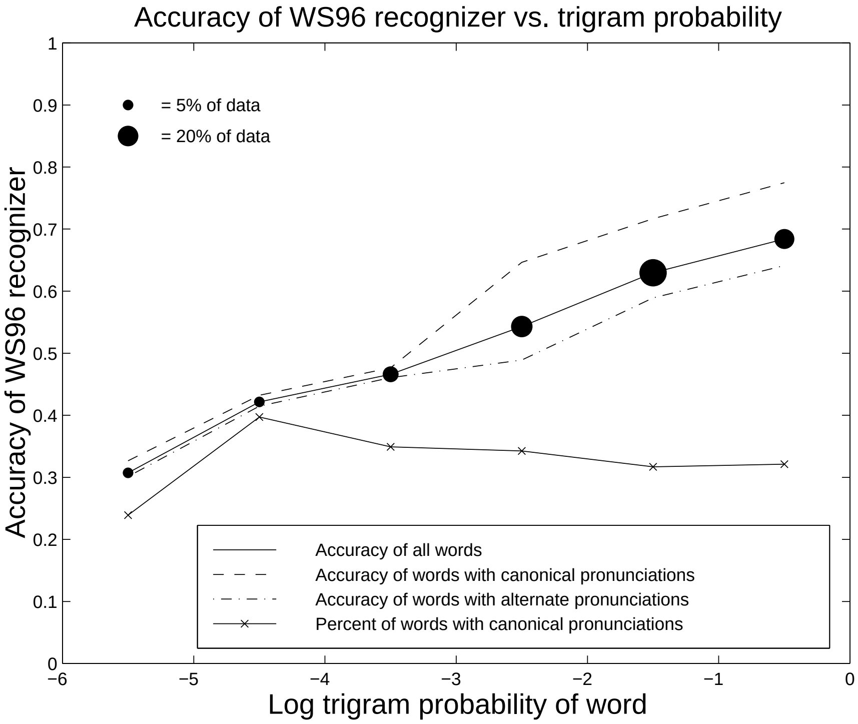


Figure 4.1c. Trigram probability

Figure 4.1: Accuracy of WS96 Switchboard recognizer dependent on several factors, represented as the fraction of words recognized correctly. In these graphs, the solid line indicates the overall accuracy trend as each factor changes. The size of the dots indicates the proportion of data found in that particular histogram bin. The dashed and dot-dashed lines indicate recognizer scores when the hand transcription of the word did or did not match the canoncial (recognizer) pronunciation, respectively. The solid line with crosses indicates the percentage of words that had canonical pronunciations for that histogram bin.

more likely, the percentage of words with canonical pronunciations decreases, as indicated by the line with crosses.[2] For higher probability words (*i.e.,* $\log_{10}$(trigram)>-3), canonically pronounced words are recognized much more accurately than non-canonically pronounced words. On the other hand, for low probability words the language model in the recognizer dominates the error, and it does not matter as much whether the pronunciation is canonical or not. The trend of increasing recognizer accuracy with increased log trigram probability is discordant with the decreasing number of pronunciations agreeing with the recognizer dictionary; this conflict provides some insights with respect to the behavior of the recognizer. In this system, either the language model is heavily favoring likely words, or the acoustic model has broadened to account for the increased variation in frequent words. Both of these hypotheses are probably true to some extent. A better model of pronunciations could allow the language model to discriminate against infrequent words less, as well as allowing for sharper acoustic models.

### 4.1.3 Summary

In Switchboard, non-canonical pronunciations pervade the landscape; only 33% of words are canonically pronounced. The question is: are the acoustic models of the WS96 recognizer accommodating the pronunciation variation observed in this corpus? The models are certainly not doing the job completely, since words with non-canonical pronunciations have an 11% absolute increase in word error over canonically pronounced words.

Furthermore, word error correlates with transcribed syllable rate. Faster speech goes hand-in-hand with increased errors, as has been observed in other corpora [Fisher, 1996b]; much of this error can be attributed to the increase in pronunciation variation at fast rates. In the realm of word predictability, more likely words are recognized with better accuracy, since they are better modeled by the acoustic and language models. Pronunciation variation affects the recognizer's performance only for highly probable words, whereas for unpredictable words, alternatively pronounced instances are recognized with an error rate similar to that of canonical baseforms.

## 4.2 Recognizer error analyses for automatic transcription of Broadcast News

While knowing the correspondence between alternative pronunciations in linguistic transcriptions and recognizer performance is important, many pronunciation modeling systems use automatic transcription methods to determine possible word pronunciations. Do the same patterns seen in the hand transcriptions of Switchboard carry over to an auto-

[2] It appears that the the probability of canonical pronunciations drops for low probability words because these words tend to be longer, so *a priori* there is an increased chance of a single phone changing in a word. In fact, this is confirmed when the probability of canonical pronunciation is calculated at the phone level. For low unigram probabilities (log unigram < -3), the probability of canonical phone pronunciation is roughly constant at 75%, whereas for log unigrams between -2 and -3, the canonical phone probability is 67% and 50% for log unigrams above -2. The class of infrequent words (log unigram < -5) makes up 5% of the words in the test set.

matic learning paradigm? Chapters 5 and 6 describe such a pronunciation modeling scheme in the Broadcast News (BN) domain. Using an early version of this system, I duplicated the Switchboard studies with the Broadcast News corpus using automatically determined transcriptions rather than hand alignments in the analysis. An added advantage to working with this corpus is the mixture of speaking styles: the effects of speaking rate and word predictability could be examined for both spontaneous and planned speech.

### 4.2.1 The corpus

The Broadcast News corpus [NIST, 1996] is a collection of speech from American radio and television news broadcasts, such as the National Public Radio program *All Things Considered* or *Nightline*, televised in the U.S. on the ABC network. These shows comprise a wide range of speaking conditions, from planned speech in studio environments to spontaneous speech in noisy field conditions over telephone lines. The (possibly multi-sentence) segments are divided into seven different focus conditions representing different acoustic/speaking environments;[3] in this study, I primarily investigated two main conditions that make up the majority of the data in the set: planned studio speech and spontaneous studio speech.

For this study, I used the results from the SPRACH hybrid neural network/HMM recognizer developed at Cambridge University, Sheffield University, and ICSI [Cook *et al.*, 1999]. This system combines a recurrent neural network trained on PLP features from Cambridge, a multi-layer perceptron trained on modulation-filtered spectrogram features [Kingsbury, 1998] from ICSI, and HMM decoder technology from Sheffield. This investigation used an intermediate version of the evaluation recognizer described by Cook *et al.* [1999]; this recognizer performed with roughly 20% more errors than the evaluation system. The acoustic models of the system described here were trained on 100 hours of Broadcast News speech. The lexicon of the recognizer used context-independent pronunciations from the Cambridge 1996 ABBOT recognizer [Cook *et al.*, 1997]; the trigram grammar was trained on 286 million words of text from transcriptions of broadcasts and newswire texts. The system was tested on a 173 segment subset of the 1997 Broadcast News DARPA evaluation test set, corresponding to roughly a half-hour of speech (also known as Hub4E-97-subset).

The detailed phone-level transcriptions that we had for Switchboard were not available for Broadcast News (BN). In order to find an approximation to the phonetic hand transcription, I used the SPRACH BN recognizer in a phone-constrained decoding. The recognition of the BN training set used monophone acoustic models in the phone recognizer; a phonotactic phone-bigram grammar[4] provided probabilities for each phone following other phones. For each utterance, the SPRACH BN recognizer generated an automatic

[3] The shows in the Broadcast News training set have been segmented by NIST; the first 100 hours of this data are labeled with the focus condition applicable for the segment. The test data, when first presented to sites participating in the DARPA evaluation, came unsegmented (sites were responsible for devising their own segmentation), but the scoring files containing the actual transcriptions do have segmentations marked. Throughout this thesis, I use the NIST provided answer segmentations for the test set; the influence of automatically segmenting the shows is therefore not included in word error results.

[4] The phone-bigram grammar was trained using the phone transcription from a Viterbi alignment of the training set to the BN recognizer dictionary.

phone transcription, which was subsequently aligned to the word transcription. As in the Switchboard analysis, test set words were annotated with whether the recognizer correctly identified them, substituted other words for them, or deleted them. In addition, the alignment was used to determine whether the word was pronounced canonically according to the recognizer's acoustic models.

While there is no guarantee that the phone transcription produced by the above procedure will match the decisions of human transcribers,[5] it does provide a clue to which acoustic models best match the phonetic content of the waveform. Since the job of a pronunciation model is to facilitate matching between the acoustic models and word hypotheses in a recognizer, and since several researchers use phone recognition as a source for pronunciation alternatives (as described in Chapter 2), it is appropriate to investigate the effects of the extra-segmental variables on the automatic phonetic alignment.

### 4.2.2 Mismatch in pronunciations as a correlate of error

Using the Broadcast News database, I examined 173 (possibly multi-sentence) segments from the 1997 evaluation test set, which provided roughly the same number of words as the Switchboard test set. The difference between recognition rates for canonical versus non-canonical pronunciations is more marked for Broadcast News (Table 4.2); this is not unexpected, since the same acoustic models used to recognize the speech also determined whether a pronunciation is canonical in the BN analysis, as opposed to the Switchboard analysis, which uses phonetic labelings provided by linguists to make this determination. Both systems see a large increase in deletion rates in alternatively pronounced words, but the increase in substitutions for these words is much greater for the SPRACH BN system — possibly due to the automatic phone transcription or to the larger overall error rate of the Switchboard system. It is also interesting to note that a similar proportion of words were judged to be pronounced canonically in each system, although the difference is significant ($p$ <0.0001) — 33% for Switchboard and 28% for Broadcast News.

### 4.2.3 Relationships between dynamic factors and recognizer error

The most notable difference between the language model graphs for Broadcast News (Figure 4.2) and those for Switchboard was the increasing percentage of words having canonical pronunciations as words became more frequent (cf. Figure 4.2b to Figure 4.1b). This is probably an effect of the acoustic models: the recognizer is likely better at recognizing words found frequently in the training set, so the automatic phonetic transcription reflects this bias. The curves showing the effects of trigram probability are rather flat (Figure 4.2c), particularly for canonically pronounced words, although alternative pronunciation scores increase for the highest frequency and drop off for the lowest — the latter shows the influence of the language model in recognition.

The graph for unigram probabilities appears strange at first glance (Figure 4.2b);

[5] I found via inspection of samples that the automatic phone recognizer usually produced intuitive transcriptions; however, conditions in which the acoustic models fare poorly, such as noisy speech, often degraded the phonetic transcript.

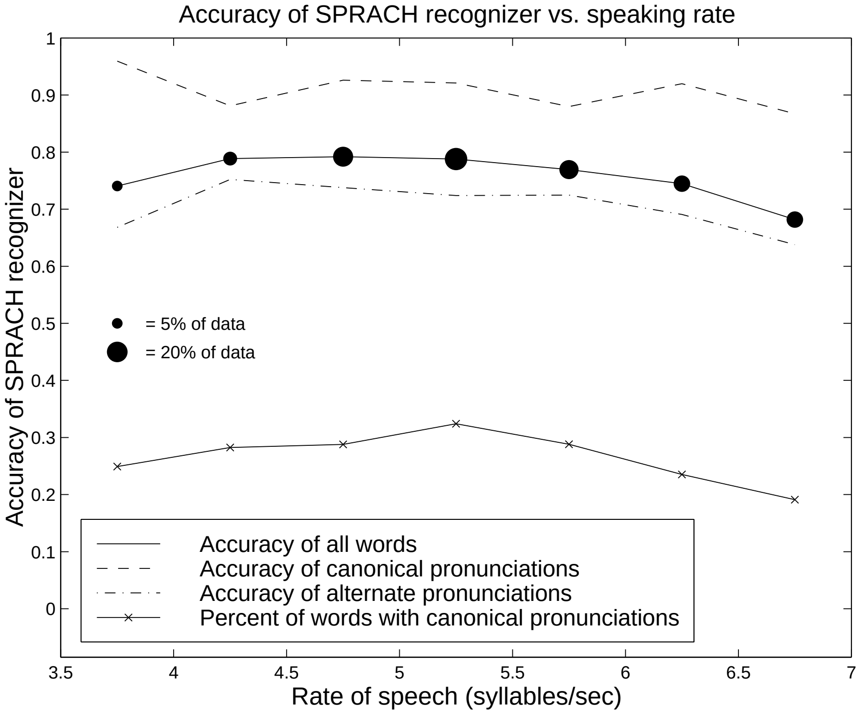


Figure 4.2a. Speaking rate

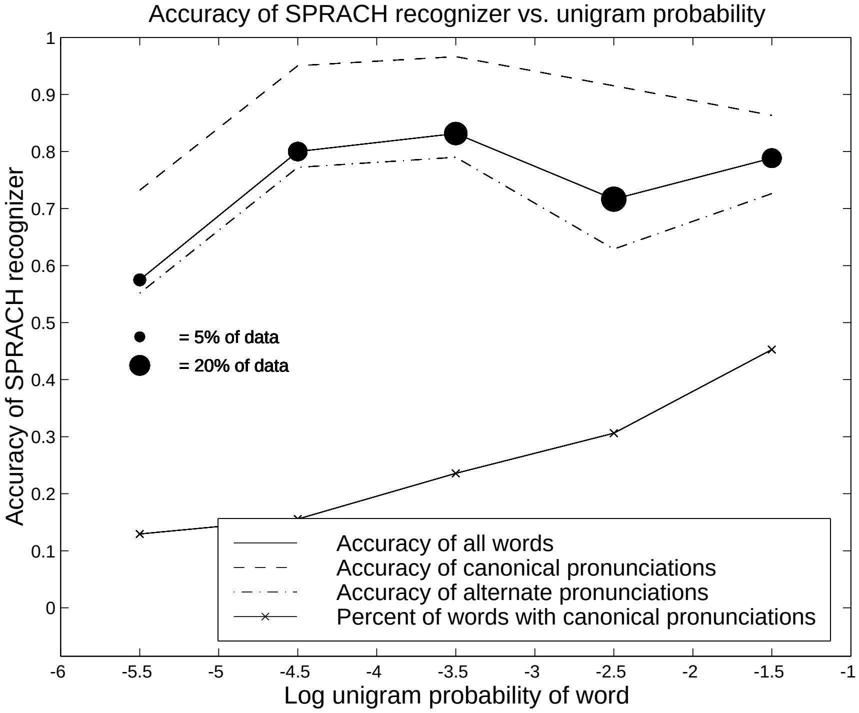


Figure 4.2b. Unigram probability

| | Overall | Canonical Pron. | Alternative Pron. |
|---|---|---|---|
| % correct | 76.4 | 90.8 | 70.9 |
| % deleted | 4.7 | 1.4 | 6.0 |
| % substituted | 18.9 | 7.8 | 23.1 |
| # of words | 5840 | 1607 | 4233 |

Table 4.2: Breakdown of word substitutions and deletions with Hybrid Broadcast News Recognizer for Canonical and Alternative Pronunciations.

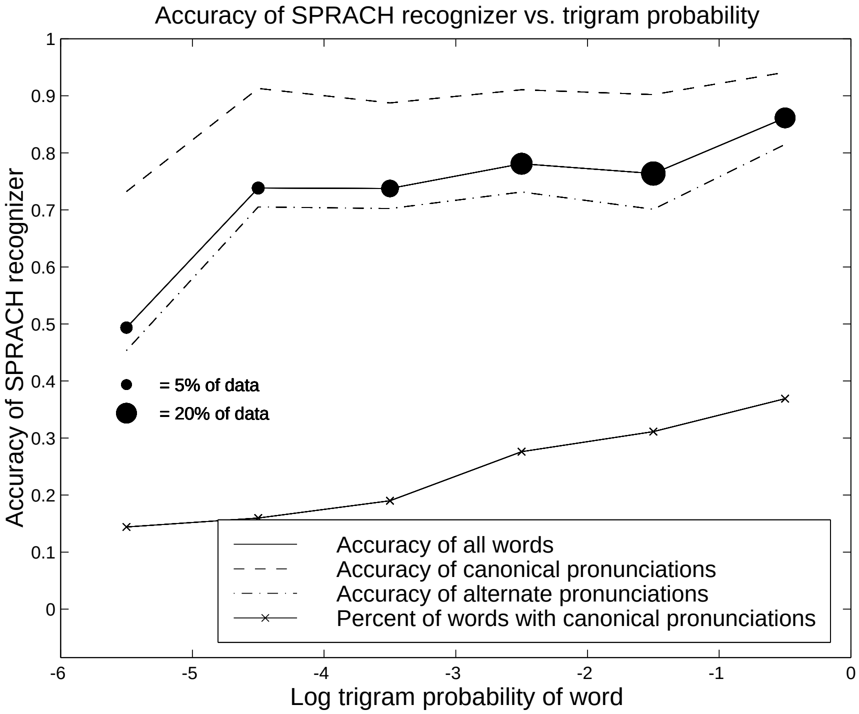


Figure 4.2c. Trigram probability

Figure 4.2: Accuracy of SPRACH Broadcast News recognizer dependent on varying factors.

instead of the smooth graph seen for Switchboard, the recognizer accuracy unexpectedly dips for words with a log unigram probability between -3 and -2. Further investigation revealed that the highest bin contained seven unique words[6] that are highly predictable from context. The second highest bin was dominated by a larger set of words that are less predictable, such as *is, this, it, who, well,* and *years*, but are common enough that they would not normally receive extra emphasis in speech. The third bin held many "content" words that probably received stress in the sentence, such as *morning, crime, campaign, economic,* and *American.* The third bin had more polysyllabic words (1.75 syllables/word average versus 1.19 syllables/word for the second bin); function words tend to be monosyllabic, while content words will range over a broader distribution.[7] It is likely that speakers emphasized these words more; stressed words are often clearer and consequently easier to transcribe automatically.

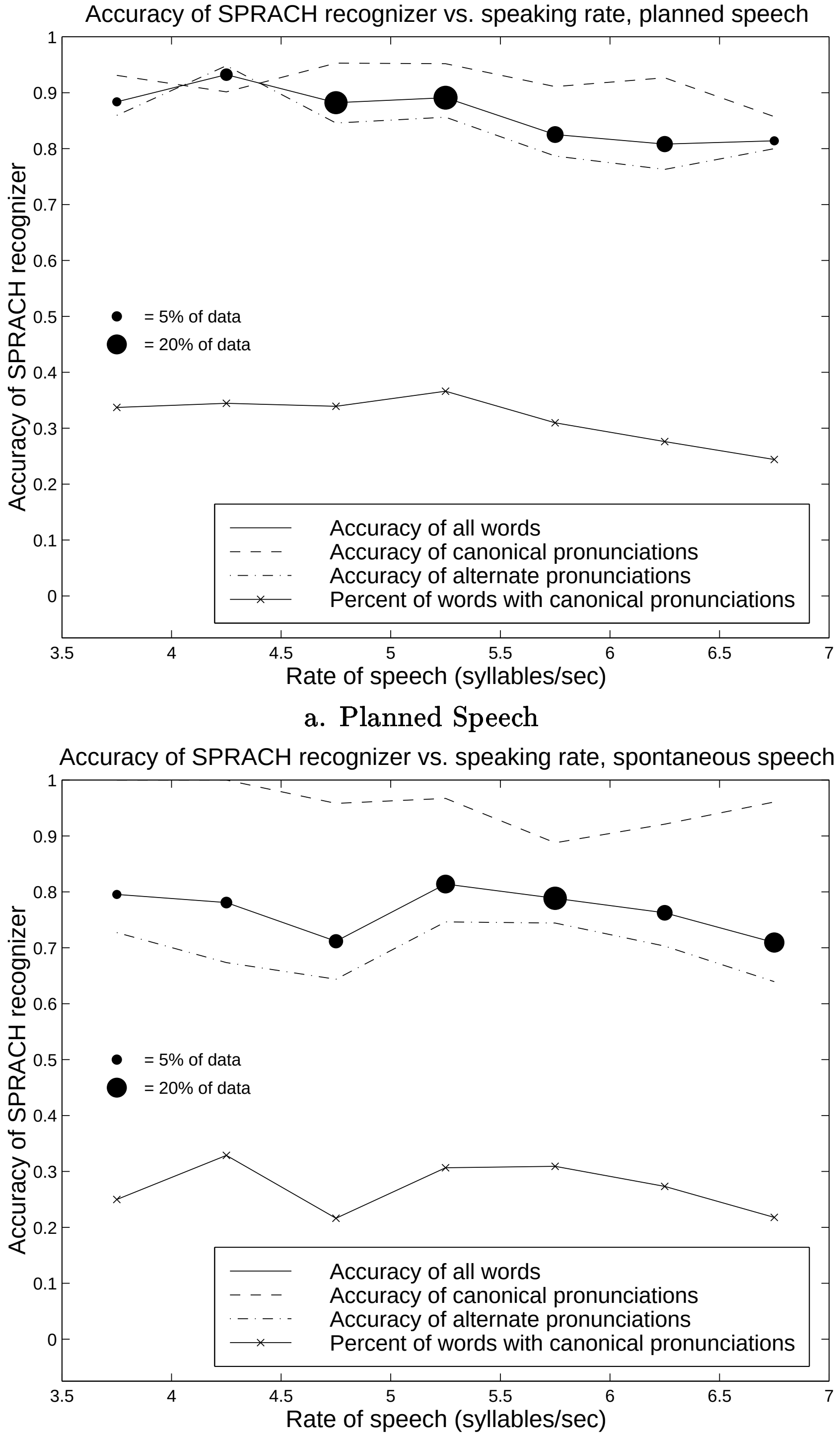
For speaking rate (Figure 4.2a), the percentage of words pronounced canonically peaks in the middle rates (5 to 5.5 syllables/second) and roughly tracks overall recognizer performance. There are several possible explanations for the shape of the curve: (1) the acoustic models are best when the speaking rate is roughly the mean, (2) the recognizer pronunciation model is geared toward mean speaking rates, or (3) the speech is clearest in the mean speaking rates. From these data one cannot distinguish among these hypotheses, and it is likely that all are true to some extent. Recognizer performance suffers at both extremes; this is different from the behavior of the Switchboard system, which performed well on slow speech, but much worse for fast speech. It is not clear why there is a discrepancy between these corpora in recognition error rates for slow speech, but the variability in performance for slow speech has been noted for other corpora as well [Mirghafori *et al.*, 1995].

When the data are separated out into planned and spontaneous conditions (Figure 4.3), some of the differences in recognizer performance between these two speaking modes become apparent. For planned speech the difference between canonically and non-canonically pronounced words is much less than for spontaneous speech, as demonstrated by the distance between the dashed and dot-dashed lines. In one histogram bin for the planned speech condition, words with alternate pronunciations were even recognized slightly more accurately than canonical words. Spontaneous speech is recognized much less consistently by this recognizer, and the performance gap between canonical and alternate pronunciations is very large, particularly for slow rates.

For an automatic phone-transcription system, the robustness of acoustic models has a serious impact on the pronunciation learning system. In Figure 4.4, I have broken up the test set into different focus conditions and show the recognizer accuracies for each condition. The canonical pronunciation percentage parallels recognizer performance relatively well; good performance of the acoustic and pronunciation models of the recognizer at the word level correlates with better matching of the phone transcript from the acoustic model to the pronunciation model. Recognizer performance for noisy conditions is some-

[6] These were *the, a, to, and, in, of,* and *that.*

[7] Switchboard exhibits similar characteristics in its unigram grammar, although it is not as marked; for instance, 13 words occupy the most frequent unigram bin, and the number of syllables per word for the second bin (1.11) is still less than for the third bin (1.51).



a. Planned Speech

b. Spontaneous Speech

Figure 4.3: Accuracy of SPRACH Broadcast News recognizer dependent on speaking rate for planned and spontaneous speech.

what lower than for planned or spontaneous conditions; this is also reflected in the lower percentage of canonical pronunciations. While far from conclusive, this suggests that the lack of acoustic model robustness to noise may be the cause of the poorer matching of the phonetic transcription to canonical models.

### 4.2.4 Summary

Despite the dependency of the automatic phone transcription system on recognizer acoustic models, there are distinct correlations in the Broadcast News database between pronunciation variations and recognizer error similar to those in the Switchboard database. Fast speaking rate again yields increased differences in the phonetic transcription, although, unlike in Switchboard, distinct changes from the baseform dictionary were observed for slow speaking rates as well. The word predictability results are tied much more tightly to the automatic transcriptions: unlike in Switchboard, more likely words are transcribed canonically far more often.

Analyzing data from the Broadcast News test corpus allows one to compare spontaneous to planned speech. The drop in recognition accuracy for non-canonical pronunciations is much larger in the more casual speaking style. In examining the response of the recognizer to a wider range of acoustic conditions, it is clear that pronunciation models generated by the acoustic models are less canonical when speech from noisier conditions is used for generation. The pronunciation model is therefore modeling not only linguistic phenomena, but also variations seen in the acoustic model due to noise.

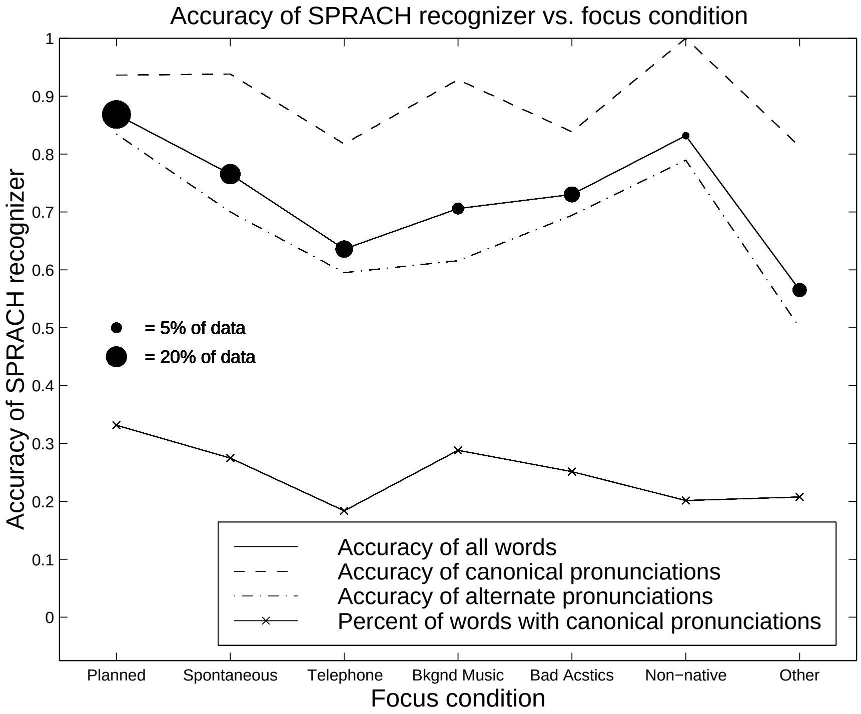


Figure 4.4: Accuracy of SPRACH Broadcast News recognizer for different focus conditions.

# Chapter 5

# Static Dictionaries

## 5.1 Introduction

As mentioned in the introduction to this thesis, the typical automatic speech recognizer contains a dictionary that describes how words map to the phone acoustic models. If the models in a pronunciation dictionary are fixed at the run-time of the recognizer, this is a *static dictionary.* This nomenclature is to distinguish these baseline models from the *dynamic* pronunciation models discussed in the next chapter that change pronunciation probabilities in response to several factors, including the word context and speaking rate.

In this chapter,[1] I detail experiments I conducted in building a new baseline dictionary for recognition of broadcast news reports. Three aims motivated this work. First and foremost, in order to fairly evaluate the dynamic pronunciation models in the next chapter, I wanted to build the best possible static dictionary, for use within the first-pass recognizer.

[1] This chapter contains some experiments previously reported in [Fosler-Lussier and Williams, 1999] and [Fosler-Lussier, 1999].

Since dynamic dictionaries require a second decoding pass over lattices or $n$-best lists of hypotheses generated by a first recognition pass, a good static dictionary was also necessary to generate the first-pass word hypotheses.

Another goal of this work was to evaluate design choices that any pronunciation modeler has to face. The central issue of pronunciation modeling is how to decide which baseforms to include in the dictionary and which ones to exclude. In these experiments, I used phone recognition to generate new pronunciations, but found that the models induced by phone recognition had many spurious pronunciations, decreasing their usefulness in the recognition system. I therefore focused my efforts on finding pronunciation selection criteria to constrain the possible variation in the models. One technique developed to improve pronunciation selection limited the number of possible variations that could be produced by the phone recognizer. In order to build an efficient model (in terms of recognizer run-time), it was also necessary to prune pronunciations from the dictionary; in this chapter I discuss various techniques for accomplishing this task. I also touch on the effects of increased training data and the robustness of pronunciation models to improvements in the acoustic model.

Finally, a more pragmatic reason behind this work was the need for improved dictionaries in the 1998 SPRACH System for Broadcast News transcription. Researchers at Cambridge University (CU), the University of Sheffield, and ICSI collaborated on a system for the 1998 DARPA Broadcast News Evaluation [DARPA, 1999]. The new pronunciation models developed in this work were integrated with acoustic models provided by CU and ICSI, language models from CU, and decoder technology from Sheffield to produce a complete system. Since the pronunciations were being developed as a semi-independent module, one goal was to provide an improvement in the dictionary that would be robust to changes in the acoustic model.

The 1998 DARPA evaluation also had a secondary decoding condition in which systems were restricted to operating within 10× real-time — besides recognition performance, fast decoding was also crucial. Pronunciation modeling can affect decoding time, since additional pronunciations increase the size of the lexical tree within the decoder, corresponding to longer search times. Therefore, in these experiments I evaluated the increase in decoding time for new dictionaries, as well as the word error rate metric. Timings were done on an Sun Ultra-30 (or one processor of a comparable 2-CPU Tatung Ultra-60 clone) with at least 768 MB of memory.[2]

The paradigm used for determining new pronunciations in this chapter is derived from collaborative work done at the Johns Hopkins Large Vocabulary Continuous Speech Recognition Summer Research Workshop in 1996 (WS96) [Weintraub *et al.*, 1997; Fosler *et al.*, 1996], described more fully in the next chapter. This model is based on the idea of *stream transformation* introduced in Chapter 2: a noisy channel model that captures the variations in how each phone is pronounced. In this model, the expected canonical pronunciations of each word are mapped to the realizations of the pronunciations in the corpus. For WS96, the realization phone sequence was produced by the acoustic model via phone recognition. The name *stream transformation* describes how this model operates

[2] Decoding processes never reached the size at which virtual memory was invoked.

— by probabilistically transforming a stream of canonical phones into a second stream of realizations.

In this chapter, the technology of stream transformation underlying the construction of static dictionaries is similar to the WS96 model, although particular implementation details are different. For instance, neural networks, rather than Gaussians, are used as acoustic models; $n$-ary phonetic features (described below) based on linguistic categories are utilized rather than the binary phonetic features in the WS96 model, similar to Riley's [1991] work. The model has also been extended to allow for selection of pronunciations based on acoustic confidence, as well as inclusion of baseforms for words not occurring in the original dictionary.

## 5.2 Phone recognition

For automatic machine learning of an ASR dictionary, one must have a source of pronunciations from which models are chosen. Most systems use *phone recognition* to generate a set of alternative pronunciations for words in the dictionary. This procedure is also called *phone constraint decoding* or *phone loop recognition.*

This procedure employs the acoustic models of the recognizer in order to generate new pronunciations. The technique as a whole is not novel — it has been used successfully in many systems [Humphries, 1997; Weintraub *et al.*, 1997]. Since it forms some of the basic building blocks for later work, and implementations vary across recognizers, I discuss in this section the particular choices made in building my system.

In the hybrid HMM-ANN system described in Chapter 2 (page 12), phone recognition is performed by substituting phones for words in the recognition system. Recall that the decoder takes four basic forms of information as input:

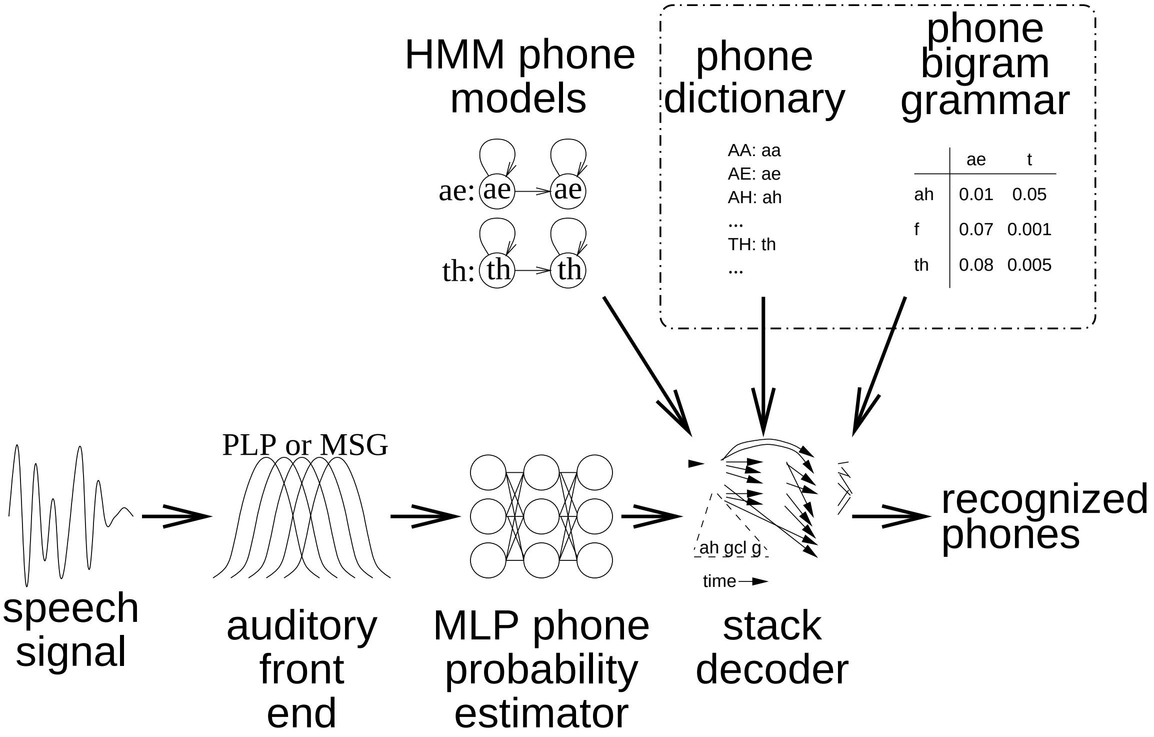
**Scaled likelihoods** The neural network provides posterior probabilities of phones given the acoustics ($P(Q|X)$), which are subsequently normalized by the priors ($P(Q)$) to give a scaled version of $P(X|Q)$.

**Phone models** These models give the basic HMM topology of each phone; in our system, these carry minimum duration information by repeating a number of states for each phone.

**Dictionary** For every linguistic unit (typically words), the dictionary lists the possible phone pronunciations that can represent that unit, with associated probabilities.

**Grammar** The grammar provides the prior probability of particular dictionary unit sequences (*e.g.*, sequences of words).

When the ICSI recognition system is utilized as a phone recognizer (Figure 5.1, the neural network still provides the acoustic probabilities $P(Q|X)$, and the phone models remain the same. The other two information sources change when phones rather than words are used as the linguistic units: the dictionary, instead of ranging over a vocabulary consisting of 65,000 words, has only 54 "words" (really phones), where each context-independent



| | ae | t |
|---|---|---|
| ah | 0.01 | 0.05 |
| f | 0.07 | 0.001 |
| th | 0.08 | 0.005 |

Figure 5.1: Phone recognition configuration for ICSI ASR system. The models appearing in the dashed box are the ones that differ from the word recognition system.

phone has one dictionary entry. The $n$-gram grammar also changes: instead of an $n$-gram over words, the phone recognizer utilizes a phone $n$-gram (*i.e.*, giving the probability of one phone following another).[3] This gives the system a model of the *phonotactics* of English; for example, *st* is a likely consonant cluster in English, but *nb* is not.

*a. Phone recognition output*

s ah m ax bcl b ey bcl b ao l s ah f ax bcl b eh z bcl b ao l

*b. Forced Viterbi alignment with canonical pronunciations*

| | | | | | |
|---|---|---|---|---|---|
| s ah m | aa | bcl b ey s bcl b ao l | s tcl t ah f | aa | bcl b ey s bcl b ao l |
| *some* | *uh* | *baseball* | *stuff* | *uh* | *baseball* |

*c. Mapping phone recognition to canonical pronunciations*

| | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| s | ah | m | ax | bcl | b | ey | | bcl | b | ao | l | s | | | ah | f | ax | bcl | b | eh | z | bcl | b | ao | l |
| s | ah | m | aa | bcl | b | ey | s | bcl | b | ao | l | s | tcl | t | ah | f | aa | bcl | b | ey | s | bcl | b | ao | l |
| *some* | | | *uh* | *baseball* | | | | | | | | *stuff* | | | | | *uh* | *baseball* | | | | | | | |

*d. Resulting alternative transcription*

*some* s ah m
*uh* ax
*baseball* bcl b ey bcl b ao l
*stuff* s ah f
*uh* ax
*baseball* bcl b eh z bcl b ao l

Figure 5.2: Building word pronunciations from phone recognition (episode b960529)

The output from the phone recognizer is a sequence of phones; no word breaks are inserted into the stream of phones (Figure 5.2a). Thus, the next task is to insert these word boundaries by aligning the *alternative* transcription provided by the phone recognizer against a *canonical* transcription that has word boundaries. The canonical transcription is obtained from a forced Viterbi alignment of the reference word sequence to the training data using a baseline lexicon (Figure 5.2b).

Since every phone in the canonical reference transcription is associated with a word, pairing these phones with phones in the alternative transcription will effectively insert word boundaries into the phone recognition. The procedure for phone pairing (described previously in Section 3.1.3) uses a generalized string-edit-distance algorithm to align two sequences of phones, where the distances between phones used by the algorithm are determined by the difference in phonetic features [Tajchman, 1994]. This alignment technique ensures that the minimum amount of phonetic variation is used to account for differences between phone streams: vowels are usually mapped to vowels, stop consonants to stop consonants, and so forth.

[3] In these experiments, I used a phone bigram grammar provided by Gethin Williams.

In the example in Figure 5.2, the baseline dictionary has only one pronunciation of *baseball*, namely `[bcl b ey s bcl ao l]`. However, the phone recognizer has recommended two new pronunciations: `[bcl b ey bcl b ao l]`, deleting the internal `[s]` sound, and `[bcl b eh z bcl b ao l]`, which monophthongizes the first vowel and voices the `[s]`.

### 5.2.1 Building a new dictionary

When the entire training corpus is transcribed in this manner, one can collect all of the pronunciation examples and build a new dictionary. Prior probabilities for the pronunciations of each word are estimated from the frequency counts of the pronunciations seen in the corpus.

For a first experiment, I automatically transcribed 100 hours of the 1997 Broadcast News training data [NIST, 1996]. The acoustic models, provided by Cambridge University, were a combination of four different Recurrent Neural Networks (RNNs). Two RNNs were trained on all 100 hours of the training data; one network was trained running forward in time, the other used a reversed waveform, looking backwards in time. The other two networks were trained only on the planned and spontaneous studio speech[4] in the first 100 hours.

After generating the phone recognition transcript for the BN97 training set, I calculated two new dictionaries containing the words present in the training set. The first dictionary allowed new pronunciations if there were at least two instances in the phonetic transcript (labeled as mincount=2); the second was more restrictive, requiring at least seven exemplars before inclusion (mincount=7).[5] Since not every word in the 65k vocabulary occurred in the training set, I combined the new dictionaries with a baseline dictionary in several ways.

For comparative purposes, I used the ABBOT96 dictionary as a baseline. This dictionary was derived from the 1996 ABBOT Broadcast News transcription system [Cook *et al.*, 1997] and contained an average of 1.10 pronunciations per word for the 65K vocabulary. I included new pronunciations provided by the phone recognition dictionaries in the recognizer in two ways: in the first strategy, the ABBOT96 pronunciations for words in the training set were replaced by the new dictionary. In the second scenario, the pronunciations from the ABBOT96 and new dictionaries were merged together, using the interpolation formula:

$$P_{\text{merged}}(\text{pron}|\text{word}) = \lambda P_{\text{ph.rec.}}(\text{pron}|\text{word}) + (1-\lambda)P_{\text{abbot}}(\text{pron}|\text{word}) \tag{5.1}$$

The value of the empirically determined smoothing parameter $\lambda$ did not affect results much within a broad range of values, so I set $\lambda = (1-\lambda) = 0.5$. Since the weighting factor

[4] Planned and spontaneous speech are designated by focus conditions F0 and F1 in the corpus, respectively.

[5] The choice of two and seven examples as thresholds is somewhat arbitrary, although there is some reason for each choice. I chose two for one of the thresholds in order to eliminate singleton events; if something occurs more than once, it's less likely to be spurious. The selection of seven as the other threshold is to match the (arbitrary) thresholding chosen by Weintraub *et al.* [1997] for their experiments with the Switchboard corpus. One would probably wish for a more statistically principled threshold criterion, but the point of this experiment is to get a baseline for performance of other commonly used techniques.

| Lexicon | Combination Style | % WER | Timing |
|---|---|---|---|
| Baseline (ABBOT96) | | 29.9 | 1.81× RealTime |
| PhoneRec mincount=2 | *replace* | 34.9 | 9.83 × RT |
| | *merge* | 29.2 | 9.93 × RT |
| PhoneRec mincount=7 | *replace* | 32.6 | 4.30 × RT |
| | *merge* | 29.7 | 4.19 × RT |

Table 5.1: Word error rate on Hub-4E-97-subset for unsmoothed phone recognition dictionaries using narrow (7-hypothesis) decoding parameters.

can be interpreted as a measure of trust in the source of a word's baseforms, a possible strategy would be to make $\lambda$ dependent upon frequency of a word's occurrence in the training data, using deleted interpolation [Jelinek and Mercer, 1980]. Section 5.3.4 touches on an experiment in which this technique is used to combine dictionaries.

With these new dictionaries, I decoded the Hub4E-97 subset defined previously by the SPRACH project. Since the acoustic models had improved since the lexicon training, the latest model was used for phone probability generation: a combination of the four-RNN system described above and a 4,000 hidden unit Multi-Layer Perceptron (MLP) trained on Modulation Spectrogram-Filtered (MSG) features. The NOWAY decoder was run with a narrow beam-width in the search (dubbed 7-hyps[6]) for expedited testing.

### 5.2.2 Results and discussion

In Table 5.1, we see that the dictionary replacement techniques introduce a significant number of errors. The more permissive "mincount=2" dictionary increases error rate by 16% (29.9% word error rate (WER) to 34.9%); using fewer pronunciations ("mincount=7") induces less error. On the other hand, smoothing with the prior dictionary pronunciations does slightly improve recognition performance.

These poor results illustrate an important point: pronunciation data derived from phone recognition can be very noisy, so it is important to introduce constraints to reduce the influence of noise. Figure 5.3 presents the variation seen in the pronunciations of the word *themselves* derived from phone recognition. Some learned variations are linguistically plausible, *e.g.*, the [m] sound being recognized frequently as [n] (likely as a consequence of place assimilation with the following [s]). Other variations, such as the replacement of [eh] with [ow] in the phone recognition dictionary, are very implausible; the poor word recognition results suggest that some way to disallow implausible variations should be employed.

While the canonical pronunciation [dh eh m s eh l v z] is rare in the ABBOT96 dictionary, it never occurs in the phone recognition. In fact, the most likely pronunciation

[6] 7-hyps has been the ICSI-internal catch-phrase for a narrow search; only 7 hypotheses are allowed to end at any particular time. Reduced search beam-widths are also employed; in other words, a hypothesis has to be much closer to the best current hypothesis (in log probability) to be kept within the search. The wider beam-width parameters are employed in a 27-hyp decoding.

| Phone Recognition | | Baseline | |
|---|---|---|---|
| PhoneRec mincount=2 | | ABBOT96 | |
| Prob. | Pronunciation | Prob. | Pronunciation |
| 0.24 | dh ax n s eh l z | 0.96 | dh ax m s eh l v z |
| 0.16 | dh ax n s ow z | 0.04 | dh eh m s eh l v z |
| 0.12 | tcl t ax n s eh l dcl d z | | |
| 0.12 | dh ax n s eh l | | |
| 0.12 | dh ax m s ow z | | |
| 0.08 | tcl dh ax m s eh l z | | |
| 0.08 | dh ax n s eh l s | | |
| 0.08 | dh ax m s eh l z | | |

Figure 5.3: Pronunciations of *themselves* from phone recognition and baseline dictionaries.

from ABBOT96 does not appear in the phone recognition dictionary either. It appears (from further examination of the phone recognition output from all words in the training set) that the [v] model is rather "weak" in that the phone recognizer frequently does not transcribe [v] sounds. This may be because [v] is an infrequent sound in English, so the bigram phone grammar may be discriminating against it, or that [v] is often a low-energy sound that was often absorbed into surrounding acoustic models (in this case, probably the neighboring [z]). [b] is often substituted by the recognizer for [v] (and vice versa); sometimes the transcription symbol is deleted altogether. Deletion of the [v] in *themselves*, though, produces a baseform homophonous with *them sells*. Intuitively, at least one representation of *themselves* should keep the [v] sound.[7]

Pronunciation models created by the acoustic model are thus not infallible. Averaging the two dictionaries reintroduces the [v] version of *themselves*; adding these averaging constraints across the entire dictionary reinforced canonical models and reduced the variance of the phone recognition models. Dictionary averaging is probably the simplest constraint that can be employed to suppress noise in phone recognition. In the next few sections, I will discuss other constraints that can be employed.

One other point: Table 5.1 indicates, in addition to word error rate, the decoding time required by the recognizer with various dictionaries. Including many different pronunciations increases the confusability of words in the corpus; thus, decoders must search larger spaces and rely more on the language model to disambiguate words and phrases that become homophonous due to the increase in the number of ways that words can be

[7] *Themselves* and *them sells* also have different stress patterns that could be used to differentiate the similar pronunciations; in addition, there are phonetic timing cues that can discriminate between these two phrases (see Church [1987] for more examples of nearly homophonous phrases that can be distinguished by phonetic cues such as aspiration). These cues are not usually employed within ASR systems because although humans use stress to differentiate word boundaries, the exact acoustic correlates corresponding to stress are not known. Research into automatic detection of stress may prove fruitful in determining pronunciation patterns [Silipo and Greenberg, 1999].

pronounced. A larger set of pronunciations translates into increased decoding time — the question is: how much longer? The best dictionary in terms of word error from the phone recognition experiments (mincount=2) takes five times as long as the baseline dictionary to run.[8] Since speed is often an issue in ASR systems, the tradeoff between accuracy due to better models and the run-time of the system should be kept in mind.

## 5.3 Smoothed phone recognition

In order to contain some of the model variability introduced by phone recognition, one needs to develop a sense of what alternatives should be allowed. Statistical techniques can tell us what variations are likely within a corpus. Learning these statistics on a word-by-word basis, however, is difficult because of the sparsity of data. Many systems use phone decision trees to learn how individual phones vary in pronunciation, dependent on the context. Infrequently seen alternatives are pruned from the trees (that is, disallowed), so that automatic retranscription of the training set is possible using a constrained set of possible phone sequences. Since the hypotheses of the phone recognizer are "smoothed" by the corpus statistics, I refer to this process as *smoothed phone recognition.*

### 5.3.1 Building decision trees

To initialize the pronunciation model, phone recognition was performed as above, producing alternative transcriptions and alignments to the canonical pronunciations. A new dictionary was *not* constructed, however. Instead, a model was trained to predict which transcribed phones corresponded to canonical dictionary phones. This model was trained using the alignment between the canonical and alternative transcriptions for training patterns.

For these experiments, I chose to use *decision trees* as a pronunciation predictor. Since decision trees (or d-trees) are well described in the literature (see, *e.g.*, Breiman *et al.* [1984]), I provide here only an intuitive description of the decision tree learning algorithm, providing examples from the pronunciation modeling domain.

A d-tree is a simple classifier that recursively finds optimal partitions of training data (according to a given criterion) to improve the classification of the data in a greedy manner (*i.e.*, at every step some criterion is maximized without regard to finding a global optimum of the criterion). In Figure 5.4, we see a collection of samples of the pronunciation for `[ey]` situated at the root of the tree. In this example, a binary question has been posed about the identity of the phone following `[ey]`; the data are partitioned into two sets based on the answer to the question. The learning algorithm chooses the *best* question that partitions the database (how to find this is described below); each subset of data is then recursively partitioned. Thus, the order of the questions in a d-tree is determined by which question is best at each level of recursion. Given this simple algorithmic structure, there are three main issues in building d-trees:

[8] We ran these experiments with a narrow pruning beam-width; with standard evaluation parameters, this system would probably be untenable.

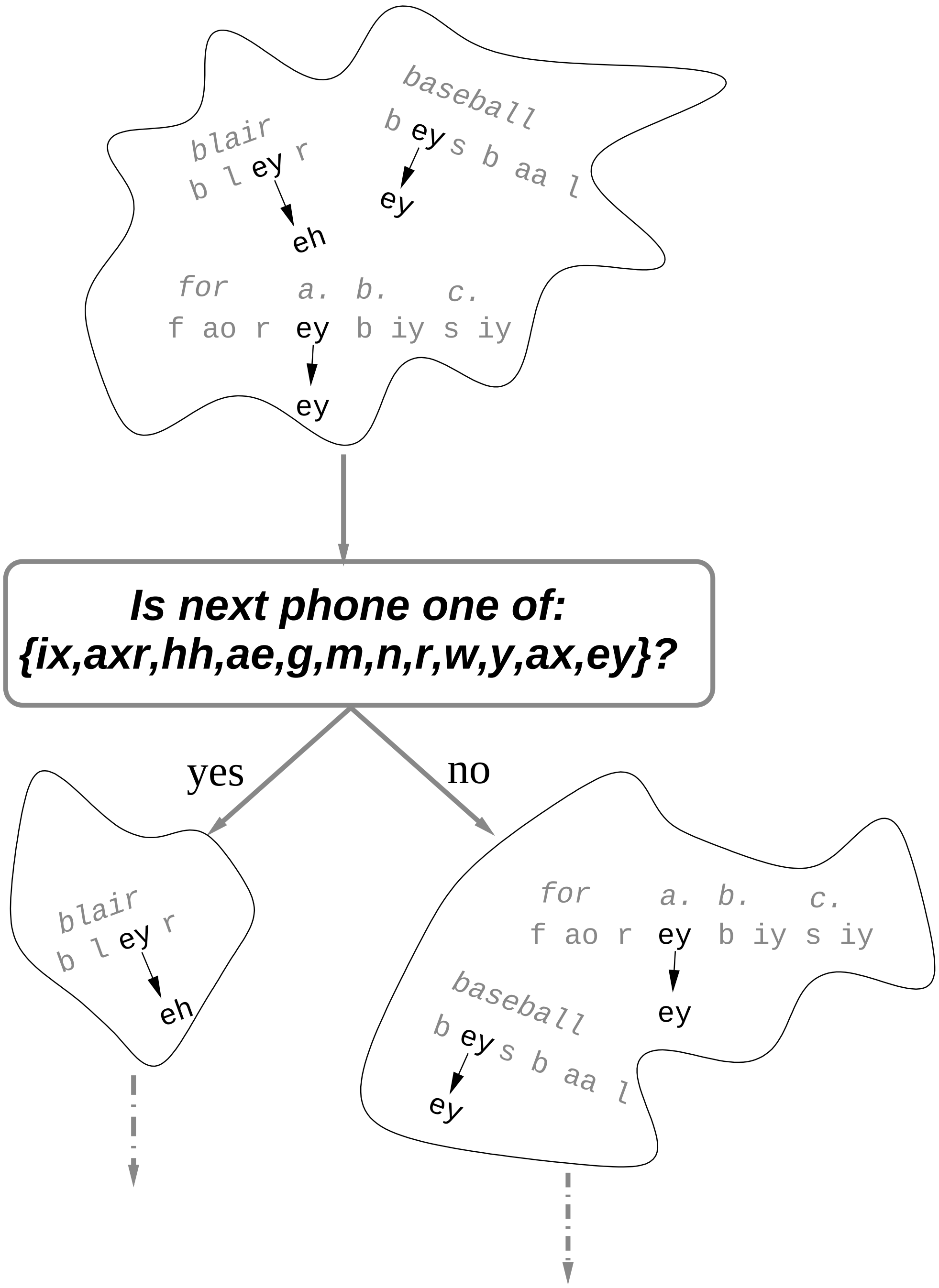


Figure 5.4: Sample partition question for the phone `[ey]`

1. What are the questions d-trees can ask in partitioning?
2. How do we find the *best* question that partitions the data?
3. When do we stop partitioning?

### D-tree questions

Like many other researchers [Young *et al.*, 1994; Odell, 1992; Riley, 1991], I have used linguistic concepts in decision tree formation. The algorithm was allowed to ask questions about the dictionary phone being modeled and the neighboring baseform phones from the dictionary-to-phone recognition mapping (Figure 5.2). These questions included:

**Phonetic identity:** the symbolic representation of the phone.

**Consonant manner:** the articulatory manner of the phone, if it was a consonant. Choices included *voiceless stop, voiced stop, silence, approximant, syllabic, voiceless fricative, voiced fricative, nasal.* Vowels were marked with *n/a.*

**Consonant place:** the articulatory place of the phone, if it was a consonant. Choices included *labial, dental, alveolar, post-alveolar, palatal, velar, glottal.* Vowels were marked with *n/a.*

**Vowel manner:** included *monophthong, w-diphthong, y-diphthong* for vowels. Consonants are marked with *n/a.*

**Vowel place:** encoded the height (*high, mid-high, mid-low, low*) and frontness (*front, mid, back*) of vowels. For diphthongs, this feature indicates the starting point of the vowel. Consonants are marked with *n/a.*

**Syllabic position:** the position of the phone within the structure of the syllable. Onset consonants were indicated with an *O*, coda consonants with a *C*, and vowel nuclei with a *N*. Silences carried the syllabic position distinction of *X*.

**Boundary markings:** indicated whether the phone started or ended a word (or both).

Every dictionary phone in the training set, with its corresponding realization in the phone recognition transcription, was annotated with these features, as well as the features of the previous and next dictionary phone forming a set of training patterns (or training instances) for the decision tree. A feature function $f_j \in \mathcal{F}$ is defined to return for each training pattern a value in the range $R_j$, where $j$ corresponds to a particular feature; for instance, when $j$ corresponds to the current phone's consonant place, $R_j = \{\text{labial}, \text{dental}, \text{alveolar}, \text{post-alveolar}, \text{palatal}, \text{velar}, \text{glottal}\}$. A partitioning question $Q_{j,S}$ is defined as a Boolean variable that is true if the feature corresponding to $j$ is in the set $S \subset R_j$, or, more formally, $Q_{j,S} = (f_j(\cdot) \in S)$. For example, in Figure 5.4 $f_j$ is the identity of the next phone, and $S = \{\texttt{ix,axr,hh,ae,g,m,n,r,w,y,ax,ey}\}$. Every training pattern is also associated with a class $c(i) \in C(\mathcal{I})$ corresponding to the phone recognition transcription for that sample. In this case, $C(\mathcal{I}) = \{\texttt{ey,eh}\}$.

### Finding the best partition (or, asking the best question)

In this implementation, all of the training patterns are divided into clusters, one for each dictionary phone. A separate d-tree is constructed for each dictionary phone by recursively partitioning the data for that phone. Partitions are chosen automatically by the algorithm using the predefined features (and questions) in the training data, determining the values that maximize a decision criterion called a *purity function*. Given a set of training instances $\mathcal{I}$, the d-tree algorithm chooses the best partitioning question $Q^*$ (from the set of all possible questions $\mathcal{Q}$), which is the question that maximizes the purity function. An example purity function, called the *information gain*, is described in Equation 5.2.[9]

$$Q^* = \operatorname*{argmax}_{Q_{j,S} \in \mathcal{Q}} H(\mathcal{I}) - [P(Q = \text{true})H(\mathcal{I}_{Q=\text{true}}) + P(Q = \text{false})H(\mathcal{I}_{Q=\text{false}})] \quad (5.2)$$

The subscripts on $\mathcal{I}$ indicate the subsets formed by the partition question $Q$, and $H(\cdot)$ is the entropy of the classes of realization phones at a particular node in the tree:

$$H(\mathcal{I}) = \sum_{c \in \{C(\mathcal{I})\}} -P(c) log_2 P(c) \quad (5.3)$$

Entropy is low when one class dominates the node and highest when all classes are equiprobable. Thus, one constraint placed by this function is to try to make the nodes further down the tree as *pure* as possible — to go from higher entropy to lower entropy. However, another constraint is at work as well. The entropies in each branch are weighted by the proportion of examples in each leaf. An even split with less entropy reduction in each node can still reduce the overall entropy more than a split that has a few examples with very low entropy in one node, and only a minor reduction in entropy for the vast number of examples in the other. Other criteria that weight this tradeoff differently are also possible (see, *e.g.*, the GINI index described in Breiman *et al.* [1984]).

Partitioning continues on each sub-node, where a different question can be employed; after each partitioning, the probability distribution $P(C)$ can be calculated at the leaves based on the training data left in that node.

### When to stop partitioning

There are no hard and fast rules for when to stop partitioning the data. Usually, heuristics applied to the depth of the tree or the number of samples in each node are used to determine ending criteria. In the experiments run here, I set a minimum example count for each node (at five examples per node); splits are not allowed if either half of the partition would have less than the minimum number of samples. While this is a bit sparse for probability estimation, Breiman *et al.* [1984] report that their best results occurred when the tree was overgrown and then pruned back using a cross-validation technique. In

[9] The maximized quantity here is also known as the *mutual information* between the instances $\mathcal{I}$ and the partitioning question $Q_{j,S}$; an alternative formulation of this quantity is $I(\mathcal{I}; Q_{j,S}) = H(\mathcal{I}) - H(\mathcal{I}|Q_{j,S})$.

particular, they use a 10-fold jackknife technique, determining the appropriate tree size by growing the tree on 90% of the data. The algorithm then prunes back the tree, optimizing classification on the remaining 10% of the data. The procedure is repeated nine more times, rotating the data used for training and test. The parameters are then averaged, and a tree is then regrown using all of the data, stopping when the learned parameters are reached.

For my experiments, I used 10-fold cross-validation, but then cross-validated again on an independent test set for further pruning, using 90% of the entire training set in the jackknife procedure, with another 10% used to reprune the trees.

### 5.3.2 Smoothed alignments

Given the set of learned trees, I was then able to relabel the training set acoustics by generating a constrained set of pronunciations with the trees. Every phone in the Viterbi alignment of the baseline dictionary to the training set (Figure 5.2b) was transformed into a realization phone sequence with the trained decision trees. The d-tree evaluation algorithm used attributes of each baseform phone and its immediate neighbors to navigate the set of questions encoded in the tree for that phone, starting at the root of the tree and taking the branch at each node corresponding to whether the answer to the question at that node was true or false. Every dictionary phone was therefore associated with a particular decision tree leaf containing the probability of alternative phones (Figure 5.5: Tree evaluation).

The distributions from the d-trees were then compiled into a finite state grammar (FSG) of alternative phone pronunciations by the following algorithm: for the $n$th phone in the canonical transcription, the appropriate tree distribution $d$ was found. Between nodes $n$ and $n+1$ in the FSG, an arc was added for every recognition phone in $d$, labeled with the appropriate probability. Phone deletions were accommodated through the insertion of null transitions. Phone insertions, corresponding to a phone pair in the leaf, were accommodated by adding extra nodes in the graph; the first arc retained the pronunciation probability, while subsequent arcs were given a probability of 1.0 (see `[bcl_b]` in Figure 5.5 for an example). Some smoothing was applied during this FSG construction by disallowing any transitions with below-threshold probabilities (the threshold was arbitrarily set to 0.1).

Following d-tree training and FSG compilation, a new static lexicon was created. The compiled FSG was realigned to the training data to obtain a *smoothed* phone-constraint decoding. Since the FSG decoder produced both a word and phone alignment, the new alternative transcription was easily converted into a new static lexicon by gathering all of the examples of each word and determining the probability distribution of pronunciations for that word.

### 5.3.3 Building Broadcast News models

I constructed phone trees for the 1997 Broadcast News training set with the IND decision tree toolkit [Buntine, 1992]. One tree was grown for each dictionary phone; all examples of the same dictionary phone in the corpus were collected into one database. The ICSI system typically represents stop closures and bursts as separate phones; I built only one tree for each stop consonant, concatenating the phone recognition targets for closures

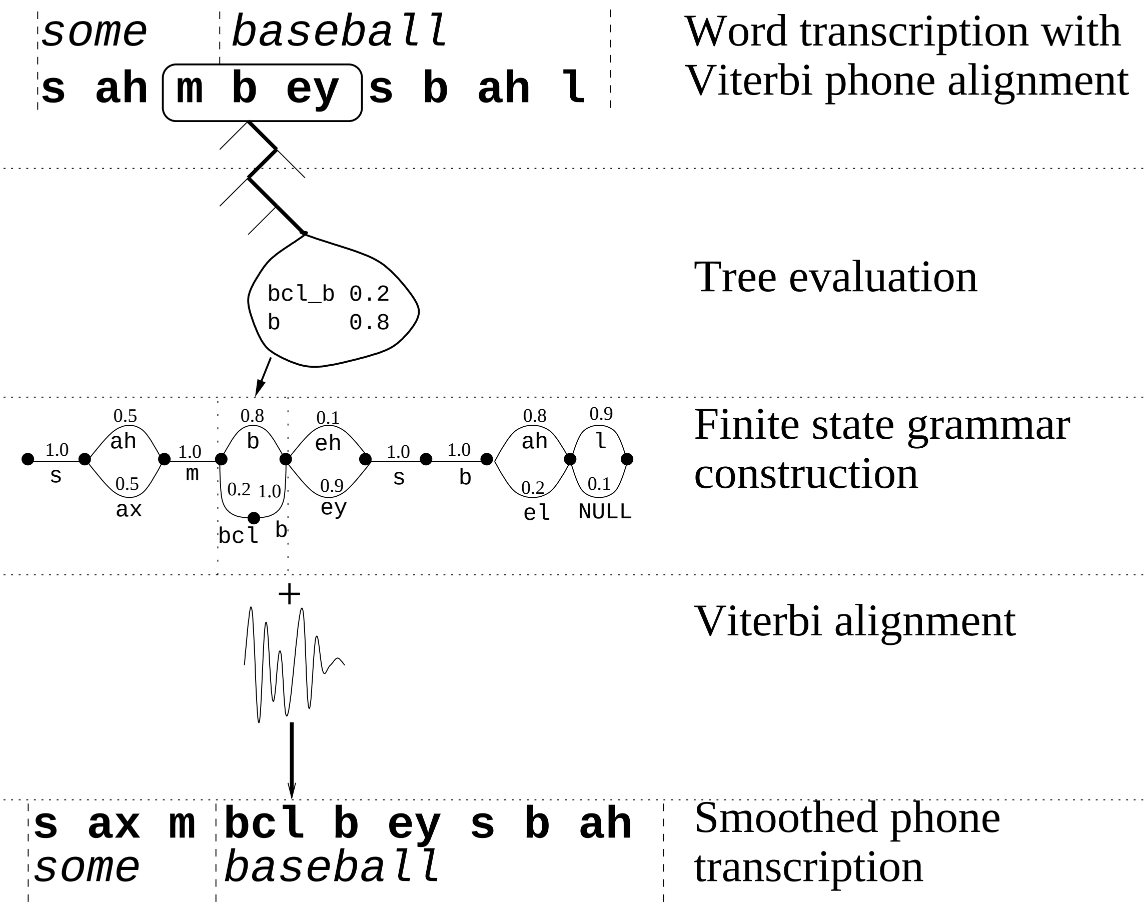


Figure 5.5: Building smoothed alignments from phone trees.

and bursts. Thus, if the phone recognition produced `[bcl b ey bcl b ao l]`, `[bcl_b]` would be used as a target for both instances of /b/. A sample tree is given in Figure 5.6.

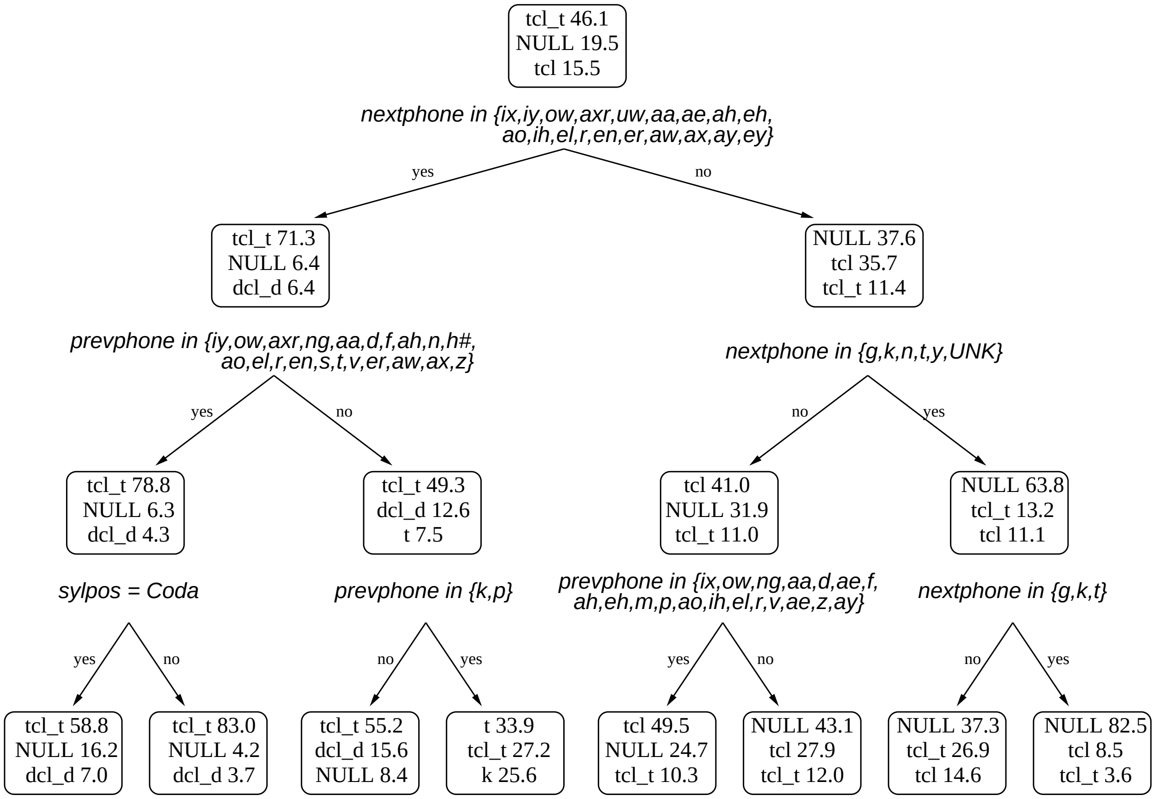


Figure 5.6: A sample tree grown from the phone recognition outputs from the phone `[t]`, artificially pruned at depth 3. The three most likely phones at each node are shown.

The dictionary resulting from the collection of pronunciation alternatives in the smoothed phone transcription was still somewhat noisy, particularly for infrequently occurring words. However, as Figure 5.7 shows, the distribution of pronunciations for the word *themselves* is much more peaked. Two baseforms account for 88% of the examples of the corpus, one of which (`[dh ax m s eh l v z]`) is found in the ABBOT96 dictionary. The main phone variation exhibited in these examples is a substitution of `[n]` for `[m]` — possibly representing a place assimilation of the nasal to the following `[s]`.[10] Meanwhile, the `[v]` of *themselves*, which was deleted completely in the original phone recognition, occurs much more frequently, being deleted in only 4% of the examples.

As in the phone recognition dictionary experiments, I merged the newly obtained pronunciations with those from the baseline ABBOT96 lexicon to smooth the pronunciation probability distributions, particularly for words with low counts. The interpolation

[10] Because these are acoustic models, not humans, transcribing these instances, one should not be too hasty in jumping to conclusions here. However, it is interesting that these data do parallel a known linguistic phenomenon.

| | Decoding Parameters | |
|---|---|---|
| Lexicon | 7-hyp. WER (%) | 27-hyp. WER (%) |
| Baseline: ABBOT96 | 29.9 | 27.5 |
| Phone Recognition: mincount=2,$\lambda$=0.5 | 29.2 | - |
| Smoothed Trees: $\lambda$=0.5 | 28.9 | 27.1 |

Table 5.2: Word error rate on Hub-4E-97-subset for static lexica.

parameter was set to $\lambda = 0.5$ (Equation 5.1).

| PhoneRec mincount=2 | | Smoothed phone trees mincount=2 | |
|---|---|---|---|
| Prob. | Pronunciation | Prob. | Pronunciation |
| 0.24 | dh ax n s eh l z | 0.55 | dh ax n s eh l v z |
| 0.16 | dh ax n s ow z | 0.33 | dh ax m s eh l v z |
| 0.12 | tcl t ax n s eh l dcl d z | 0.04 | dh ax n s eh l v s |
| 0.12 | dh ax n s eh l | 0.02 | dh ax n s eh l |
| 0.12 | dh ax m s ow z | 0.02 | dh ax n s ah l v z |
| 0.08 | tcl dh ax m s eh l z | 0.02 | dh ax m s eh l v s |
| 0.08 | dh ax n s eh l s | 0.02 | dh ax m s eh l |
| 0.08 | dh ax m s eh l z | 0.02 | d ax n s eh l v z |

Figure 5.7: Pronunciations of *themselves* from smoothed and unsmoothed phone recognition

In a first experiment, I replaced the dictionary in the NOWAY decoder with the tree-based dictionary and decoded with the narrow pruning beam-width parameters (Table 5.2: 7-hypothesis decoding). The Smoothed Trees dictionary outperformed both the ABBOT 96 dictionary and the previous best phone-recognition dictionary, although not by a statistically significant margin.

The improved performance means that the new dictionary matched the acoustic models better when the search was very restricted. Wider decoding beam-widths, however, reduce the gain provided by the new dictionary, as seen by the 27-hypothesis (evaluation quality) results in Table 5.2. This is not unexpected, as wider decoding beam-widths translate into allowing the decoder more chances to guess at the right word. In this situation, the acoustic models will (on average) be more likely to match the baseline dictionary pronunciation, so the effect of adding new pronunciations is reduced. Another way to look at this result is that the new dictionary is a closer match to the acoustic models, because reducing the search space does not induce as great a reduction in performance.

### 5.3.4 Dictionary pruning

The Smoothed Trees dictionary described above increased the number of pronunciations per word to 1.67 from ABBOT96's 1.10. This large rise in the number of pronunciations

| Lexicon | Pruning Style | % WER | Timing |
|---|---|---|---|
| Baseline (ABBOT96) | n/a | 29.9 | 1.81× RT |
| Phone Recognition | mincount=2, merged | 29.2 | 9.93× RT |
| Smoothed Trees | no pruning | 28.9 | 6.69× RT |
| prune low probability prons | $p_{\text{pron}}<0.1*p_{\max}$ | 29.5 | 2.50× RT |
| | $p_{\text{pron}}<1.0*p_{\max}$ | 31.4 | 1.85× RT |
| Count-based pruning | log count $\alpha$ =1.2 | 28.8 | 2.72× RT |
| + Deleted Interpolation | log count $\alpha$ =0.5 | 28.9 | 3.79× RT |

Table 5.3: Word error rate on Hub-4E-97-subset for various pruning methods using narrow (7-hypothesis) decoding parameters.

increased decoding time almost four-fold over the ABBOT96 dictionary (Figure 5.3). While the decoding time of 6.69× real-time was better than the 9.93× real-time provided by the best Phone Recognition dictionary, the long decode time was still devastating for the development of the SPRACH 10× real-time system, particularly since the system was running with a very narrow search. The next avenue of research was therefore to attempt to keep the improvements of the new lexicon, while pruning unnecessary pronunciations from the dictionary.

In order to reduce the decoding time, I investigated two dictionary pruning techniques. In the traditional dictionary pruning scheme at ICSI, baseforms were removed from the lexicon if they had a prior probability that was less than some fraction of $p_{\max}$, the prior probability of the most probable baseform for the word. While this significantly reduced decoding time, it also halved the gains from the new dictionary, even for low pruning values (Table 5.3). Reducing the lexicon to a single baseform per word ($p_{\max} = 1.0$) also significantly hurt performance with no corresponding speedup relative to the ABBOT96 baseline.

Since high-frequency words usually have more pronunciation variants in continuous speech, I developed a new pruning technique based on the number of occurrences of the word in the training data. In this second scheme, the maximum number of baseforms $n_i$ for each word $w_i$ was determined by

$$n_i = \lfloor \alpha \log_{10} \mathrm{count}(w_i) \rfloor + 1 \quad , \tag{5.4}$$

where $\alpha$ is a parameter that can be tuned to adjust the number of baseforms allowed. The $n_i$ most likely baseforms for each word were included in the dictionary, with probabilities rescaled to account for the probability mass of the removed pronunciations. As shown on the count-based pruning line of Table 5.3, this method facilitated lower decoding times (only 1.5 times that taken by the ABBOT96 dictionary) without any increase in word error rate relative to the unpruned lexicon.

The above results used a lexicon that integrated the new pronunciations with the ABBOT96 lexicon using even weighting (Equation 5.1). In order to determine if the even weighting affected results, I reconstructed the lexicon using deleted interpolation [Jelinek and Mercer, 1980] to set the parameters. In deleted interpolation, the smoothing parameter

| Dictionary | WER (%) | Decode time |
|---|---|---|
| Baseline (ABBOT96) | 27.5 | 21.73 x RT |
| Smoothed Trees: no pruning | 27.1 | 72.03 x RT |
| log count (SPRACH98) | 26.9 | 33.07 x RT |

Table 5.4: Word error rate on Hub-4E-97-subset for various pruning methods using full (27-hypothesis) decoding parameters.

$\lambda$ is set based on the count of each word in the corpus; all words with similar frequency in the training corpus share the same $\lambda$. An excellent description of this algorithm can be found in [Jelinek, 1997]. As the last line of Table 5.3 shows, the use of deleted interpolation did not improve recognition results.

The results in Table 5.4 show that gains in both recognition performance and speed provided by the log count pruning scheme carry over to the wider beam (27-hypothesis) decoding condition. A lexicon pruned using this second scheme was therefore selected for use in the SPRACH98 system. It is interesting to note that the increase in decoding times over the baseline for the 27-hyp (wide-bandwidth) condition is almost identical to the 7-hyp condition (3.3× for the unpruned lexicon, 1.5× for the log count dictionary). This implies that the larger lexicon is not causing a larger fan-out in the search of the recognizer; the increase in decode time is almost completely accounted for by the increase in stack size.[11] The improvements from this lexicon, while modest, were duplicated across test sets (including the full 1997 Hub4 Evaluation, where the small difference is statistically significant) and with different acoustic models.

### 5.3.5 Choosing pronunciations via confidence scores

The lexicon pruning experiments showed that sub-selecting a good set of new pronunciations from data is important for good performance, from both word error rate and decoding time points of view. However, it is not clear that selecting the most likely pronunciations is the best criterion for building a dictionary. In this section,[12] I describe an alternate technique for pronunciation selection based on the acoustic confidence of the pronunciations.

In the smoothed phone recognition paradigm, the dictionary is built from the Viterbi alignment of a phone-based finite state grammar (FSG) to the training set. This means that a choice of the best pronunciation for a word is forced for every instance. However, some matches may be better than others; an instance that is clearly pronounced in a particular way is counted with the same weight in the new dictionary as an example that may be modeled poorly. Moreover, if two pronunciations for an instance are close (perhaps due to an acoustic model ambiguity), only the “better” pronunciation influences the dictionary, since Viterbi is a winner-take-all strategy — the second-place pronunciation,

[11] Decode time increases roughly as the square of the number of hypotheses allowed.

[12] This section represents joint work with Gethin Williams. I am indebted to Williams for evaluating the confidence of all of the word pronunciations from the previous section in the 1997 Broadcast News training set.

even if close in the acoustic score, contributes nothing to the final model.

One way to describe the relative differences among models is to use an *acoustic confidence score*, which gives an estimate of model quality for a given segment of speech. Williams [1999] describes a technique for computing acoustic confidences for individual phones and complete words using a neural network acoustic model. The multi-layer perceptron acoustic probability estimator used in the ICSI recognition system produces posterior probabilities of the form $P(q_k^n|X^{n\pm c})$, where $q_k^n$ is a phonetic state for phone type $k$ at time $n$, and $X^{n\pm c}$ is the acoustic feature vectors centered at $n$ with a context window of $\pm c$ frames.[13] With the estimate of the posterior probabilities of phones, Williams defines a normalized posterior probability confidence score for a phone, given acoustics between start time $n_s$ and end time $n_e$:

$$nPP(q_k, n_s, n_e) = \frac{1}{n_e - n_s} \sum_{n=n_s}^{n_e} logP(q_k^n|X^{n\pm c}) \tag{5.5}$$

This measure, the $nPP$ score, is the duration-**n**ormalized log **p**osterior **p**robability for the phone given the acoustic segment.

In order to evaluate confidence for an entire word pronunciation, the $nPP$ scores must be combined in some fashion. Williams' experiments showed that the following algorithm worked reasonably well:

1. Determine the pronunciation for the word $w_j = q_1 q_2 \ldots q_L$.
2. Given acoustics $X^{n_s \cdots n_e}$, determine the best (Viterbi) alignment of $q_1 \ldots q_L$, producing a segmentation $X^{n_{s1} \cdots n_{e1}}, X^{n_{s2} \cdots n_{e2}}, \ldots X^{n_{sL} \cdots n_{eL}}$.
3. Compute the normalized posterior probability for the word $w_j$:

$$nPP_w(w_j, n_s, n_e) = \frac{1}{L} \sum_{l=1}^{L} nPP(q_l, n_{sl}, n_{el}). \tag{5.6}$$

Note that the $nPP_w$ function requires three arguments: a word pronunciation $w_j$, and the start and end times of the associated acoustics. To compute the confidence of all of the pronunciations of word $w$, one must determine where the instances of $w$ are located in the corpus. Thus, the following algorithm was used to determine average confidence for all pronunciations of the word $w$:

1. For all instances $1 \ldots I$ of word $w$ in the training set ($w_i$):
   (a) Determine $n_s$ and $n_e$ for $w_i$ through forced Viterbi alignment.
   (b) For each pronunciation $p_j$ of $w$:
   $\mathrm{Conf}(p_j) = \mathrm{Conf}(p_j) + nPP_w(p_j, n_s, n_e)$

[13] The RNN used by Williams computes a similar probability, estimating $P(q_k^n|X^{1..n})$; the only difference with the quantity estimated by the MLP is the length of the acoustic context used to estimate the probability of $q_k^n$.

| Dictionary | WER (%) | Decode time |
|---|---|---|
| Baseline (ABBOT96) | 27.5 | 21.73 x RT |
| Smoothed Trees: frequency log count (SPRACH98) | 26.9 | 33.07 x RT |
| confidence log count | 26.6 | 30.45 x RT |

Table 5.5: Word error rate on Hub-4E-97-subset by method of pronunciation selection, using full (27-hypothesis) decoding parameters.

2. For each pronunciation $p_j$ of $w$:

$$\text{AvgConf}(p_j) = \frac{\text{Conf}(p_j)}{I}$$

Using this word acoustic confidence, we could then re-rank pronunciation alternatives for each word by the average confidence of the alternatives in the 1997 Broadcast News training corpus. The ranking affected only the selection of which alternatives should be included in the model. The pronunciation probabilities, on the other hand, were determined as before by the Viterbi counts in the corpus. We did not adjust the $\alpha$ parameter that dictated the log scaling of the corpus counts.

Compared to the SPRACH98 dictionary, different pronunciations were chosen in the confidence-based dictionary for 3343 of the words. I scanned the differences in pronunciation between the two dictionaries; the confidence-based pronunciations made sense linguistically slightly more often, although it was often difficult to judge which pronunciation was "correct."

Table 5.5 shows that confidence-based selection of pronunciation alternatives gives a slight boost in performance over the frequency-based method. In addition, decoding was also slightly faster than in the previous experiment, indicating perhaps that the improved dictionary model better matched the acoustic model.

Acoustic confidence is a very promising alternative metric for determining the pronunciations used in a system dictionary. This technique moves away from the winner-take-all strategy employed previously and appears to perform slightly better at the selection task, even without optimizing the log-scaling selection parameter. On the other hand, it is relatively expensive to compute (one must evaluate *every* pronunciation for *every* instance of the word). The computational expense, in addition to the comparative lateness of the introduction of this technique into the research paradigm, prohibited using confidence-based evaluation in most of the later experiments described in the thesis. In future experiments, however, I hope to employ confidence measures more thoroughly in pronunciation learning.

### 5.3.6 Including more training data and better acoustic models

Development of an automatically learned pronunciation model is, of course, dependent on the acoustic model used to generate the pronunciations.This section details experiments testing the robustness of the new static dictionaries to changes in acoustic models and decoders, and then examines the effect of retraining the pronunciation model using pronunciation data generated by improved acoustic models.

During the period of rapid system development for the 1998 Broadcast News evaluation, the SPRACH team produced new acoustic models almost daily. Two options were available for developing new pronunciation models: either fixing the acoustic model at some stable point, or always using the best available model at the time. The former option (which I favored for these experiments) facilitates comparisons across changes in the pronunciation modeling technique, while the latter reduces the mismatch between the pronunciation model and the ever-changing acoustic models.

Holding the acoustic model constant does increase the danger that the learned dictionary is tuned to the particular probability estimator and therefore will not work when used with improved models. In this section, I describe experiments conducted with updated acoustic models to ascertain whether the modest performance increase from the new pronunciation dictionary is independent of the acoustic model. I also consider the effect of retraining the pronunciation model using the new acoustic model, along with a larger training set size — the original dictionary was trained using only half of the available Broadcast News training data. Third, since the phone recognition I used for training samples was provided by the NOWAY decoder, it remained to be seen whether switching to a different decoder would affect results. During the SPRACH system development, the CHRONOS "time-first search" decoder became available from our partners at SoftSound Limited [Robinson and Christie, 1998];[14] I compared results from the two decoders in hopes that the pronunciation models would be robust to the different search strategies in each decoder.[15]

### Pronunciation and acoustic models

As described in the initial experiments of the previous sections, I used an acoustic model from an intermediate stage in the SPRACH system's development. I combined the 1997 ABBOT PLP-based recurrent neural network (RNN) context-independent phone classifier with a 4,000 hidden unit multi-layer perceptron (MLP)using modulation-filtered spectrogram (MSG) features. Both networks were trained only on the 1997 BN training data. For these experiments, I will refer to this combined acoustic model as A-Model I. To build the dictionary for the SPRACH system, I used A-Model I to generate a smoothed phone transcript on the 100-hour 1997 BN training set.

After the 1998 evaluation, I retrained the pronunciation models using an improved acoustic model (A-Model II) that combined a PLP-based RNN (trained only on the 1997 data), and two 8,000 hidden unit MLPs, trained on PLP and MSG features calculated for the full 200 hour training set, respectively. All 200 hours of the 1997 and 1998 BN

[14] We have used two different decoders in our experiments because the CHRONOS decoder is an order of magnitude faster than NOWAY, but has the shortcoming of producing only a single best hypothesis, not the lattice of hypotheses required for the dynamic dictionary recognition experiments in the next chapter. The CHRONOS decoder tends to outperform NOWAY by a few tenths of a percent (WER) with the particular parameter settings we are using, although this is not consistent.

[15] It is not obvious that improvements in pronunciation models would be consistent across decoders, because CHRONOS orders its search differently than NOWAY and has different pruning methods, so adding new pronunciations could affect recognition performance by altering the search space.

| | Acoustic Model/Decoder | | | |
|---|---|---|---|---|
| | A-Model I | | A-Model II | |
| Dictionary | NOWAY | CHRONOS | NOWAY | CHRONOS |
| ABBOT96 (baseline) | 27.5 | 27.5 | 24.2 | 24.0 |
| SPRACH98 (BN97 training) | 26.9 | 27.2 | 23.7 | 23.4 |
| BN97+98 training | 26.9 | 26.7 | 23.5 | 23.2 |

Table 5.6: Word error rates for Hub4E-97-subset with varied dictionaries, acoustic models, and decoders

training sets were transcribed using the smooth phone recognition procedure; the resulting dictionary is labeled "BN97+98 training" in these studies.

### Experimental results

Table 5.6 describes the results of varying the three major parameters in these experiments: three different dictionaries, two acoustic models, and two decoders. In general, A-Model II has about a 3.5% absolute lower word error rate than A-Model I. Comparing the SPRACH98 dictionary to the ABBOT96 dictionary, we see that the change in acoustic models makes no difference in the improvement between A-Model I and II in most cases (0.5-0.6%). Changing the decoder also makes very little difference in the results; one can see that modifying the dictionary does change the search space somewhat (*cf.* the difference between ABBOT96 and SPRACH98 for NOWAY and CHRONOS with A-Model I), but the SPRACH98 dictionary does improve recognition in all cases.

Thus, the automatically derived dictionary is at least *somewhat* independent of the acoustic models from which it was derived; one does not have to retrain the pronunciation models every time the acoustic models are changed. Nonetheless, the dictionaries probably still depend on the corpus and overall recognition system.

Doubling the amount of pronunciation training data had only a small effect on performance. Comparing the last row of Table 5.6 (BN97+98 training) to the SPRACH98 results shows only a 0.2% absolute gain with A-Model II. The results for A-Model I are less clear; there was no improvement with NOWAY, while the CHRONOS numbers improved to make up for the poorer performance of the SPRACH98 dictionary in that condition. This is not surprising, since the BN97+98 training data were generated with A-Model II, so a dictionary constructed from that data should match A-Model II better. All of these gains are under the statistical significance margin, so all that can be concluded is that there is no great effect from increasing the training data. However, the BN97+98 dictionary is the best in all conditions, so it became the default dictionary in the ICSI BN recognition system.

I decoded the full 1997 Hub4E evaluation set with these three dictionaries (using CHRONOS and A-Model II) (Table 5.7), and found that the improvement pattern for the full set closely resembled that for the subset, with BN97+98 only just edging out SPRACH98. The improvement of the BN97+98 system over the ABBOT96 dictionary is significant at $p < 0.02$; SPRACH98's results are significant at $p < 0.05$.

| | | Focus Conditions | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Dictionary | Overall | F0 | F1 | F2 | F3 | F4 | F5 | FX |
| ABBOT96 (baseline) | 23.0 | 14.6 | 24.4 | 31.8 | 31.3 | 27.0 | 22.9 | 35.4 |
| SPRACH98 (BN97 training) | 22.4 | 14.2 | 23.6 | 31.4 | 30.7 | 25.3 | 23.2 | 35.1 |
| BN97+98 training | 22.3 | 14.2 | 23.6 | 30.9 | 31.3 | 25.1 | 22.4 | 35.1 |

Table 5.7: Word error rates for Hub4E-97, using A-Model II and the CHRONOS decoder. The focus conditions for Broadcast News include Planned Studio Speech (F0), Spontaneous Studio Speech (F1), Speech Over Telephone Channels (F2), Speech in the Presence of Background Music (F3), Speech Under Degraded Acoustic Conditions (F4), Speech from Non-Native Speakers (F5), All Other Speech (FX).

I also determined the focus conditions for which improvements were shown in the full evaluation set. Comparing ABBOT96 to BN97+98, we see decreases in word error rate in almost every category (except F3, speech with background music), leading to a significant improvement overall. Most of this gain was obtained from training with the first half of the data; between SPRACH98 and BN97+98, no difference in error rate was seen in the planned and spontaneous studio focus conditions (F0 and F1), from which the majority of the test data is drawn.[16]

Automatically derived models, one must remember, capture various sources of variation in acoustic models. The primary reason to build a new dictionary is to capture linguistic variability, such as the [n]/[m] alternation in *themselves* illustrated in Figure 5.7. One would expect that more casual speaking conditions (*e.g.*, focus condition F1) would benefit most from the new dictionary. However, the dictionary learning algorithm also captures changes in acoustic model output due to channel conditions. Both new dictionaries (Table 5.7) improve results in F2 and F4 (telephone speech and degraded acoustics).[17]

In summary, both derived dictionaries improved word error rate in almost all focus conditions, not just ones that were marked for casual speech. Most of the benefit of dictionary learning was from the first 100 hours of training; the additional training data used to train BN97+98 improved the recognition of non-primary focus conditions (F2-FX) slightly, and the primary conditions F0 and F1 not at all.

## 5.4 Do probabilities matter?

The best static dictionary produced in this line of research (in terms of joint error-rate and speed performance) was built by adding a pruned set of automatically generated pronunciations to the baseline dictionary. Pruning is an engineering-oriented operation on a statistical model; in essence, pruning a pronunciation means that one believes that it has zero probability of occurring. This runs counter to general statistical practices: for instance, in $n$-gram language modeling, much effort has gone into finding ways to estimate

[16] The word sequences provided by each system were not identical even though the word error rates were similar; for example, the "error rate" of BN97+98 when scored against the SPRACH98 results was 3.6%.

[17] The BN97+98 gain in F2 is not significant, while F4 is significant at $p < 0.05$.

| Dictionary | Standard probs | Probs=1 | Quantized | Quantized & removed infrequent |
|---|---|---|---|---|
| ABBOT96 (baseline) | 24.0 | 24.4 | 24.4 | 24.4 |
| SPRACH98 (BN97 training) | 23.4 | 27.2 | 24.0 | 24.1 |
| BN97+98 training | 23.2 | 27.0 | 23.5 | 23.6 |

Table 5.8: Effects of removing pronunciation probabilities on word error rates for Hub4E-97-subset (decoded with CHRONOS and A-Model II).

low-probability events using backoff schemes like Good-Turing discounting [Good, 1953; Katz, 1987].[18]

It is reasonable to ask whether the gains from the pruned static lexicon are mainly due to model selection, or whether the probabilities of the pronunciations play a role in determining performance. This hypothesis can be tested by setting all of the pronunciation probabilities for a word equally, and then evaluating the recognition performance. The result of this experiment is dependent on the decoder parameters; some decoders weight the pronunciation model more heavily by scaling probabilities by an exponential factor. In the NOWAY and CHRONOS decoders, the pronunciation model is given roughly the same weight as the language model.

The decoding strategy of the recognizer will also affect results. In a Viterbi decoder (one that finds the best path at every time point and eliminates all others), one should set all pronunciation probabilities to 1, since each pronunciation is effectively treated as a separate word by the decoder. In a full-forward probability decoder, path probabilities are summed across pronunciations at the end of a word, so it is more appropriate to set the priors of the $n$ pronunciations of each word to $\frac{1}{n}$ so that words with more baseforms do not get extra weight in decoding.

For a first experiment, I took the ABBOT96, SPRACH98, and BN97+98 dictionaries and stripped out the pronunciation priors, replacing them with a probability of 1 (since NOWAY and CHRONOS are both Viterbi-based decoders). I then recognized the Hub4E-97 test subset with all of the new dictionaries, decoding with CHRONOS; A-Model II provided the acoustic probabilities. The results (shown in Table 5.8: Probs=1 column) indicate that removing probabilities severely degrades performance in the new dictionaries while only slightly affecting performance of the ABBOT96 dictionary.[19]

If one examines the distribution of pronunciation probabilities in the dictionaries (Figure 5.8), the reason for the disparate dictionary results becomes clear. 80% of the pronunciation models in the ABBOT96 dictionary, corresponding to 91% of the words, have a probability of 1.0, *i.e.*, most words have a single pronunciations. However, many fewer pronunciations in the automatically derived dictionaries have such high probability — several

[18] In any case, the language model developer must decide which words to include in the model; thus, some pruning goes on in this domain as well.

[19] Similar results were seen with the NOWAY decoder. When A-Model I was used for decoding, a similar pattern was seen for the ABBOT96 and BN97+98 dictionaries; however, the SPRACH98 models improved from 27.2% word error to 26.7%. There does seem to be some interaction between quality of the acoustic models and this phenomenon, although it is difficult to say why this is so.

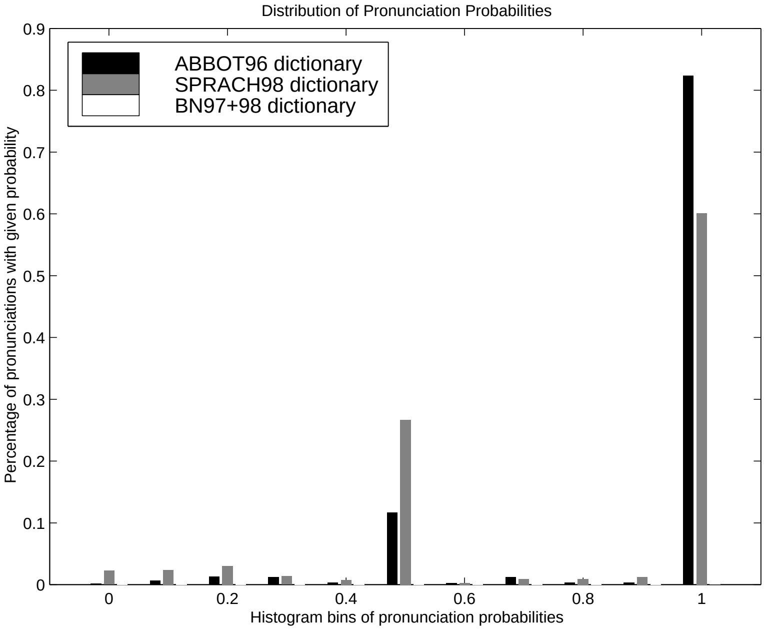


Figure 5.8: The histogram distribution of pronunciation probabilities in the three dictionaries.

other peaks in the distribution appear at much lower probabilities (0.5 for SPRACH98, and 0.25, 0.5, and 0.75 for BN97+98) due to the dictionary averaging scheme. Thus, it is not surprising that setting all probabilities to 1.0 had a disastrous effect on recognition results with these lexica.

One possible explanation for the large increase in error with the derived dictionaries is that greatly increasing the probability of very unlikely pronunciations devastated recognition performance. In order to reduce the influence of low-probability baseforms, I quantized the pronunciations in the following manner: all dictionary forms that had a probability of less than 0.1 in the original lexicon were set to a probability of 0.01. All other pronunciations received a weighting of 1.0. As the "Quantized" column in Table 5.8 shows, downgrading the influence of low probability baseforms almost completely compensated for the previous increase in error rate. Removing those infrequent pronunciations from the dictionary did not affect error rates greatly ("Quantized and removed infrequent" column). From these results, it appears that probabilities are important for determining the appropriate pronunciations in the very broad sense of model selection and to discourage use of infrequently seen pronunciations in the decoder, but the exact values of probabilities are not critical. These results therefore imply that carefully refining probabilities with more accurate statistical models will likely lead to small improvements in performance. This result is not conclusive — because of the scaling factor between the acoustic model and language model used in most recognizers, it is also possible that finer-graded probabilities are useful only within a certain range of scaling factors. It may also be the case that improvements in the pronunciation model will show up only after a large improvement in another part of the system (*e.g.*, in the acoustic model).

## 5.5 Confidence-based evaluation of novel word pronunciations

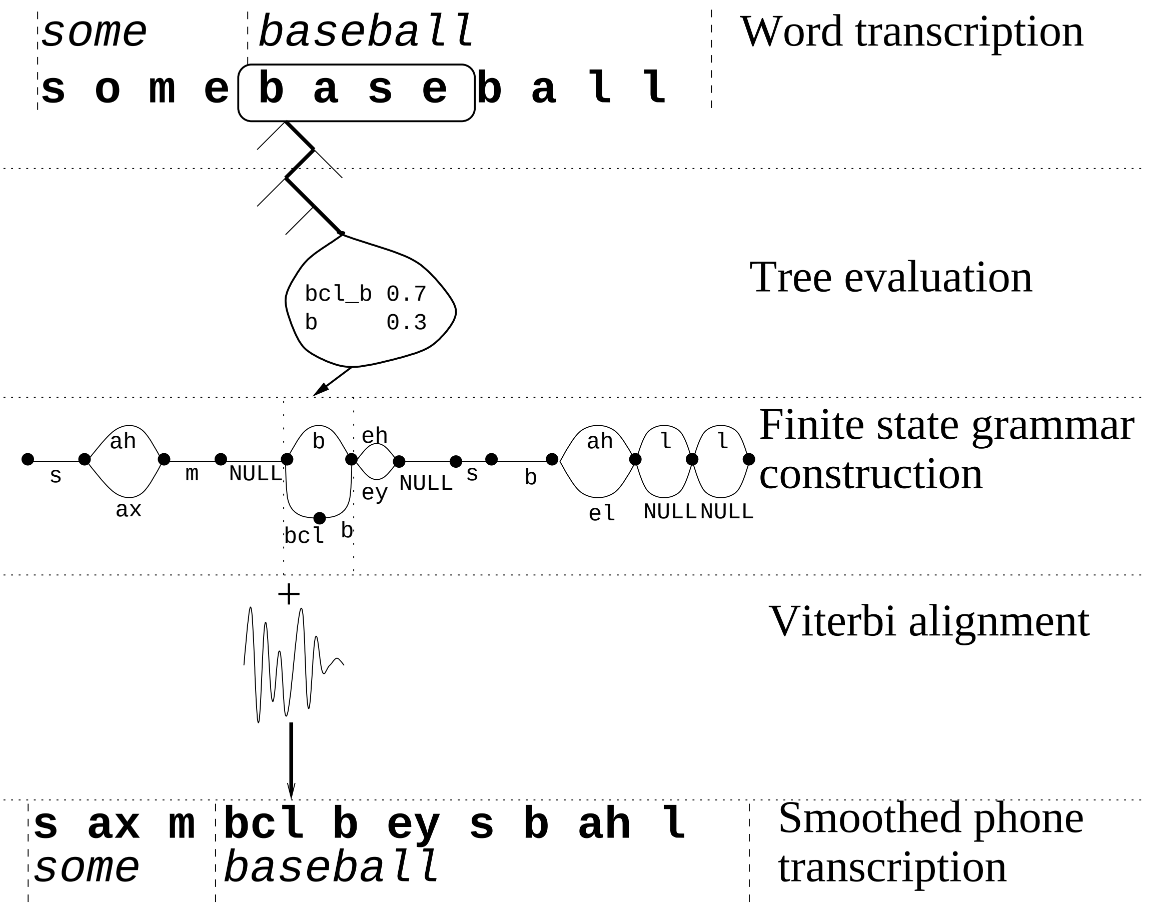


Figure 5.9: Constructing pronunciations from letter-to-phone trees.

Speech recognition systems are usually built to handle a bounded vocabulary (a typical limit is 65,000 words, so that word indices can be encoded as a 16-bit short integer). Thus, system designers must choose what vocabulary to represent for the task at hand, using criteria such as out-of-vocabulary rates on independent data representative of the target domain. In the Broadcast News domain, news stories come and go; as the system progresses from year to year, it must take new words into account. The 1998 SPRACH Broadcast News system, for instance, needed representations for 7,000 new words like *Lewinsky* and *Tamagotchi* due to new subjects in the news.

For the pronunciation modeler, new words in the vocabulary mean new baseforms are needed in the dictionary. The techniques discussed in the previous sections provide new baseforms only for words for which we already had a baseline representation: because of the mapping component of the smooth phone recognition models, the automatic learning system requires a forced-Viterbi alignment from the baseline dictionary. One could provide phonetic transcriptions for all of the words by hand, but this method produces results that are not self-consistent, or not consistent with the other pronunciations in the dictionary, due to the frailties of human transcribers. Also, in the SPRACH98 case, we had only a two-week period to create the 7,000 new pronunciations; modeling these words by hand in that short time period was infeasible. In this section, I discuss extensions to the smooth phone recognition algorithm that allow for generation of pronunciations for novel words in a short amount of time.

The problem of text-to-phone generation has been explored extensively in the literature. The approach most systems employ is to find a statistical mapping between a letter and a phoneme given the context of the letter,[20] a process very similar to the phone trees described in this chapter. Decision trees [Lucassen and Mercer, 1984; Jiang *et al.*, 1997; Ngan *et al.*, 1998; Kuhn *et al.*, 1998] and neural networks [Sejnowski and Rosenberg, 1987; Adamson and Damper, 1996] have been popular tools for this task, although Boltzmann machines [Deshmukh *et al.*, 1997], Hidden Markov Models [Parfitt and Sharman, 1991], and $n$-gram based rules [Fisher, 1999] have also been utilized. All of these systems are very similar in behavior and bear similarities to the model described in this section.

Lucassen and Mercer [1984] describe interfacing letter-to-phone (LTP) rules with a phone recognizer to determine new pronunciations for words. This provides two sets of constraints on pronunciations: the orthography and recorded exemplars of the words. Lucassen and Mercer claim that

> The two parts of the system complement each other. The phoneme recognizer generally gets vowels correct but has trouble with consonants while the spelling-to-baseform channel model generally gets consonants correct but has trouble with vowels. [page 42.5.3]

The algorithm developed here uses a similar technique. In the first phase of training, the letters in the baseline dictionary (ABBOT96) are aligned to phones using a Hidden Markov Model.[21] As a first guess, the HMM maps each letter in the dictionary to one phone. These mappings are re-estimated using the Expectation-Maximization (EM) algorithm [Dempster *et al.*, 1977] until the labels converge (*i.e.*, mappings do not change between one step of the algorithm and the next).[22]

Once the correspondences between letters and phones were determined, I trained decision trees on the mapped dictionary. I split the dictionary into a training set (90%) and

[20] This is in comparison with writing a large compendium of letter-to-sound rules by hand, as was done by Allen *et al.* [1987] in the MITalk (later DECtalk) system.

[21] Thanks to Tony Robinson and SoftSound Ltd. for providing the pronAlign HMM software.

[22] Descriptions of the EM algorithm and its applications to HMMs can be found in Rabiner [1989] and Jelinek [1997].

a test set (10%). For every letter of the alphabet, I trained a d-tree to predict the mapped phone given the three letters to the left and right (giving a total context window of seven letters). Each word was treated in isolation (*i.e.*, not in the context of other words); context windows that occurred on the boundaries of words were padded with spaces. For example, if the pattern in Figure 5.9 were used in training, the three left phones for the initial "b" would be spaces.

Once trees are grown, it is easy to see by analogy to the smoothed phone recognition how the letter-to-phone rules are employed (Figure 5.9). Given a set of letter-to-sound trees, it was then possible to construct a (bushy) pronunciation graph for each novel word. The best pronunciation for the word can be found by a simple lattice search, but having a recorded example can improve models quite a bit (as Lucassen and Mercer found). When such an example is available, the pronunciation graph is aligned to the waveform using the finite state grammar decoder.[23] The matching of the pronunciation graph to the acoustic models is the critical component of this technique; using a text-to-speech system that does not incorporate information from acoustic models would likely produce pronunciations with different properties than those in the baseline dictionary.

For the new pronunciations needed by the 1998 SPRACH system, I built letter-to-phone trees from the ABBOT96 dictionary. Recordings were made of all 7,000 novel words; these were aligned against finite state grammars constructed from the orthography of the new words. Using the recorded words, I also computed a normalized posterior confidence score ($nPP_w$) for each word. These confidence scores were critical for evaluation of the new pronunciations: most of the words were proper nouns (and a large subset of these were foreign words); these words have a higher variance in the relationship between spelling and pronunciation than the common nouns that dominated the LTP training set. The pronunciations produced by the procedure were far from perfect as a group; however, spot checks of the high-confidence novel baseforms showed them to be more reliable than the low-confidence ones.[24] Therefore, hand-correction efforts were focused on lower-confidence pronunciations.

With my hand-correction efforts, it appeared to me that most of the low-confidence (and poorly modeled) words were of foreign origin. Improvements to the model could be effected by including more proper names in the training dictionary (as is done by Ngan *et al.* [1998]), as well as a mixture model where the etymological origin of the word is identified as a feature (*e.g.*, *Tamagotchi* is Japanese, *Pagnozzi* would be marked Italian, and *Ovsyannikov* could have a Slavic marker). Identifying language source, however, could be as difficult as the pronunciation generation problem itself.

Since the goal of this small project was to quickly produce a large number of new pronunciations for the SPRACH system, I did not evaluate this paradigm in any of the standard ways (*e.g.*, testing on some held-out portion of a known dictionary). Rather, the point here is that the stream-transformation paradigm is flexible and modular enough that implementation of a letter-to-phone system required only a small amount of effort. While the system does not produce perfect results, confidence scores produced by alignment against

[23] The baseline pronunciations were used in the graph-building procedure for words already in the dictionary.

[24] The confidence measures also isolated occurrences of misrecordings, improving quality control.

recorded acoustics greatly reduced dictionary development time.

## 5.6 Summary

This chapter explained the design of static dictionaries for use in our first-pass recognizers, NOWAY and CHRONOS. The work presented here represents an improved baseline for comparison against dynamic modeling techniques in the next chapter. A second goal of the studies presented here was to extend the dictionary from the ABBOT96 Broadcast News system for use in the 1998 DARPA Broadcast News evaluation.

The initial experiments presented in this chapter extended the ABBOT96 baseline dictionary by allowing a phone recognizer to suggest new pronunciations. Replacing the baseline pronunciations with these new models increased errors significantly — a 9 to 17% relative increase depending on how many training examples of a new pronunciation were necessary before inclusion in the dictionary. Merging new phone recognition pronunciations with the original ABBOT96 dictionary helped to decrease the influence of poorly modeled pronunciations, improving recognition by about 2% relative over the baseline. The added constraints of dictionary smoothing are important for suppressing spurious pronunciations produced by the automatic learning technique.

Pronunciation dictionaries generated by phone recognition had the drawback of vastly increasing decoding time, which created problems for the SPRACH project's desire to compete in the DARPA evaluation, which stipulated a 10× real-time condition. Smoothed phone recognition techniques provided more constraints on new pronunciations by using the statistics of the entire corpus to guide transcription. This technique improved word error results slightly over the phone recognition dictionary, while recognition was almost 50% faster than when using the phone recognition dictionary. Further increases in speed were achieved using a log-based pruning algorithm (265% faster than the phone recognition dictionary, 146% faster than the unpruned smooth phone recognition dictionary) without sacrificing the error rate improvements of the unpruned dictionary. Appropriate pronunciation selection, by means of the log count pruning algorithm, was the key to achieving this result.

It is important, therefore, not only to generate a set of new pronunciations, but also to carefully select models to maximize both speed and recognition performance. Confidence measures can play an important part in this selection process by improving model selection, as the experiments with pronunciation confidence score pruning showed. Acoustic confidence scores also facilitated validation of pronunciations for novel words generated by letter-to-phone trees. The flexibility of the pronunciation learning paradigm allowed generation of 7,000 new baseforms in about two weeks; confidence scores directed hand-tuning efforts, greatly reducing the amount of time needed to produce the new pronunciations.

The new static dictionaries were derived using a particular set of acoustic models; it seemed possible that the gains developed on the subset of the 1997 Hub-4 evaluation test set would not carry over to other decoding conditions. Fortunately, it appears that the automatically learned static dictionaries are robust to changes in acoustic models, decoders, and test sets; if this were not the case, then the lexicon would have to be retrained with every new acoustic model.

Unfortunately, experiments with doubling the amount of training data did not prove fruitful. The modest improvements may result from the fact that the actual values of the pronunciation probabilities mean very little; in the studies conducted here, probabilities on pronunciation models appear to be most important for model selection. As long as low-probability baseforms are not over-emphasized, setting all probabilities equal harms recognition only slightly.

In sum, the two critical components for building a new dictionary automatically are a method for constraining suggestions of pronunciations from the acoustic models, and a selection criterion for choosing the appropriate models. Here, we used phonetic context to provide constraints on phone recognition; probability and confidence measures were used for selection criteria. In the next chapter, the concept of context is expanded to include features describing words, syllables, speaking rate, and duration, and the selection criteria will become *dynamic*, allowing pronunciation models to change during recognition dependent on these variables.

# Chapter 6

# Dynamic Dictionaries

## 6.1 Introduction

In the previous chapter, we saw that using context to constrain pronunciation variations suggested automatically by the speech recognition system allowed for better static dictionaries. The primary purpose of these contextual models was to facilitate baseform selection. In this chapter, I hypothesize that if the choice of appropriate pronunciations is made *dynamically* (*i.e.*, the model is allowed to change with context, even cross-word), then recognition performance may improve.

The definition of context is also expanded in this chapter. In the previous chapter, I used information about neighboring phones to determine smoothed phone recognition mappings. However, the contextual influence on phonetic realizations reaches beyond the surrounding phones. In Chapter 3, I demonstrated that pronunciations in the Switchboard corpus depend heavily on other factors in addition to phonetic context. In particular, the frequency of a word influences the extent to which reduction processes are correlated with speaking rate: more frequent words have more variation at high rates of speech. Syllabic structure also plays an important part in determining which phones are more likely to vary; coda consonants are much more likely to be non-canonical than onset consonants.

The influence of syllables and words in these studies suggest that an orientation toward larger linguistic units (as opposed to phones) may prove beneficial in pronunciation modeling. This is easy to implement in our paradigm; instead of d-trees modeling one phone each, they can model one syllable or one word each. One can integrate different forms of context into syllable or word trees than in phone trees. For example, because phones can occur in so many different words, in phone trees the identity of a neighboring word (as opposed to a neighboring phone) will probably not have much effect on the pronunciation of the phone overall, whereas the co-occurrence of two words may have a profound influence on the pronunciation.

In this chapter, I will examine ways of incorporating long-range context into prediction of changes in the pronunciation of syllables and words. Section 6.2 describes how decision trees can be used to model words and syllables and how these d-tree models are employed to rescore potential word sequences postulated by the recognizer. The following section introduces the phone-based rescoring model developed at the 1996 Johns Hopkins Summer Research Workshop, describing results on the Switchboard corpus. In Section 6.4, I present descriptions of the additional context features used to model pronunciations on the Broadcast News corpus. Several methods of model evaluation are described in the subsequent section, followed by a summary of experiments using dynamic pronunciation models.

## 6.2 Rescoring with dynamic dictionaries

In Chapter 2, I laid out the mathematical motivation for a dynamic pronunciation model based on stream transformations. In a static model, $P_P(B|M)$, the baseform pronunciations for word $i$, are fixed — *i.e.*, they depend only on the distributions $P_P(b_i|m_i)$ (Equation 2.13). The dynamic pronunciation model $P_{DP}(Q|B, M, R, LM, \ldots)$ allows the baseform pronunciations to vary based on more information than just the word being modeled, in this case word context ($M$), speaking rate ($R$), language model probabilities ($LM$), and possibly other features. This section describes how these features are incorporated into the d-tree models of Chapter 5, and how these models are employed during recognition.

### 6.2.1 Incorporation of extra features into dynamic trees

Changing the model $P_{DP}(Q|B, M, \ldots)$ to have a dependence on additional factors requires including these factors in the training data of the decision trees. For categorical features such as the identity of neighboring words, one can just add extra attributes to the input data, as described in Section 5.3.1. Real-valued attributes can be pre-quantized (*e.g.*, speaking rate can be quantized into slow, medium, and fast); however, determining the dividing points between these categories *a priori* may not reveal the best possible divisions for capturing pronunciation variation.

The IND decision tree package [Buntine, 1992] has the ability to find the optimal dividing point of real-valued attributes automatically. The algorithm is a simple extension of the categorical partitioning scheme. Training samples are sorted by the attribute value $f_j$, giving a sequence of examples $(e_1 \ldots e_n)$; for every neighboring pair $(e_i,\ e_{i+1})$, a cutoff $c_i$ is defined to give the partitioning question

$$Q_{j,c_i} = (f_j(e) < c_i), c_i = \frac{f_j(e_i) + f_j(e_{i+1})}{2}, \tag{6.1}$$

similar to the categorical question $Q_{j,S}$ in Section 5.3.1. These new questions can be employed in the decision tree search algorithm (Equation 5.2). In practice, the space of possible cutoffs is subsampled in order to reduce the number of possible questions.

### 6.2.2 $N$-best rescoring as smoothed phone recognition

Once decision trees have been constructed, employing them in a rescoring paradigm is very similar to the smooth phone recognition process. A second recognition pass for the dynamic rescoring of hypotheses is necessary because in a dynamic pronunciation model, the pronunciation of each word or syllable depends on both previous *and* subsequent words and baseforms. In the first pass of decoding, it is expensive to evaluate the pronunciation of a word based on some feature of the following word, due to the time-synchronous nature of sentence processing in many decoders — earlier words in the hypothesis (word sequence) are fixed before later words. One *can* re-evaluate previously decoded words (*e.g.*, the penultimate word in a current hypothesis) online with a dynamic pronunciation model, although this can cause a large increase in processing time due to a significant augmentation of the search space.

To get around this limitation, one can compute a list of the $n$ top hypotheses suggested by the recognizer, and then re-rank them according to some criterion. This permits the dynamic evaluation of a word's pronunciation within the context of both the previous and subsequent words. In the system used at ICSI (Figure 6.1), the NOWAY first-pass stack decoder [Renals and Hochberg, 1995a] outputs a hypothesis *lattice*. This lattice is a word graph where vertices of the graph correspond to particular time points; edges in the graph correspond to a word that starts at the time of the first vertex and ends at the second vertex. Each edge is also annotated with the word's acoustic likelihood, the pronunciation variant used, and the pronunciation model prior. This graph can be converted into an $n$-best list by a separate decoder (not shown) that takes the language model into account.

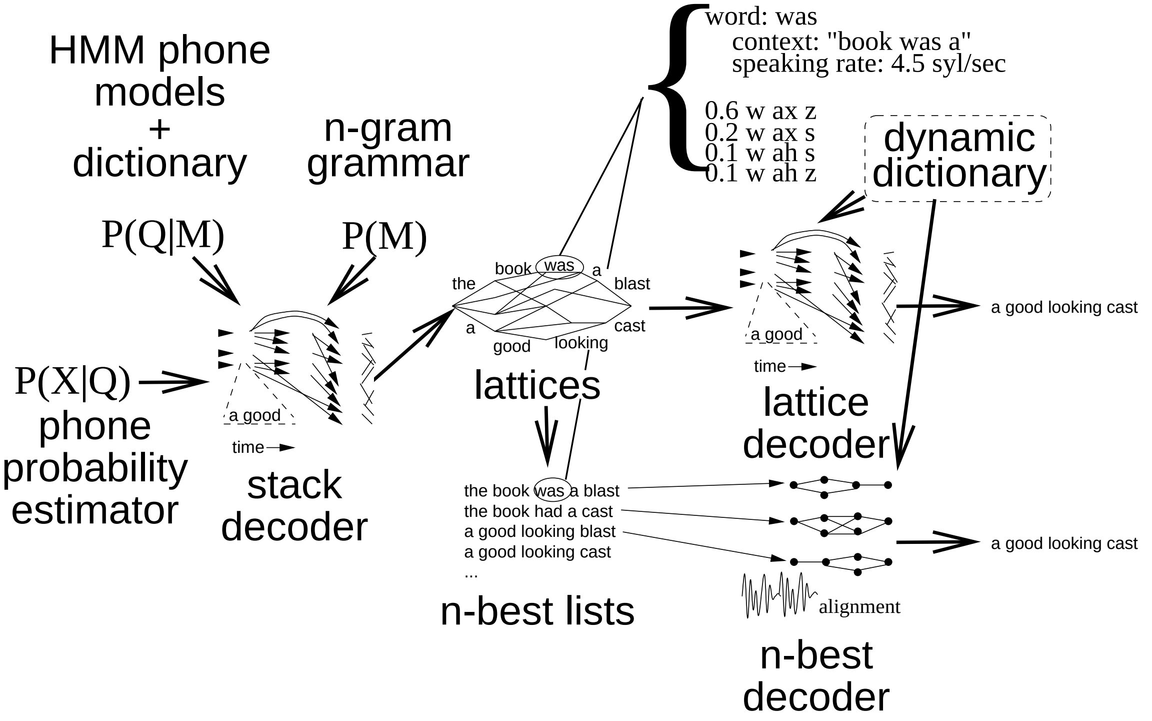


Figure 6.1: The dynamic pronunciation rescoring paradigm.

This $n$-best rescoring process has the advantage that the smoothed phone recognition software (Section 5.3.2) can be reutilized for this task. Every hypothesis $h$ suggested by the recognizer has an acoustic score $A(h) = \log P_{A+P}(X|h)$ (the probability of acoustics given the hypothesis) associated with it, as well as a language model score $L(h) = \log P_L(h)$.[1] In rescoring with the d-trees, the hypothesis is turned into a finite state grammar and aligned against the test acoustics. The acoustic model score from this realignment, $A_{\mathrm{dyn}}(h)$, is retained and combined with the language model score to give a new probability for the hypothesis:

$$\log P(h, X) = A_{\mathrm{dyn}}(h) + L(h). \tag{6.2}$$

For smoothing purposes, one can also interpolate old and new acoustic model scores, using

$$\log P(h, X) = \lambda A_{\mathrm{dyn}}(h) + (1 - \lambda)A(h) + L(h). \tag{6.3}$$

With these new estimates of utterance probabilities, the hypothesis satisfying $\underset{h}{\mathrm{argmax}} P(h, X)$ can be returned as the best guess of the decoder.

### 6.2.3 Extending to lattice decoding

Rescoring $n$-best lists also has a disadvantage: hypotheses within the list often differ by only a few words, so for very long utterances (of 100 words or more) rescoring $n$-best lists will possibly make very little difference in the final recognition error rate. Furthermore, in the first-pass search, early choices in the hypothesis search can affect which words are selected later in the word sequence through the influence of the language model. If a new pronunciation model selects a different set of initial words, it would be advantageous to incorporate this fact into the hypothesis search.

A middle ground between $n$-best decoding and full first-pass decoding is to operate directly on the lattices. This still requires the decoding structure found in a first-pass recognizer, but limits the search space by allowing searching only along paths in the lattice. The lattice, in essence, becomes a finite state grammar that guides the search,[2] providing additional constraints to the $n$-gram grammar.

In order to understand how a dynamic pronunciation model can be used in a lattice decoder, it is instructive to highlight the differences between a normal decoder and a dynamic pronunciation lattice decoder. The basic search strategy for NOWAY is shown in Figure 6.2; I have stripped away most of the bells and whistles that make NOWAY efficient for first-pass decoding, such as the least-upper-bound calculation for the A* search [Renals and Hochberg, 1995b] and the tree-based acoustic score calculations.[3] NOWAY is a time-

[1] See Section 2.1, page 11 for a description of the probabilistic models used in ASR.

[2] The recognition system at AT&T [Mohri *et al.*, 1998], in fact, treats lattices and $n$-gram grammars as FSGs; the intersection of these grammars provides a weighted lattice, where $n$-gram probabilities are imposed on the lattice structure. Pronunciation models and acoustic models can also be viewed as FSGs, so that one can completely decode utterances via the operations of FSG intersection and a best-path search through a weighted FSG.

[3] Using the NOWAY libraries, I created a lattice decoder with this algorithmic structure (called NOHOW) as a first step toward building the dynamic pronunciation lattice decoder.

```
• Initialize stack $S_0$ with the null hypothesis ($h = ()$)
• For time t = 0 to n-1
    – While $S_t$ is not empty
        * Pop hypothesis $h$ from $S_t$
        * For all words $w$ s.t. start($w$) = $t$
            · Let $h_{new}$=$h + w$
            · Set score($h_{new}$) = score($h$) + AcousticScore($w$) + LanguageModel($w|h$)
            · Push $h_{new}$ onto stack $S_{\text{end}(w)}$
• While $S_n$ is not empty
    – Pop hypothesis $h$ from $S_n$
    – Let $h_{new}$=$h$ + end_of_utterance
    – Set score($h_{new}$) = score($h$) + LanguageModel(end_of_utterance|$h$)
    – Push $h_{new}$ onto stack $S_{final}$
• Pop best hypothesis from stack $S_{final}$
```

Figure 6.2: The NOWAY stack management algorithm for processing utterances from time 1 to $n$ [Renals and Hochberg, 1995a].

```
• Initialize stack $S_0$ with the null hypothesis ($h = ()$)
• For time t = 0 to n-1
    – While $S_t$ is not empty
        ∗ Pop hypothesis $h = (h_1, h_2, \ldots, h_m)$ from $S_t$
        ∗ For all words $w$ s.t. start($w$) = $t$
            · Let $h_{new}$=$h + w$
            · $\mathbf{Let\ h_{old}} = (\mathbf{h_1 \ldots h_{m-1}})$
            · Set score($h_{new}$) = score($h_{old}$) +
              $\mathbf{DynamicScore(h_m|h_{m-1}, w)} + \mathbf{LanguageModel(h_m|h_{old})}$ +
              AcousticScore($w$) + LanguageModel($w|h$)
            · Push $h_{new}$ onto stack $S_{\mathsf{end}(w)}$
• While $S_n$ is not empty
    – Pop hypothesis $h = (h_1, h_2, \ldots, h_m)$ from $S_n$
    – Let $h_{new}$=$h$ + end_of_utterance
    – $\mathbf{Let\ h_{old}} = (\mathbf{h_1 \ldots h_{m-1}})$
    – Set score($h_{new}$) = score($h_{old}$) +
      $\mathbf{DynamicScore(h_m|h_{m-1}, w)} + \mathbf{LanguageModel(h_m|h_{old})}$ +
      LanguageModel(end_of_utterance|$h$)
    – Push $h_{new}$ onto stack $S_{final}$
• Pop best hypothesis from stack $S_{final}$
```

Figure 6.3: The JOSÉ stack management algorithm. Differences from the NOWAY algorithm that allow for dynamic rescoring of the penultimate word in hypotheses are highlighted in boldface.

synchronous stack-based decoder; in plain English, this means that partial hypotheses (time-aligned strings of words with probabilities) that end at the (discrete) time $t$ are kept in a stack[4] associated with time $t$. Each stack $S_t$ is ordered by the hypothesis path probability $P(h_1^t, X_1^t)$, corresponding to the acoustic and language model scores of the partial hypothesis up to time $t$. To start decoding, the first stack ($S_0$) is initialized with a hypothesis containing no words and probability 1. Then, for every time step $t$, the hypotheses in $S_t$ are extended by single words starting at time $t$. In lattice decoding, this corresponds to extending hypotheses with all of the words in the lattice that start at time $t$.[5] Each extension is filed into a stack corresponding to the end time of the new partial hypothesis; the new score for the extension is the sum of the score for the unextended hypothesis, the acoustic score for the extended word (found in the lattice), and the language model score for the word given the previous hypothesis.[6] When the end of the utterance is reached (time $n$), a separate check for end of utterance probability is done, and the best hypothesis after this check is output.

The critical difference in the implementation of the dynamic pronunciation model is how the acoustic scores for words are calculated. Figure 6.3 shows the search algorithm internal to JOSÉ,[7] an acoustic-rescoring lattice decoder created for this purpose. Hypotheses are created by the same extension procedure as in NOWAY: when a new hypothesis is created, each word $w$ is added in to the hypothesis with the pre-computed acoustic score from the lattice. The *penultimate* word is then rescored with the new pronunciation model, using the neighboring words (and other features) to determine new pronunciations. The new dynamically derived score replaces the original acoustic score provided by the lattice.

The rescoring procedure does introduce possible inaccuracies into the search, particularly when the search is pruned. There are two separate pruning methods for NOWAY-style search algorithms. The primary strategy is to limit the stack depth, so that only a certain number of hypotheses can end at any particular time. Any new hypotheses that score worse than the top $n$ in the stack are not inserted. In addition, a beam-width can be used to prune any extensions that are worse than the best hypothesis by some factor. Since these pruning cutoffs are implemented in JOSÉ using the lattice acoustic scores from NOWAY, rather than the dynamically derived scores, it is possible that the best dynamic-dictionary hypothesis is pruned too early in the search. In practice, though, this is not a significant problem as long as the search beam-width is large enough.

In both the lattice and $n$-best rescoring paradigms, averaging the rescoring model with the original lattice acoustic score in a multistream-like approach improved results. Because of the large amount of time required to run experiments, I did not tune the combination parameter until the final set of experiments; all of the initial runs weighted each acoustic score evenly.

[4] In actuality, all that is required is some sort of ordered list of hypotheses; the data structure need not be a stack. For efficiency, NOWAY uses a priority queue with hypotheses in reverse order of score, so that the worst hypothesis can be easily deleted from the queue when the queue reaches maximum size.

[5] This is a bit of an oversimplification; it is possible that more than one lattice node may have the same start time. In this case, all one has to do is annotate each partial hypothesis with its corresponding ending lattice node. NOWAY, however, only produces lattices in which nodes have unique associated times.

[6] These scores are log probabilities, which justifies the addition (rather than multiplication) of scores.

[7] So named because NOWAY produces lattices for it in the first pass.

### 6.2.4 Modeling larger linguistic units

In the smoothed phone recognition experiments of the previous chapter, pronunciations were modeled on a phone-by-phone basis. Each baseform phone was associated with a decision tree that predicted how the phone would be realized in context. The decision tree leaf corresponding to the context was then compiled into a small piece of a finite state grammar, which was then concatenated with other phone grammar fragments into an FSG for the entire utterance, as shown in the phone tree illustration of Figure 6.4.

One problem with this technique of straight concatenation is that model decisions are independent. In the figure, the "ball" portion of *baseball* has four possible pronunciations, yet the pronunciation `[b el l]` is not likely at all; the system should be able to learn that when `[ah]` is realized as `[el]`, the subsequent `[l]` is very likely to be deleted. Riley's [1991] solution is to include a dependence on the previous decision tree output. This makes FSG generation more difficult, although no more complicated than introducing context-dependency into a pronunciation graph [Riley *et al.*, 1997]. Riley found that including the left context gave a substantial improvement in the predictive power of the trees; this information was used subsequently by Riley *et al.* [1998] to build phone trees for the Switchboard corpus.

Weintraub *et al.* [1997] used a different approach to handle context issues. Their technique added $n$-gram constraints on the phones using a maximum entropy model. The probability distributions provided by the d-trees were modified to penalize pronunciation sequences that were determined unlikely by an $n$-gram phone grammar. This extra information degraded recognizer performance significantly in initial experiments, though this research took place within the context of a six-week workshop and the authors were hopeful that further research along these lines might prove fruitful. To date, no further experiments have been conducted along these lines, so whether $n$-gram constraints on phones can improve modeling is an open issue.

The strategy I adopted toward solving this problem was to model the distributions of phone pronunciations jointly, at the syllable and word levels. This captures many of the coordinated phone pronunciation variations not handled by the independent phone trees. Since phones at segment boundaries still vary with context, pronunciations in these models include dependencies on the neighboring baseform phone. As suggested in the introduction to this chapter, other forms of context, such as word identity, can also be included in the model.

#### Word models

The method for determining word pronunciations dynamically is a cross between automatic baseform learning techniques and the d-tree statistical learning used for phone models. The training data, provided by (smoothed) phone recognition, is pooled by word to give a possible set of pronunciations. One tree is constructed for each word, with the word pronunciation predicted from context.

During recognition, the FSG construction algorithm (Figure 6.4: word trees) generates a grammar for each word by instituting a separate path for each pronunciation and

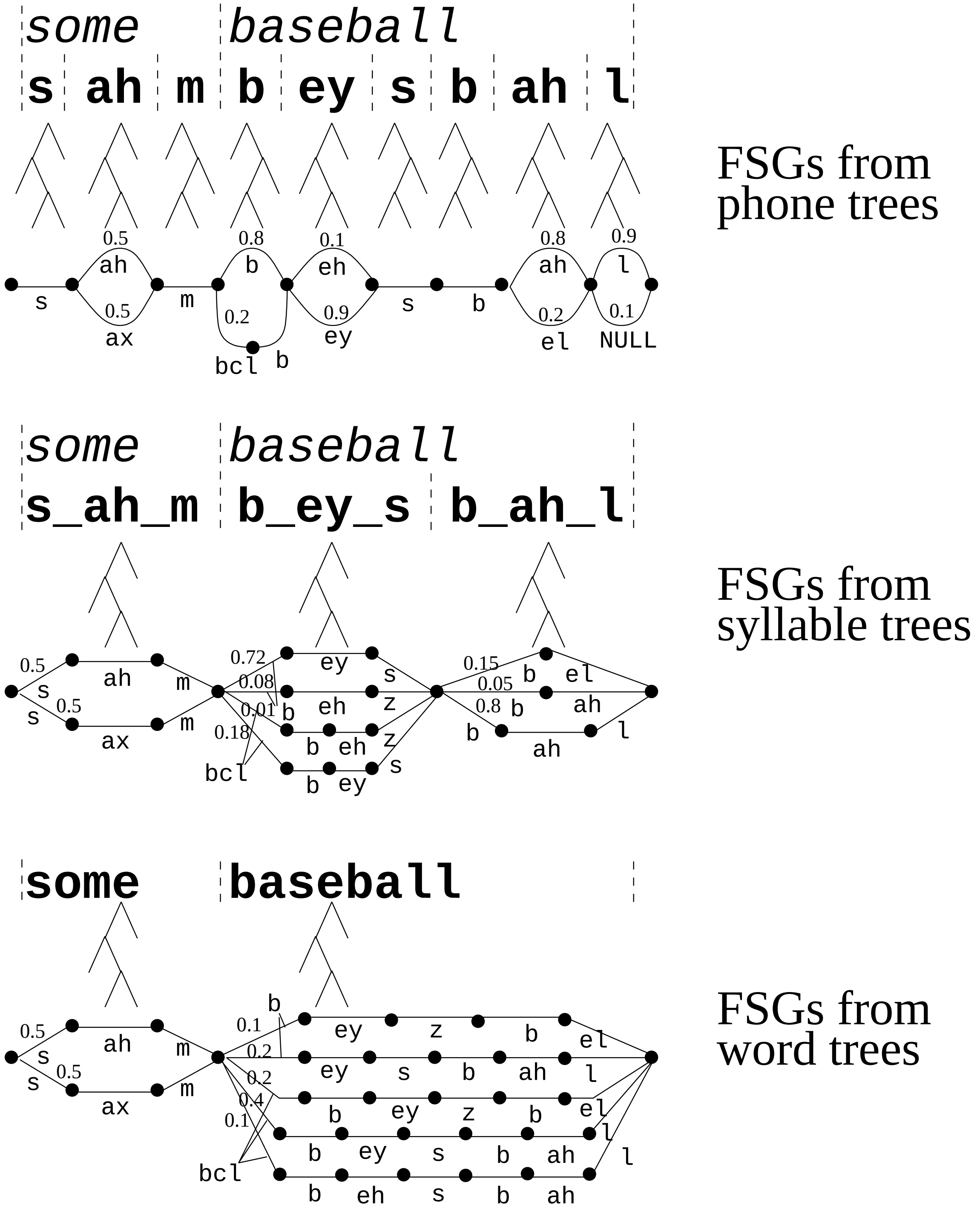


Figure 6.4: Example expansion of "some baseball" into finite state grammars with phone, syllable, and word d-trees. Arcs unlabeled for probabilities have probability 1.

tying the paths together at a start and end state; probabilities assigned by the trees are added to arcs leaving the initial state for the word. In some instances (such as *some* in the figure), the individual path probabilities are the same as in the phone trees. On the other hand, in *baseball* path likelihoods differ greatly from the phone tree FSGs, due to the phone coordination effects described above. Word trees can also allow different pronunciation alternatives than do the phone trees, as exemplified by the `[s]`/`[z]` substitution in *baseball*; these patterns may be word-specific or variations due to the statistical nature of the learning algorithm.

### Syllable models

Building separate decision tree models for each word does have the drawback that only words with enough training data can be modeled, whereas for phone trees one can model every phone in the corpus. A way to increase coverage is to use syllable models, so that words like *baseball* and *football* could share pronunciation models for their shared syllable. This can be implemented as a straightforward adaptation of the word-tree modeling, although there is no canonical way to choose the appropriate syllable models for each word. Ideally, one would like to model multiple pronunciations within the syllable model, leaving the choice of syllable model dependent on only the word. For instance, the word *some* has two pronunciations in the baseline dictionary, `[s ah m]` and `[s ax m]`, but it would not be beneficial to model this word with two separate syllable models — the variation between these alternatives should be provided by one syllable model. I developed the following algorithm to determine a single syllabic transcription for each word from a baseline dictionary:

1. Map "stop-closure stop" sequences to the stop consonant (*e.g.*, `[tcl t]`⇒`[t]`).
2. Syllabify all pronunciations of the word.
3. Find the longest pronunciation (in syllables) for the word.
4. Align all other pronunciations to the longest baseform, providing a list of phone alternations.
5. Remove all unstressed vowel alternatives if a stressed vowel exists.
6. Remove all NULL alignment alternatives.
7. Bind all remaining alternations into a "superphone" (*e.g.*, a phone that is pronounced either as `[t]` or `[dx]` becomes `[t/dx]`).
8. Link syllable-internal phones to give the syllable model name.

An illustration of this algorithm operating on the word *automatically* is shown in Figure 6.5. This example is particularly interesting because *automatically* can be pronounced with five or six syllables. Step 3 in the algorithm, by choosing the longest baseform, assures that this type of alternation is modeled as a shortening process.

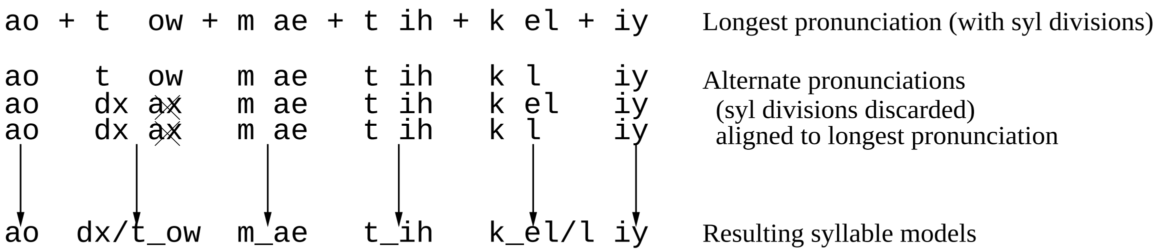

ao + t ow + m ae + t ih + k el + iy
Longest pronunciation (with syl divisions)
ao t ow m ae t ih k l iy
ao dx ax m ae t ih k el iy
ao dx ax m ae t ih k l iy
Alternate pronunciations (syl divisions discarded) aligned to longest pronunciation
ao dx/t_ow m_ae t_ih k_el/l iy
Resulting syllable models

Figure 6.5: Selection of syllable models for the word *automatically.* Reduced phones (**[ax]**) are eliminated if unreduced variants exist; similar phones are clustered into superphones.

## 6.3 The WS96 dynamic model

At the Johns Hopkins 1996 Summer Research Workshop, I participated in the Automatic Learning of Word Pronunciation from Data group, headed by Mitch Weintraub of SRI. At WS96, we built a dynamic pronunciation decoder for recognition of the Switchboard corpus. The dynamic models were used both to generate static dictionaries (as in the last chapter) and to dynamically rescore $n$-best lists of hypotheses. In this section, I will briefly summarize some of the results from our work [Weintraub *et al.*, 1997].

Researchers at the workshop trained the HTK (Hidden Markov Model Toolkit) Recognizer [Young *et al.*, 1995], available from Entropic, on Switchboard data as the baseline recognizer for the workshop. The system had multiple mixtures of Gaussians estimating triphone densities and was trained on 60 hours of training data that were judged (via automatic alignment) to be acoustically "good."[8] The baseline, using a trigram grammar, achieved a 46.0% word error rate on the entire WS96 development test set.

Starting from this point, our group utilized the baseline HTK system as a phone recognizer, automatically transcribing 10 of the 30 hours of training data. Using the recognized phones as our surface forms, we then aligned the output to our baseform pronunciations, as outlined in Section 2.4.2. This provided us with a new set of dictionary pronunciations, notated in Table 6.1 as "HTK PhoneRec." We only allowed new pronunciations if they occurred more than seven times in the training data. Replacing the HTK pronunciation dictionary with the new PhoneRec dictionary, the 100 top hypotheses from the baseline recognizer were rescored on a 200-sentence subset of the development test, for which the baseline performance was 46.4% word error. The new dictionary improved the word error rate by about 2% relative. This is an interesting contrast to the Broadcast News phone recognition results of the previous chapter, where replacing the dictionary resulted in a 9% relative *increase* in word error. The primary difference in this experiment is that the dictionary was used in a rescoring paradigm where the $n$-best lists were chosen with the original dictionary. Indeed, when the BN phone recognition dictionaries were smoothed with the original dictionary, an improvement resulted that was on the same order as that

[8] The acoustic likelihood of the training data determined which data were used in the training set; data which fell below a cutoff likelihood were excluded from training.

| Experiment | Word error rate<br>Rescoring 100 best hyps |
|---|---|
| Baseline HTK | 46.4% |
| HTK PhoneRec mincount=7 | 45.5% |

Table 6.1: Results of rescoring of 100-best hypotheses by phone recognition dictionary for a 200 sentence subset of 1996 Switchboard Development Test Set.

| Experiment | Word Error Rate<br>Rescoring 75-100 Best Hyps |
|---|---|
| Baseline HTK | 46.4% |
| Modified HTK: no-sp | 45.7% |
| DT1-Pron | 45.2% |
| DT2-Graphs | 45.5% |

Table 6.2: Results of rescoring of 75- to 100-best hypotheses by decision tree models for the 200 sentence Switchboard subset.

of the WS96 system.

From the alignments between the baseline dictionary and the HTK phone recognition we also trained a first set of phone decision trees, called DT1. The d-tree construction procedure was almost identical to that of Section 5.3.1; there were only three differences in the training data. First, only binary questions about the baseform phones were used, instead of the $n$-ary questions derived from [Riley, 1991]. Thus, instead of asking "What is the consonant place of the next phone?" the tree-construction algorithm asked the more specific question "Is the next phone an alveolar consonant?" In addition, questions about the lexical stress of the syllable were allowed, in part because the baseline lexicon (Pronlex [Linguistic Data Consortium (LDC), 1996]) marked vowel nuclei with stress marks.[9] All in all, 140 binary questions were used (compared to the 7×3 $n$-ary questions used in the BN trees of the previous chapter).

In the middle of the experiments at the workshop, we discovered that the short pause ([sp]) phone, which is appended to every word and is intended to capture short silences, was actually sometimes modeling longer non-silence sequences (up to 400 ms). We decided to delete the sp phone and explicitly model silence via an extra "pause" word, which improved recognition performance by 0.7%. All subsequent recognition experiments were thus conducted using this "no-sp" model.

Using the DT1 trees, we generated a new set of static (per-word) pronunciations, which we used to replace the HTK dictionary (Figure 6.2: DT1-Pron). We also used the DT1 trees to realign the training set and perform an additional iteration of d-tree training (DT2). These decision trees were used to generate pronunciation graphs (DT2-Graphs) for each hypothesis string, thus (we hoped) capturing some of the cross-word regularities in pronunciation.

[9] This information was not encoded in the ABBOT96 dictionary for the BN experiments.

The explicit static dictionary replacement (DT1-Pron) worked slightly better than using pronunciation graphs to dynamically rescore the $n$-best lists (DT2-Graphs), although the difference was not significant. We did not retrain the acoustic models at the workshop, although this should have given us some improvement. Murat Saraclar, another student at the workshop, did retrain the acoustic models after the workshop by adapting only the triphone building stage of the acoustic models [Saraclar, 1997]. The new acoustic model led to a slight increase in errors (46.6%); it is not clear whether full acoustic retraining will help, but this should be investigated.

## 6.4 Broadcast News model design

The rest of this chapter focuses on efforts to extend the WS96 model. Segmental context, while important, is certainly not the only factor that affects pronunciations; this fact engendered the studies in Chapters 3 and 4 that suggested that including speaking rate and word predictability into pronunciation estimation can help to build better models.

As I indicated in the previous chapter, the effort of building the SPRACH Broadcast News recognizer at ICSI coincided with my research. An interesting question presented itself: can a recognition system with half the error rate of the Switchboard system still benefit from improved pronunciation modeling, or were the gains from the WS96 pronunciation model just compensating for poorer acoustic models on a harder task?[10] Broadcast News, with its test sets divided into separate focus conditions, could also shed light on the differences between modeling for spontaneous versus planned speech. With these factors in mind, I decided to build dynamic syllable and word pronunciation models for the SPRACH Broadcast News system.

To build models for larger linguistic units, I added a number of features to the decision trees in hopes of improving modeling. In the remainder of this section, I discuss the contextual features used in d-tree construction, including linguistic segment context, estimates of speaking rate, estimates of word predictability, and other features.

### 6.4.1 Linguistic context

Most of the phonetic features used to construct phone-based d-trees (Section 5.3.1) were employed in building syllable and word d-trees. Features representing the extended context of syllables and words were also added to the algorithm, as discussed in this section.

#### Phonetic features

As described previously, decision trees were allowed to select for the consonant manner, consonant place, vowel manner, vowel place, and phone identity of the last phone of the previous syllable or word (depending on the model type), or the first phone of the next syllable or word. Syllabic positions (onset, nucleus, coda) were also included,

[10] This is not to suggest that the modeling technique in the WS96 HTK system is sub-standard, rather, that all systems seem to perform more poorly in the Switchboard domain compared to Broadcast News.

although boundary markings were not included because of redundancy — for word models, the previous/next phone always occurred at a boundary, while for syllable models, this was encoded by a position feature, described below.

### Syllable and word features

An obvious syllable-tree correlate to the phone identity feature is the identity of the neighboring syllables. I also allowed the algorithm to query the position of the syllable in the word (either *initial, final, initial-and-final,* or *not-initial-or-final*), as well as the syllable count of the word. Since lexical stress was an important feature in the WS96 model, I attempted to mark all syllables with the markers *stressed, secondary-stressed,* or *unstressed* by mapping the syllabic patterns in the ABBOT96 dictionary to the Pronlex and CMU dictionaries.[11] For words not found in either dictionary, syllables with reduced vowels (*e.g.*, schwa) were marked as *unstressed.* If there was only one syllable left unmarked in the word, then it received the *stressed* designation, otherwise all remaining syllables were marked with *unknown* stress.

Syllable and word trees were also allowed to depend on the identity of the neighboring word. This feature bears an interesting relationship to the multi-word work of Finke and Waibel [1997b], in which the authors find that the mutual information criterion is appropriate for determining which words to pair for learning pronunciations. The decision tree algorithm similarly uses a mutual information criterion for determining appropriate feature selection, so including the word identity as a feature in d-trees is related to the Finke and Waibel multiword algorithm.

## 6.4.2 Speaking rate estimates

As described in Section 3.1.1, many possible algorithms exist for estimating speaking rate. I have investigated two basic features for automatic estimation of syllabic speaking rate, *recognition rate* and *mrate.* Given the differences in performance for the two measures, I have also added differential features that (perhaps) can capture some pronunciation variation.

### Recognition rate

The speaking rate can be estimated directly from a first pass of the decoder over the test material. This algorithm is relatively straightforward:

1. Get a time-aligned transcript of words for the acoustic sequence, including pauses.
2. Between every two pauses:
   (a) Let $t$ be the time between pauses in seconds.

[11] The ABBOT96 dictionary did not include lexical stress annotations.

(b) Let $n$ be the number of syllables found by looking up the transcribed words in the syllable dictionary.

(c) Compute rate $= \frac{n}{t}$

This rate measure is the *interpausal recognition rate*, as it is derived from a first recognition pass. Recognition rate does tend to underestimate the rate computed from hand transcriptions, particularly when the first-pass hypothesis has a high recognition error rate [Mirghafori *et al.*, 1996]. However, Mirghafori *et al.* used phone rate as the criterion for this study; since syllables are less mutable than phones, recognition syllable rate and transcribed syllable rate may be closer than they suggest.

### Mrate

At ICSI, we[12] have developed a measure of speaking rate that is derived from the acoustic signal [Morgan and Fosler-Lussier, 1998]. This measure, dubbed mrate for its **m**ultiple **rate** estimator components, uses a combination of three estimates: the first spectral moment of the wide-band energy envelope, a simple peak counting algorithm performed on the wide-band energy envelope, and a module that computes a trajectory that is the average product over all pairs of compressed sub-band energy trajectories. A block diagram of mrate can be seen in Figure 6.6.

The measure correlates moderately well with transcribed syllable rate,[13] although it tends to underestimate the rate for fast speech. We have noted in a number of individual cases that a high speaking rate sometimes results in the smearing together of energy peaks, even in sub-bands, which can lead to an underestimate of the syllable rate. For slow segments, high-energy phonetic onsets that are strongly correlated across bands and form distinct spectral peaks can confuse the estimator; these features are usually associated with syllable onsets. In these cases, mrate tends to overestimate the syllable rate.

Mrate also correlates less well with pronunciation reductions than the transcribed rate does. I repeated the experiment from Section 3.2.3, where for each of the 200 most frequent syllables, the pronunciations for each word were partitioned into histogram bins based on mrate. Only 54 of the 200 syllables showed significant shifts in the probability of the canonical or most likely pronunciations when mrate is used as the partitioning criterion, compared to 95 of 200 for transcribed rate. Mrate tends to underestimate the true rate when pronunciations are non-canonical, since reduced pronunciations sometimes have less sharp acoustic distinctions. When mrate matches or overestimates the true (*i.e.*, transcribed) rate, the probability of a canonical syllabic pronunciation is roughly 50%. However, as the amount that mrate underestimates the true rate increases, the canonical probability drops, reaching 33% when the rate is underestimated by 40% or more.

If recognition syllable rate is a closer match to transcribed rate, the two estimates of recognition rate and mrate may provide complementary information about pronuncia-

[12] The mrate measure was developed primarily by Nelson Morgan; I assisted in this research by providing evaluation statistics. See Morgan and Fosler-Lussier [1998] for more details about this algorithm.

[13] On the transcribed subset of the Switchboard corpus, we found that mrate calculated interpausally had a correlation with transcribed syllable rate of $\rho \sim .75$ [Morgan and Fosler-Lussier, 1998].

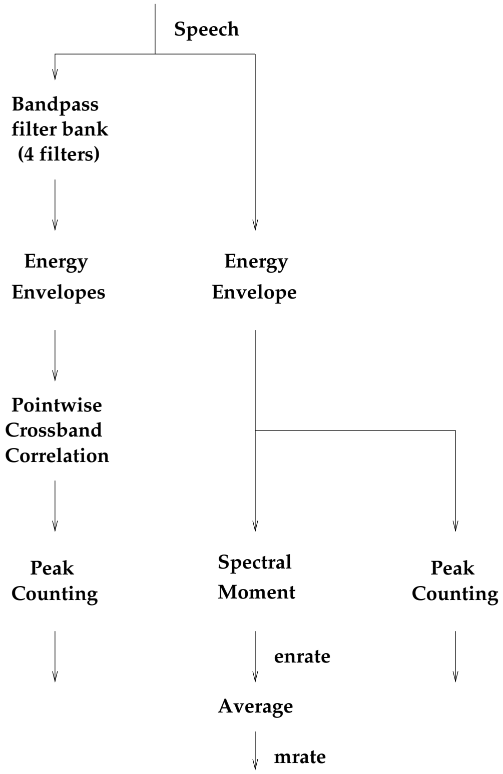


Figure 6.6: Major steps in the calculation of mrate, from [Morgan and Fosler-Lussier, 1998]. The bandpass processing currently uses steep FIR filters with band edges of (300,800), (800,1500), (1500,2500), and (2500,4000). Estimation is typically done over 1-2 seconds; for the experiments reported here, we used between-pause intervals, which varied from .25 to 6 seconds, but which were most typically between .5 and 2 seconds.

tion reduction. Thus, a differential measure may give information about when to expect pronunciation reductions. I have included the *rate ratio* (the ratio of mrate to recognition rate) and the *rate difference* (mrate subtracted from recognition rate) as two extra features in the d-trees.

#### Localized rate measures

Psychologists have suggested that many rate effects are localized [Summerfield, 1981]; I have found from initial experiments that estimating rate from the durations of a small number of syllables (between three and seven), as opposed to interpausally, sharpens some of the pronunciation distinctions described in Chapter 3. To incorporate this feature into the pronunciation model, the *local syllable rate* is computed directly from the hypothesis for each word by taking the number of syllables in the word and each of its immediately neighboring words and dividing by the total time allocated to the three words. In order to further distinguish local syllable rate from recognition rate, syllable boundaries are calculated online.[14]

An even more localized measure of rate was the word length; the length of the word in seconds was added to the feature list. Syllable lengths were unfortunately not immediately available from the first pass recognizer,[15] since segmentation was marked only at the word level in the recognition hypothesis.

### 6.4.3 Word predictability estimates

Two measures of word predictability were discussed in Section 3.1.2: the *unigram* probability and *trigram* probability of a word given a particular hypothesis. For the Broadcast News task, these measures were particularly easy to calculate, since a trigram grammar was used during the first-pass decoding.

#### Unigram probability

The unigram probability $P(w)$ — or unconditional probability — of each word in the test dictionary was pre-computed and recorded in a lookup table. For training, each word in the training data was simply annotated with the appropriate unigram probability; as each word was evaluated in testing, the unigram probability was provided by the lookup table.

[14] The online syllabification was due more to engineering choices rather than to any particular linguistic constraint, but it fortuitously showed some interesting effects when combined with recognition rate, as described in Appendix B.

[15] Online determination of the syllabic segmentation would not be difficult, requiring only a Viterbi pass over the recognized utterances. This could be implemented in a future version of the system.

**Trigram probability**

The training set was annotated with the trigram probability of each word ($P(w_i|w_{i-2}, w_{i-1})$) by feeding the word transcriptions into the NOWAY decoder running with the `-text_decode` option. The language model probability of each word obtained from NOWAY was then attached to the training word. For $n$-best list decoding, a similar process was employed: each hypothesis was scored separately by NOWAY. In the lattice decoder JOSÉ, the trigram language model scores were computed directly as part of the decoding process and passed to the pronunciation model.

### 6.4.4 Other features

One other feature that was investigated was the amount of time since the last pause taken by the speaker, as marked by silence in the recognition hypothesis. The motivation behind this feature is that, intuitively, the longer the speaker has gone without a breath, the less energy the speaker can place behind the speech, perhaps leading to more reductions. I also thought about incorporating the time until the next pause as another feature, but this feature requires knowing the entire hypothesis in advance (or at least knowing where the next pause is in the hypothesis), which is not feasible in lattice decoding.

There are a number of other features that one could imagine using in dynamic models, including energy, pitch, or phrase-level stress. I did not compute these features for this set of experiments, but they are likely candidates for inclusion as features in future experiments.

## 6.5 Dynamic decision tree models

In my initial experiments, I used the 1997 Broadcast News training set as the source of pronunciations for the word and syllable trees. The training set was automatically phonetically transcribed by means of smoothed phone recognition; the acoustic models used for training set generation were A-Model I — the combination of the two recurrent neural nets and the multi-layer perceptron described in Section 5.3.6.

550 word models were constructed from the smoothed transcriptions obtained by aligning A-Model I to the 1997 training set. The word d-trees included the phonetic, word identity, speaking rate, and predictability features described in the previous section to select appropriate pronunciation distributions, using the tree-growing algorithm described in Section 5.3.1. Slightly less than half of the trees in each case had a distribution other than the prior — that is, the constructed tree had more than one leaf. This set of trees were labeled the BN97 word trees.

I also trained roughly 800 d-trees based on syllable distributions (BN97 syllable trees). As described above, each word was given a single canonical syllable transcription, so that words with similar syllabic-internal pronunciation variations in the ABBOT96 dictionary shared the same syllable model. In addition to the features found in the word trees, syllabic tree context features included the lexical stress of the syllable, its position within the word, and the word's identity.

### 6.5.1 Tree analyses

In order to judge the quality of the trees, I extended the measurement paradigm of Riley *et al.* [1998]. In their system, the average log (base 2) probability of a held-out test set is calculated, giving a measurement related to the perplexity.[16] This score can be obtained by filtering the test set down through the trees to the leaves; as each sample reaches a leaf, the probability of that example according to the leaf distribution is recorded.

The average log probability is problematic as a metric for evaluating pronunciation models. Some test examples receive zero probability from the pronunciation model; this makes the measure unusable, as $\log_2(0) = \infty$. This can happen in two (related) situations in pronunciation modeling: test transcriptions can occur that are not covered by the model due to the nature of statistical modeling. Moreover, pronunciation modelers often explicitly introduce zeroes into baseform distributions by pruning the model. The language modeling (LM) solution to this dilemma is to never assign any model zero probability; the carry-over to the pronunciation domain would be to assign a minimum probability to unknown pronunciations. The value of the minimum probability is somewhat arbitrary (although one can use backoff techniques to obtain better estimates). More to the point, disallowing zero probabilities does not match the way models are used within an ASR system, since each word has only a finite number of baseforms. One can ignore "out-of-vocabulary" pronunciations (to use another LM term) and compute the log probability; however, this technique favors models that prune low-probability baseforms heavily; it does not penalize test set errors as much because zero-probability pronunciations are not counted.

Another problem with the $\log_2$ metric pertains to its use with scoring syllable and word trees. Riley *et al.* were building phone trees when they used this metric; every training sample in the corpus was associated with a decision tree. In my case, though, models were not constructed for some words and syllables due to lack of training data. One could substitute static dictionary priors for the missing words, but the appropriate substitution for missing syllable models is not clear.

To address these issues, I chose to compile three test-set statistics. I evaluated the average log probability only for pronunciations receiving non-zero probabilities (*i.e.*, baseforms occurring in the model). I also report the percentage of evaluated baseforms included in the scoring as the "pronunciation coverage." Additionally, the percentages of words or syllables in the test set that are actually modeled is also stated. This testing paradigm allowed me to test pronunciation models under the assumption that some pruning would be used within the ASR system. In unpruned models, pronunciation coverage remains the same no matter what features are used, but when pruning is invoked, the coverage will vary depending on which pronunciations are eliminated at each tree leaf.

Riley *et al.* also use a measure called *efficiency*, which is the relative increase in

[16] The perplexity is 2 raised to the power of the average negative log probability, that is, $PP = 2^{\frac{1}{n}\sum -\log_2 P(\mathrm{pron}|\mathrm{model})}$.

| | No pruning | | Prune < 0.1 | |
|---|---|---|---|---|
| | Avg $\log_2$ | Pronunciation | Avg $\log_2$ | Pronunciation |
| Features | Probability | Coverage | Probability | Coverage |
| None (baseline) | -0.70 | 92.58% | -0.53 | 89.39% |
| Word context only | -0.65 | 92.58% | -0.47 | 89.47% |
| Word and phone context | -0.55 | 92.58% | -0.33 | 88.70% |
| All | -0.45 | 92.58% | -0.26 | 89.42% |

Table 6.3: Test set probabilities for word trees. 58.9% of test words were modeled by the 550 trees; unmodeled words are not included in totals. Percentages of pronunciations covered for the 550 modeled words are listed under "Pronunciation Coverage." Higher $\log_2$ scores are better models; the maximum score would be zero.

log probability (LP) given some baseline, measured as

$$\text{Efficiency(model)} = \frac{LP_{\text{baseline}} - LP_{\text{model}}}{LP_{\text{baseline}}} \times 100\%. \tag{6.4}$$

Thus, a perfect model (with $LP_{\text{model}} = 0$) would have an efficiency of 100%.

### Word trees

I evaluated the effectiveness of the additional features by measuring the $\log_2$ probability on the secondary cross-validation set[17] for several sets of d-trees; these trees differed only by which features were included during construction (Table 6.3). In the baseline model, the pronunciation probabilities were set to the prior distributions over the training set. This model corresponds to a (simple) automatic baseform learning scheme; on the test set this model has a average log probability of -0.70. Comparing the unpruned to the pruned coverage numbers, roughly 3% of pronunciations in the test corpus had probabilities of 0.1 or less according to the prior model. Two metrics exist for calibrating improvement from this baseline model: increase in the log probability and in the pronunciation coverage for the pruned model.

Including just the word context (corresponding to a multi-word model) only increases average log likelihood by 0.05 (for an efficiency of only 7%). A bigger gain comes from adding in the surrounding phone context (21% efficiency); using all of the features gives the best efficiency of 35%, as one would hope. Comparing these results to those of Riley *et al.* [1998], the efficiency gain seems to be remarkably similar; they found efficiency gains of 20% to 32% depending on the training data. Yet, one must be careful in comparing these results. Riley's team was testing phone models on hand-transcribed data, whereas I am working with word models on automatically transcribed data.

When pruning is invoked, larger efficiency gains result; the trees using all features have a 51% efficiency rating. This means that, on average, it is the pronunciations that have

[17] As described in Chapter 5, the d-trees were built using 10-fold cross-validation (CV); a second test set was used to further prune the trees. In practice, the second CV set did not affect tree pruning, so this set is semi-independent from the d-trees.

higher probabilities ($p > 0.1$) in the baseline model that are increasing in likelihood due to the contextual modeling. Meanwhile, the actual percentage of test pronunciations that have a probability above 0.1 does not change significantly with the increased context (only 0.03% more of the pronunciations, representing 1% of the pruned baseforms in the baseline model, have probabilities above the pruning threshold in the d-tree with all features).

A list of the features used in the "all" trees is shown in Table 6.4, rank-ordered by the number of appearances in the tree. The number of occurrences at each depth of the tree completes the right side of the chart. A number of interesting patterns can be found in this data:

- The most frequently occurring feature is the duration of the word being modeled. This feature helps determine when vowel reductions and stop deletions are likely.
- The phone-based features appear high in the list. The features that group phones into classes (manner and place) tend to be placed early in the tree, whereas the phonetic identity features are spread throughout the tree.
- Previous and next word are prominent, but not ubiquitous.
- The following segmental context is always more prominent than the previous context. This is expected, because syllable onsets (and hence word onsets) have much less variation than nuclei and codas [Greenberg, 1998]; therefore, variations are much more likely to happen word-finally, dependent on the following segment.
- Consonantal features are more prevalent than vocalic features, probably for a similar reason — consonants occur more frequently on word boundaries.
- The extra features encoding speaking rate and predictability (save word duration) serve a secondary role to the segmental features.

### Syllable trees

I also computed test set scores for the 800 syllable trees from the BN97 training set. Unlike the 59% word coverage for the word trees, the coverage for the syllable test set was 79%, a much higher proportion of the test set. The efficiencies of the syllable models were a little higher than those for the word d-trees,[18] reaching 37% for unpruned and 54% for pruned models. The real gain, however, was in pronunciation coverage: 9% of the pronunciations lost in pruning the baseline model were recovered under the d-tree models.

The non-segmental features did not improve the model as much as in the word trees (cf. "All" to "Word/Syllable/Phone context"). The increase in $\log_2$ probability is only about half of that seen when these features are included in word tree construction.

To compare syllable models with the more conventional phone-based methods, I took the syllable training set and broke apart the syllables into phone models. Since the

[18] It is important to compare efficiencies or relative increases in $\log_2$ probability, and not actual probabilities, as the test sets are different due to coverage constraints.

| Feature | | All depths | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| current | word duration | 137 | 59 | 36 | 17 | 15 | 6 | 2 | 2 | - | - | - |
| next | consonant manner | 104 | 9 | 8 | 4 | - | - | 1 | - | - | - | 1 |
| next | phone ID | 79 | 27 | 20 | 10 | 9 | 2 | 5 | 2 | 4 | - | - |
| previous | phone ID | 45 | 10 | 10 | 11 | 3 | 6 | 2 | 2 | 1 | - | - |
| next | consonant place | 36 | 8 | 27 | 1 | - | - | - | - | - | - | - |
| next | word ID | 30 | 7 | 5 | 7 | 5 | 4 | 1 | 1 | - | - | - |
| previous | word ID | 23 | 5 | 2 | 3 | 3 | 2 | 2 | 3 | 3 | - | - |
| current | local rate | 19 | 10 | 4 | 3 | 1 | - | 1 | - | - | - | - |
| next | vowel manner | 15 | 11 | 4 | - | - | - | - | - | - | - | - |
| next | word duration | 11 | - | 2 | 2 | 3 | 1 | 3 | - | - | - | - |
| next | trigram | 11 | 1 | 3 | 4 | - | 2 | 1 | - | - | - | - |
| previous | word duration | 10 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | - | - | - |
| current | recog rate | 9 | - | 4 | 1 | - | - | 2 | 1 | 1 | - | - |
| previous | consonant manner | 8 | 3 | - | 2 | 1 | - | - | - | 2 | - | - |
| current | trigram | 8 | 1 | 3 | 2 | 1 | 1 | - | - | - | - | - |
| previous | recog rate | 7 | 2 | 2 | 1 | - | 1 | - | - | 1 | - | - |
| current | time since last pause | 7 | 1 | - | 2 | 3 | 1 | - | - | - | - | - |
| next | time since last pause | 7 | 3 | 2 | 2 | - | - | - | - | - | - | - |
| next | vowel place | 7 | 2 | 3 | - | 2 | - | - | - | - | - | - |
| previous | trigram | 6 | - | 1 | 4 | - | - | - | - | - | 1 | - |
| previous | mrate | 6 | 1 | 1 | 2 | 1 | 1 | - | - | - | - | - |
| next | recog rate | 5 | - | - | 2 | 2 | 1 | - | - | - | - | - |
| previous | consonant place | 5 | 2 | 1 | 2 | - | - | - | - | - | - | - |
| current | rate ratio | 4 | - | - | 2 | 1 | - | - | 1 | - | - | - |
| current | mrate | 4 | - | 1 | 1 | 1 | 1 | - | - | - | - | - |
| next | mrate | 4 | 1 | 1 | - | 2 | - | - | - | - | - | - |
| previous | time since last pause | 4 | 1 | - | - | 1 | - | 2 | - | - | - | - |
| previous | rate difference | 3 | - | 1 | - | - | - | - | 1 | 1 | - | - |
| next | rate ratio | 3 | - | 1 | - | 1 | - | - | - | - | 1 | - |
| current | rate difference | 3 | - | - | 1 | 1 | 1 | - | - | - | - | - |
| next | syllabic phone position | 3 | 3 | - | - | - | - | - | - | - | - | - |
| previous | vowel place | 2 | 1 | 1 | - | - | - | - | - | - | - | - |
| previous | vowel manner | 2 | - | - | - | - | 1 | 1 | - | - | - | - |

Table 6.4: Frequency of feature occurrence in word trees; columns labeled 1-10 represent the depth in the tree at which the feature was found (1=tree root).

| | No pruning | | Prune < 0.1 | |
|---|---|---|---|---|
| Features | Avg $\log_2$ Probability | Pronunciation Coverage | Avg $\log_2$ Probability | Pronunciation Coverage |
| *Syllable trees:* | | | | |
| None (baseline) | -0.70 | 95.96% | -0.46 | 91.74% |
| Word/syllable/phone context | -0.49 | 95.96% | -0.26 | 92.17% |
| All | -0.44 | 95.96% | -0.21 | 92.14% |
| *Phone trees w/syllabic-internal variants:* | | | | |
| None | -1.45 | 94.95% | -0.96 | 84.49% |
| Phone context | -0.60 | 94.95% | -0.33 | 90.42% |
| All | -0.54 | 94.95% | -0.25 | 90.39% |

Table 6.5: Syllable test set probabilities. 78.5% of test syllables were modeled by the 800 syllable trees; unmodeled syllables are not included in totals. Phone tree probabilities were combined to form syllable pronunciations, and were scored on the same subset of syllables used to score syllable trees.

syllable models contained variants (for example, the syllable `[k_l_ow_s/z]` has encoded the fact that the final phone can alternate as `[s]` or `[z]`), this would give them an advantage over regular phone models. Therefore, I built separate trees for the phone variants listed in the syllable models, *e.g.*, the final segment of `[k_l_ow_s/z]` was modeled by the phone `[s/z]`. The phone trees were then scored only on the syllable level, where pronunciations for the syllable were determined by concatenating the individual phone pronunciations from each tree; syllable pronunciation probabilities were obtained by multiplying together the phone probabilities. The subset of test syllables modeled by the syllable trees were used for scoring these models.

Without context, phone trees have a large decrease in log likelihood modeling on the phone level (an efficiency of -107% compared to the syllable tree baseline). By adding in contextual elements, the phone trees perform only a little bit worse than syllable trees, although the pronunciation coverage is significantly worse for both the unpruned and pruned cases. Syllable trees utilizing only segmental features outperform the phone d-trees with all features at their disposal. Thus, it seems that syllable models are as good, if not better, than phone models as an organizational structure for modeling the variation in pronunciations. Syllable models have the drawback of less coverage overall; one can model the entire corpus with phone models. Yet, coverage will be incomplete with syllable models. One could supplement unmodeled syllables with phone trees in a hybrid syllable-phone approach; or perhaps one could deconstruct syllables into onset models and rime (nucleus and coda) models, since most of the variation occurs within the rime.[19]

Table 6.6 lists the d-tree features used in the syllable trees. The patterns seen here are very similar to that of the word trees, with a few differences:

- Phone features are more prominent than the length of the word in these trees.

[19] One could also split the rime models into separate nucleus and coda models.

| Feature | | All depths | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| next | phone ID | 167 | 83 | 41 | 23 | 11 | 4 | 3 | 1 | 1 | - |
| current | word duration | 166 | 47 | 44 | 39 | 18 | 13 | 4 | 1 | - | - |
| next | consonant manner | 136 | 102 | 22 | 10 | 1 | 1 | - | - | - | - |
| previous | phone ID | 99 | 21 | 26 | 15 | 17 | 12 | 6 | 1 | 1 | - |
| next | syllableID | 84 | 21 | 27 | 21 | 6 | 3 | 5 | 1 | - | - |
| next | consonant place | 62 | 11 | 35 | 8 | 6 | 1 | 1 | - | - | - |
| previous | syllableID | 61 | 26 | 13 | 11 | 6 | 3 | - | 2 | - | - |
| previous | consonant manner | 22 | 6 | 4 | 5 | 2 | 2 | 1 | 2 | - | - |
| current | local rate | 20 | 1 | 5 | 2 | 8 | 1 | 1 | 1 | - | 1 |
| next | vowel manner | 18 | 14 | 4 | - | - | - | - | - | - | - |
| current | word ID | 15 | 10 | 1 | 3 | 1 | - | - | - | - | - |
| current | recog rate | 13 | 1 | 3 | 1 | 1 | 1 | 4 | 1 | 1 | - |
| current | time since last pause | 13 | - | 4 | 2 | 3 | 1 | 2 | - | 1 | - |
| next | unigram | 12 | 1 | 4 | 2 | 2 | 1 | 1 | 1 | - | - |
| next | word duration | 11 | - | 2 | 6 | 1 | 1 | 1 | - | - | - |
| next | vowel place | 11 | - | 8 | 1 | 1 | 1 | - | - | - | - |
| current | word position of syl. | 10 | 4 | 3 | 1 | - | 1 | - | 1 | - | - |
| previous | word duration | 9 | - | 1 | 3 | 1 | 2 | 1 | - | 1 | - |
| current | lexical stress | 9 | 4 | 4 | 1 | - | - | - | - | - | - |
| previous | word ID | 9 | 1 | 2 | - | 3 | - | 2 | 1 | - | - |
| next | syllabic phone position | 9 | 9 | - | - | - | - | - | - | - | - |
| current | mrate | 9 | - | 3 | 2 | 2 | 1 | 1 | - | - | - |
| next | lexical stress | 8 | 1 | 3 | 4 | - | - | - | - | - | - |
| next | trigram | 8 | - | 3 | 2 | 1 | - | 1 | - | 1 | - |
| next | time since last pause | 7 | - | 2 | 3 | - | 1 | - | - | 1 | - |
| current | trigram | 7 | - | 1 | 1 | 2 | 2 | - | - | 1 | - |
| previous | consonant place | 7 | 1 | 3 | 1 | 1 | 1 | - | - | - | - |
| current | unigram | 7 | 2 | 2 | 1 | 2 | - | - | - | - | - |
| previous | vowel manner | 6 | 1 | - | 3 | 2 | - | - | - | - | - |
| current | rate difference | 5 | - | - | 1 | - | 3 | - | 1 | - | - |
| next | word ID | 5 | 2 | 1 | - | 1 | - | - | 1 | - | - |
| previous | unigram | 5 | 1 | - | - | 1 | 3 | - | - | - | - |
| current | rate ratio | 4 | - | 1 | - | - | 1 | 1 | 1 | - | - |
| previous | syllabic phone position | 4 | 2 | 2 | - | - | - | - | - | - | - |
| next | word position of syl. | 4 | 2 | 1 | - | 1 | - | - | - | - | - |
| previous | trigram | 3 | - | - | 1 | 1 | 1 | - | - | - | - |
| previous | word position of syl. | 2 | - | - | 2 | - | - | - | - | - | - |
| previous | vowel place | 2 | 1 | - | 1 | - | - | - | - | - | - |
| current | syl. count in word | 2 | 2 | - | - | - | - | - | - | - | - |
| next | syl. count in word | 2 | - | 1 | - | - | - | 1 | - | - | - |
| previous | time since last pause | 1 | - | - | - | - | - | 1 | - | - | - |

Table 6.6: Frequency of feature appearance in syllable trees.

- Instead of word identities, syllable identities are prominent in the list.
- Trigram probabilities are less prevalent — perhaps due to the introduction of unigram probabilities, which are used more frequently.

### Individual tree analyses

In this section, I examine some of the BN97 trees, particularly to ascertain how features interact to generate pronunciation distributions at the leaves. Even though the pronunciations given in these trees are automatically derived (*i.e.*, not from hand transcriptions), one can see interesting linguistic phenomena modeled by the classifier. Decision trees for two words and two syllables are shown in the next few figures for illustrative purposes; other trees that show different linguistic processes than the ones below (*e.g.*, flapping) can be found in Appendix B. In these graphs, the features listed in ovals are the same ones found in Tables 6.4 and 6.6, although the names have been abbreviated for readability (*e.g.*, *cur, next,* and *prev* correspond to the current, next, and previous word or syllable, respectively). The shading of the tree leaves corresponds to the most likely pronunciation of the leaf — the first pronunciation has the lightest shade of gray, and the last listed baseform has the darkest.

Figure 6.7 illustrates the decision tree for the word *them*. Three pronunciations of this word are commonly found in the corpus: the canonical `[dh_eh_m]`, a variant with a reduced vowel (`[dh_ax_m]`), and a variant with a reduced vowel and final consonant change (`[dh_ax_n]`). The first split in the tree is very revealing with respect to this last variant. The phones listed at this node are mostly alveolar or dental phones, so if the first phone of the next word is one of these, the `[m]` of *them* assimilates to `[n]`, the articulatory place of the next phone. This effect is particularly strong for following `[d]` and `[t]` (shown by the next split on the left). Nowhere else in the tree does `[dh_ax_n]` have significant probability. It is clear, though, that the automatic techniques are not perfect, at least in a linguistic sense, since the phones `[m]` and `[h#]` are included with the alveolar/dental phones in the first split — an unexpected grouping.

For those examples not listed in the top split, word duration is the second most important feature. When the duration is short, it is very likely that the pronunciation is reduced. For longer words, a short time since the previous pause (roughly six syllables or fewer, including this word) also indicates a likely reduction. This contradicts the hypothesis that a longer time from the last pause will induce more reductions.

At this point in the tree, we are left with medium to long words that occur a while after the last pause, and are not followed by words beginning with alveolar/dental/`[m]`/`[h#]` phones. Longer words of this class, unsurprisingly, are more likely to be unreduced. For medium-length words, however, the identity of the next word comes into play. At first glance, the list of words that affect pronunciation seem very strange: one would not expect to hear *them were*, for instance, contiguously in a sentence.[20] This is because all of these collocations happen in a very few phrase structures. In the vast majority of training corpus

[20] as opposed to the phrase *they were*.

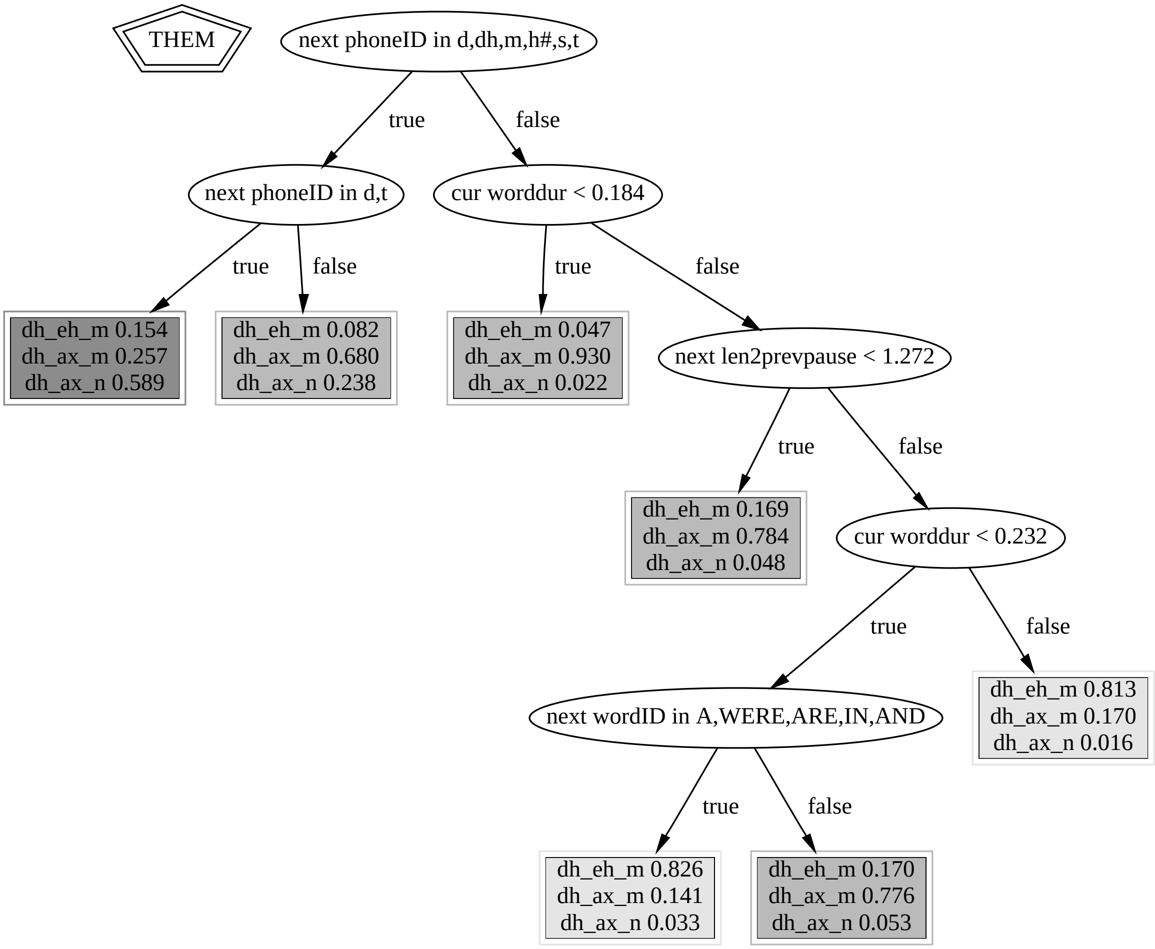


Figure 6.7: Word decision tree for *them*.

examples, the word *a* follows *them* only if there is an intervening sentence boundary[21] (example 1), as part of an indirect object expression (2), or as part of a small clause (3):

1. ... THEM <SENTENCE_BOUNDARY> A ...
2. {GOT/OFFER/HAND/GIVE/COST/MAKE} THEM A [lot of money] ...
3. ... SEEN THEM A LOT DARKER THAN THIS.

When the following word is *are* or *were*, the word *them* occurs only in a partitive phrase:

4. {SOME/NONE/ALL...} OF THEM {ARE/WERE}

When *and* follows *them*, it generally functions as something that conjoins clauses (*That's what I told them, and you'd better believe it...*), as opposed to conjoining noun phrases like *them and someone else*. It is clear that some syntactic constraints are in effect, and that word identity is compensating for the lack of phrasal information in the feature set. I do not want to hazard a guess as to exactly what the effects of syntax are here, because the data are not marked for a full syntactic analysis, and some observations are problematic: for instance, why is *in* clustered with these words, rather than *on, out,* and *from*, which appear in the other cluster? However, these data do suggest that some encoding of syntactic structure would be useful in future models.

In the case of the word *than* (Figure 6.8), another interesting linguistic phenomenon called *epenthesis* was found by the d-tree learning procedure. Epenthesis is the insertion of a sound within the phonetic stream due to the juxtaposition of two segments. In the second level of the tree, the pronunciation of *than* in the phrases *less than* or *worse than* includes an extra `[tcl]` (t-closure). This is very reasonable if one considers the articulator (tongue) position during this phrase: as the alveolar `[s]` moves to the initial `[dh]` (dental position), the tongue sometimes can make a closure at the alveolar ridge or teeth, particularly if the speaking rate is slower than normal.[22]

In most of the trees, word duration, recognition rate, and local speaking rate were much more prominent than the mrate measure that we developed. Where mrate does seem to be effective is in cases of ambiguous syllabification. Figure 6.9 shows the example of the syllables `[b_l_ax_m]` and `[b_l_ax_m_z]`, in which the `[l]` can be elongated into a syllabic `[el]` in slower speech, creating a second syllable (as in the PRAH-BUHL-EM pronunciation of *problem*). The recognition rate depends on the pre-compiled, canonical syllabification of the word. However, if an extra syllable is inserted, mrate will (usually) detect the extra syllable, since it is based on the acoustic signal rather than syllabification of the word hypothesis. This extra syllable increases mrate so the ratio of recognition rate to mrate will then be very low in this instance. This is borne out in the tree for `[b_l_ax_m]`: at the top of the tree, fast recognition rate examples are likely to occur as one syllable. Further down the tree (after other constraints on likelihood), mrate and the rate ratio appear as features.

[21] Sentence boundaries are not marked because the recognizer has no way of knowing *a priori* where the sentence boundaries lie.

[22] as typified by the longer word duration requirement in the node above *less* and *worse*.

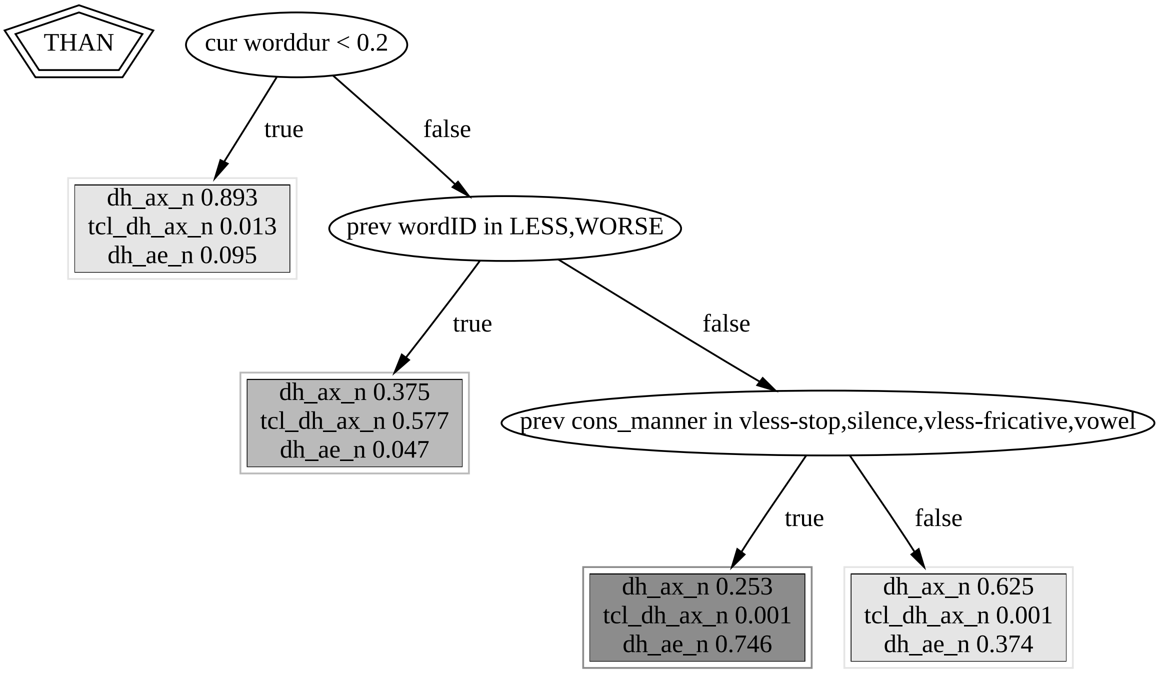


Figure 6.8: Word decision tree for *than*.

In [b_l_ax_m_z], the trend is even clearer: short duration and very high mrate indicate an extra syllable, as well as longer durations (slower recognition rates) with a low rate ratio.

These trees illustrate a significant interaction among the features. One cannot help but wonder whether d-trees are the right model for these continuous features, since early splits may disallow interactions between later features. For example, in the tree for *them*, what if there were syntactic effects having to do with the next word being *to* related to the other effects of *a, and,* and *were*? Since the d-tree partitioned off the samples of following words with initial [t] early in the tree, the syntactic influence of *to* was not exhibited in the tree. Smoother learning techniques, such as neural networks[23] or Bayesian Decision Trees [Jordan and Jacobs, 1994], may allow for better interaction between features.

### 6.5.2 $N$-best list rescoring

The effectiveness of the BN97 syllable and word d-trees can also be determined by integrating them into the ASR system. In this section, I describe the results of rescoring Broadcast News $n$-best lists with these new pronunciation models.

Lattices were generated for the 173 segment subset of the 1997 Hub4E Broadcast News test set that was defined in Chapter 4. NOWAY used the SPRACH98 dictionary and

[23] An example of an ANN system for pronunciation learning was presented by Fukada *et al.* [1999], although they do not incorporate features other than phonetic context.

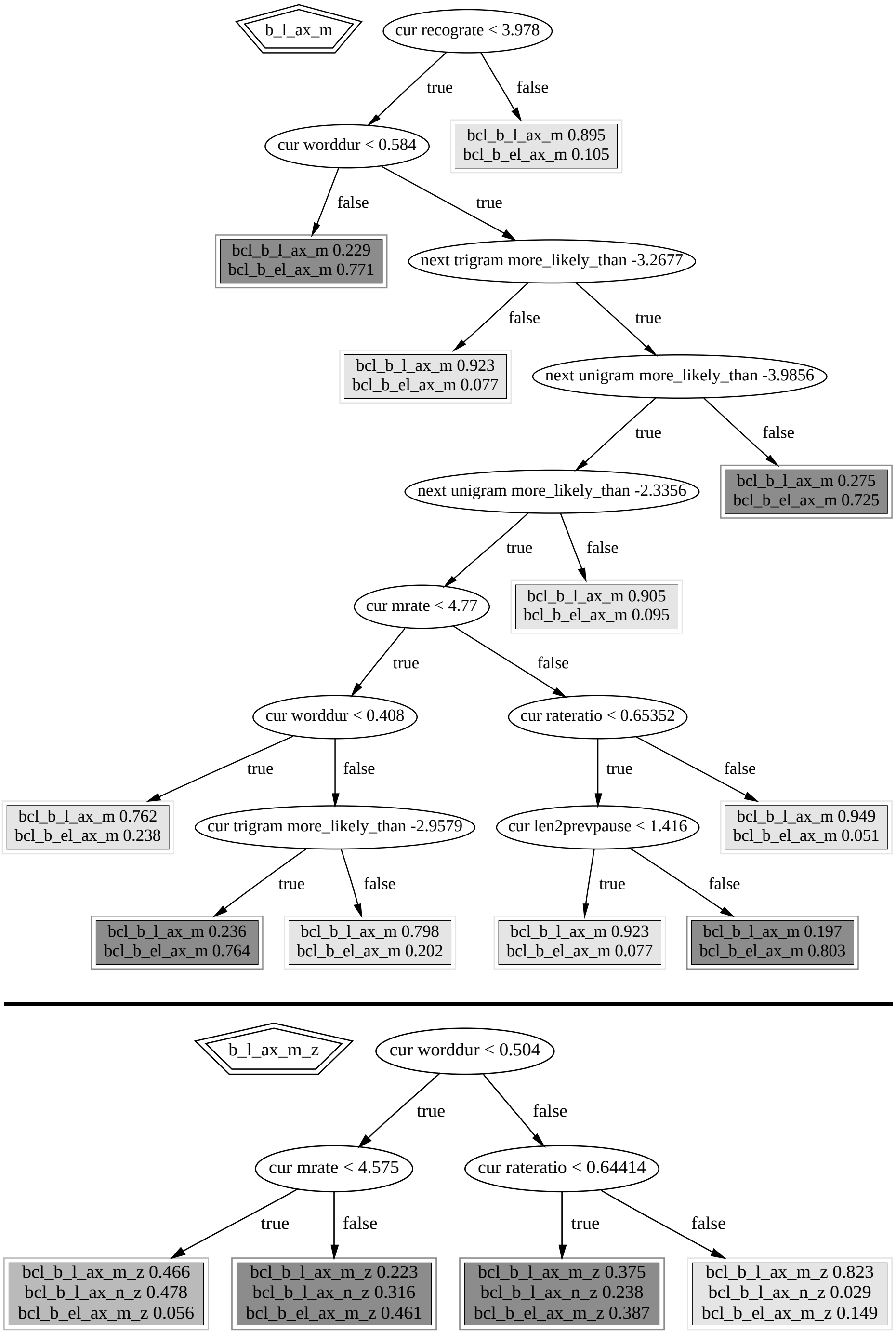


Figure 6.9: Decision trees for the syllables [b_l_ax_m] and [b_l_ax_m_z].

| Dictionary | 100-best WER |
|---|---|
| first pass NOWAY ABBOT96 | (27.5%) |
| first pass NOWAY SPRACH98 | (26.9%) |
| SPRACH98 1-best (baseline) | 26.7% |
| BN97 Word trees | 26.5% |
| BN97 Syllable trees | 26.3% |

Table 6.7: Hub4E-97-Subset Word Error Rates for dynamic tree rescoring of $n$-best lists under A-Model I.

A-Model I to generate the word graphs; the pruning parameters were the wide evaluation parameters (27-hyps). The LATTICE2NBEST program then decoded the lattice into a list of the $n$ best hypotheses. For the rescoring experiments described in this thesis, I chose $n$ to be 100 for two reasons: in initial experiments, rescoring more than 100 hypotheses did not seem to make much difference in final error rate, and allowing only 100 hypotheses reduced the decoding time necessary for experiments.

To allow for discrepancies in the decoding process, I re-evaluated the word error rate on the first-best hypothesis for each utterance from the LATTICE2NBEST decoder. Table 6.7 shows that the new baseline from this re-evaluation is 26.7%, which is very close to the NOWAY result of 26.9%.[24]

In this test, syllable trees outperformed word trees, although neither differs significantly from the SPRACH98 best hypothesis. The small performance increase of the syllable trees may be attributable to the increase in coverage provided by the syllable trees (compared to the word-based trees). Comparing BN97 Syllable trees to the first-pass NOWAY results, the overall gain for syllable trees is roughly the same as the improvement of the SPRACH98 dictionary over ABBOT96, although part of this is attributable to the difference in decoders.

As opposed to the across-the-board improvements seen with the SPRACH98 static dictionary in most focus conditions, the 0.4% difference between $n$-best decoding of the baseline and the syllable trees was accounted for almost completely by a 1.4% improvement in WER in the spontaneous broadcast speech focus condition (F1), and a 0.9% improvement for speech with background music (F3). Thus, it may be the case that improved static dictionary modeling helps in all conditions, whereas dynamic dictionaries improves performance only in particular conditions that are more difficult.

### 6.5.3 Lattice rescoring

In a second experiment, I rescored the lattices using the JOSÉ lattice decoder (instead of creating $n$-best lists from the lattices and rescoring via Viterbi alignment). The lattice decoder baseline achieved a slightly worse error rate than the $n$-best and NOWAY

[24] Obtaining similar results from LATTICE2NBEST and NOWAY required minor modifications to the LATTICE2NBEST program, as well as a week of debugging.

| Dictionary | Lattice WER |
|---|---|
| SPRACH98 (baseline) | 27.0% |
| BN97 Word trees | 26.6% |
| BN97 Syllable trees | 26.4% |

Table 6.8: Hub4E-97-Subset Word Error Rates for dynamic tree rescoring of lattices under A-Model I.

baselines.[25] Notably, the improvement from dynamic pronunciation models for both model types parallels the $n$-best rescoring results; it appears that full lattice decoding does not add much compared to $n$-best rescoring, although this is difficult to determine because of the difference in the baseline results.

### 6.5.4 Experiments with the 1997 and 1998 training sets

As in the static dictionary experiments, when the later acoustic model consisting of the RNN and two MLPs became available (A-Model II), I regenerated both sets of trees using the 1997 and 1998 training sets, providing 1300 syllable and 920 word classifiers (labeled as the BN97+98 trees). I also trained a separate set of trees on the segmental context features alone, to determine the influence of secondary features such as speaking rate.

$N$-best lists were regenerated for Hub4E-97-Subset using A-Model II and the BN97+98 static dictionary; in this experiment I rescored the lists using both the BN97 and BN97+98 trees. Table 6.9 shows that none of the trees made a significant difference in performance. However, we can see some general trends across the experiments: first, BN97 trees performed worse than both the baseline and BN97+98 trees. This suggests that dynamic models may be more susceptible to changes in the acoustic model, since the BN97 trees were trained using A-Model I. Also, in a reversal of the A-Model I experiment, the BN97+98 word trees outperformed the syllable trees. The increase in training data, which allowed for greater coverage of the corpus by the word trees, may have contributed to this result. However, the differences between the two systems are small and not statistically significant.

When non-segmental features like speaking rate and trigram probability were removed from the trees, performance improved. I have highlighted the lowest error rate in each focus condition across all seven experiments; the lowest error in five of seven focus conditions occurred with a segmental context model (lines 4 and 7). These results indicate that the measures of speaking rate and word predictability used here may not be robust enough for use in a dynamic model.

On a final note, as in the initial experiments, the modest improvements of the dynamic models (*e.g.*, BN97+98 segmental word trees) were concentrated in the non-F0

[25] Again, this is due to slight differences in the decoding strategy. For example, NOWAY guides the search with a calculation of the least upper bound on the hypothesis score [Renals and Hochberg, 1995b]; this is omitted in JOSÉ.

| | | | Overall | Focus Condition WER (%) | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Dictionary | | WER (%) | F0 | F1 | F2 | F3 | F4 | F5 | FX |
| 1. | Static: | BN97+98 (baseline) | 23.6 | 13.5 | 23.3 | 34.5 | 29.2 | 26.6 | 16.8 | 44.4 |
| 2. | Word: | BN97 | 23.5 | 13.5 | 23.5 | 34.7 | 27.8 | 26.8 | 16.8 | **43.3** |
| 3. | | BN97+98 All Features | 23.4 | 13.2 | 23.0 | 34.6 | 27.8 | 26.8 | 17.6 | 44.6 |
| 4. | | BN97+98 Segmental Context | **23.1** | 13.2 | **22.6** | 34.0 | **26.3** | **26.0** | 16.8 | 44.4 |
| 5. | Syllable: | BN97 | 23.8 | 13.5 | 23.1 | 35.5 | 29.2 | 26.8 | **15.1** | 45.9 |
| 6. | | BN97+98 All Features | 23.3 | 13.3 | 23.3 | 34.5 | 29.2 | 26.6 | 16.8 | 44.4 |
| 7. | | BN97+98 Segmental Context | 23.2 | **13.1** | 22.9 | **33.9** | 27.8 | 26.6 | 16.0 | 45.7 |

Table 6.9: Hub4E-97-Subset word error rate for dynamic evaluation of syllable and word tree models using A-Model II. The best error rate in each category is highlighted.

portions of the corpus, although, with the lack of statistical significance, one cannot draw definite conclusions about the ability of dynamic dictionaries to better model speech in these more difficult focus conditions.

### 6.5.5 Parameter tuning

Before running the final experiments on an independent test set, I took the opportunity to tune several of the parameters used in the dynamic trees. A search was conducted over the following variables:

**Included factors:** The trees that used only segmental factors instead of all features performed better in the experiments described in Table 6.9.

**Pruning threshold:** Pronunciation sequences that had a probability below the pruning threshold were discarded in each d-tree leaf.

**Acoustic model interpolation:** As Equation 6.3 suggests, the acoustic score for each utterance can be interpolated with the original (lattice) acoustic score. The tuning parameter $\lambda$ is chosen by searching over possible values between 0 and 1.

For both syllable and word d-trees, an exhaustive search was made over the included factors and pruning thresholds to optimize the word error rate on the 1997 Hub4E Subset test set. For the best two candidates from this search, a second search was conducted over possible $\lambda$ values; the best parameter settings were selected again by comparing word error rates.

Figure 6.10 shows the entire processing chain for word d-trees. Lattices were constructed by the NOWAY recognizer, using the best static dictionary from Chapter 5 (BN97+98). From these lattices, the 100 best hypotheses were derived, with a first-best hypothesis error rate of 23.6%. The best word d-trees from the optimization process proved to be the segmental-context-only trees, where pronunciations with probabilities of less than 0.01 were discarded. The optimal acoustic interpolation parameter was 0.55 (slightly favoring the dynamic acoustic scores over the static acoustic scores). Without interpolation, the dynamic pronunciation model had the same error rate as the static 1-best hypothesis. However, interpolating the two scores brought a 0.5% absolute error reduction.

Even though the word trees had the same baseline error rate as the 1-best hypothesis, an examination of the actual word sequences produced by each system found that the word hypotheses were often different. In these cases, it is often advantageous to combine hypotheses at the word level. The ROVER system from NIST [Fiscus, 1997] blends hypotheses that have words annotated with confidence scores. I integrated the first-best hypothesis, the best hypothesis from the word d-trees without acoustic score interpolation, and the best post-interpolation hypothesis; each word in all three hypotheses was annotated with the nPP acoustic confidence score (Section 5.3.5). I conducted a search over the combining parameters of ROVER;[26] the best resulting word error rate was very similar to the interpolated acoustic score result (23.0%). Despite the small gain, I suspected that using ROVER in this way would provide robustness in recognition on independent test sets.

[26] ROVER has several parameters that can be tuned; the best results were found by using the maximum confidence metric, with alpha=0.3 and Null_conf=0.7.

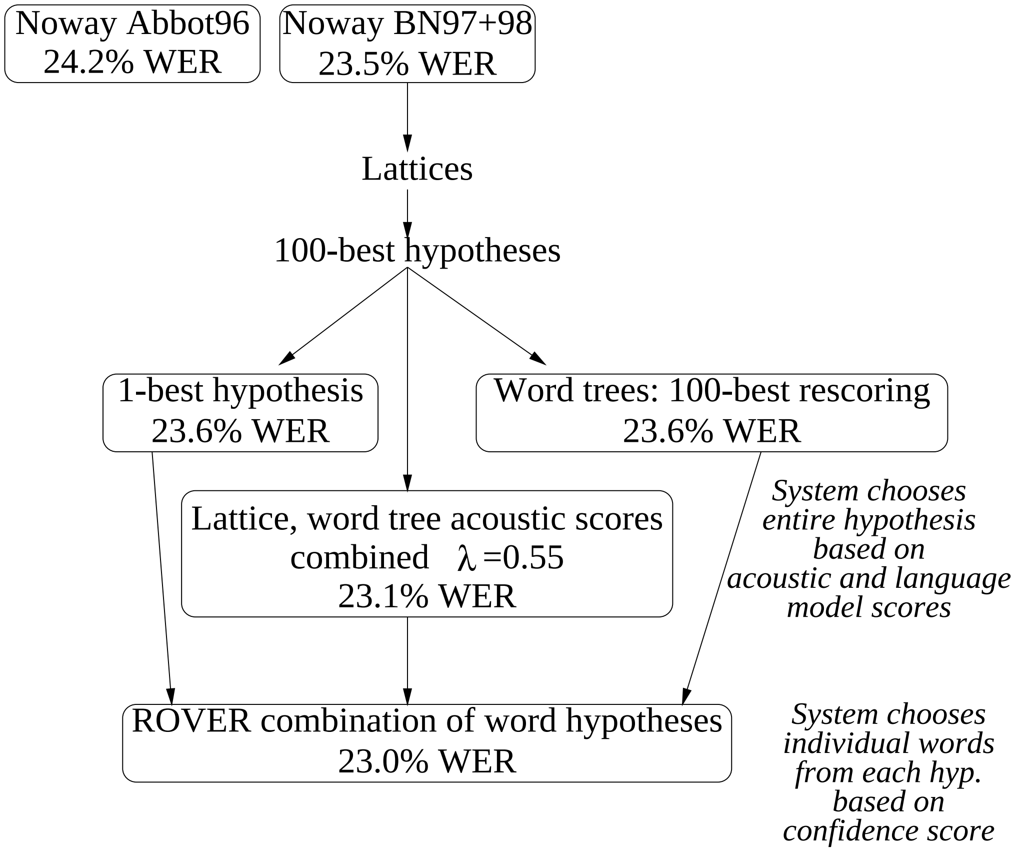


Figure 6.10: Results of tuning the combination of word-based dynamic pronunciation models with static dictionary components for 1997 Hub4E Subset.

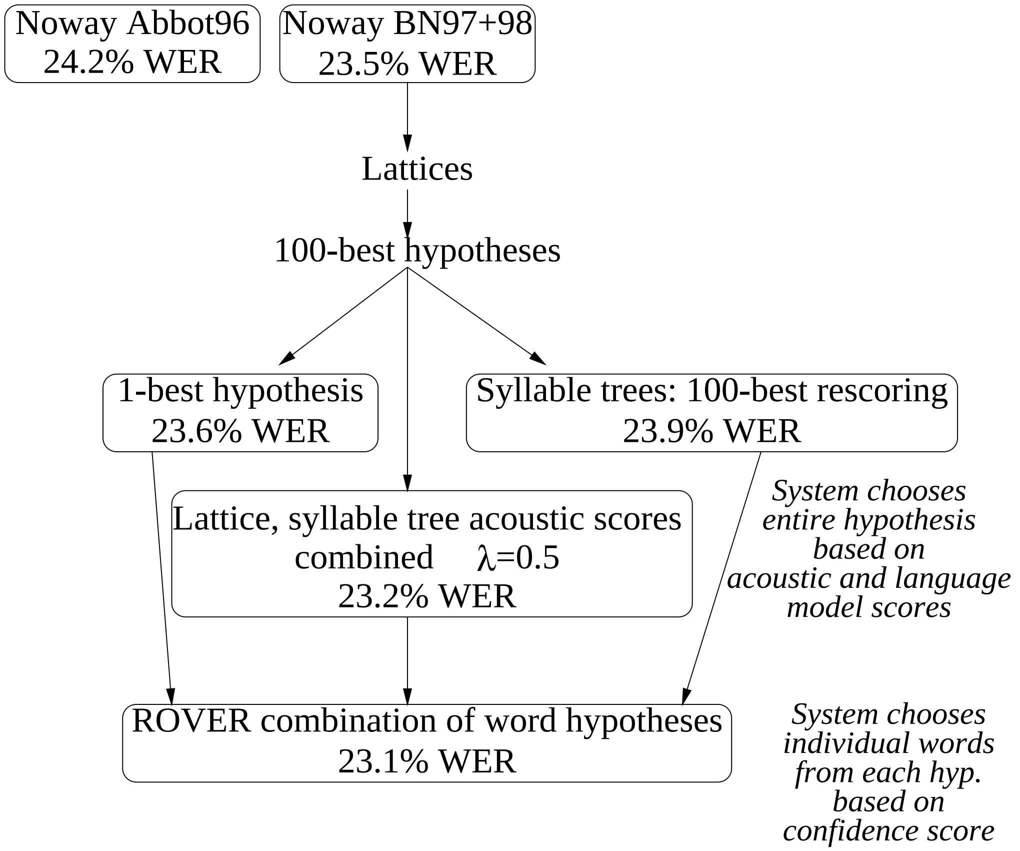


Figure 6.11: Results of tuning the combination of syllable-based dynamic pronunciation models with static dictionary components for 1997 Hub4E Subset.

I also performed the same tuning experiments on the syllable trees (Figure 6.11). The performance patterns were very similar to those of the dynamic trees. The syllable d-trees performed slightly worse by themselves than the 1-best hypothesis (23.9% WER compared to 23.6%). The optimal $\lambda$ acoustic interpolation was an even weighting between dynamic and static acoustic scores, providing 23.2% word error. The ROVER combination[27] was again almost the same as the interpolated result.

### 6.5.6 Evaluation on 1998 Broadcast News test set

I evaluated the best word d-tree system and syllable d-tree system on the 1998 Broadcast News test set [Pallett *et al.*, 1999], otherwise known as Hub4E-98. This independent test set was divided into two parts, each consisting of roughly 1.5 hours of speech. The first set contained news broadcasts from mid-October to mid-November of 1996, a time period in between the 1997 and 1998 Broadcast News training sets (containing data from 1996 and 1997, respectively). The second test set consisted of data from June 1998. The SPRACH 1998 evaluation system (reported in Cook *et al.* [1999]) included a vocabulary and language model updated to contain news stories from 1997 and 1998.[28] The dictionaries tested here, however, used the older ABBOT96 vocabulary to ensure consistency across tests. The dictionary training set is probably better matched to the first test set (Hub4E-98_1) than to the second due to the temporal overlap of the two sets.

The results of the final evaluation for the static and dynamic dictionaries are shown in Table 6.10. The static dictionary provides most of the improvement seen (0.6% overall, $p = 0.031$); including the word d-trees increases the improvement to 0.9% ($p = 0.014$ compared to ABBOT96). Compared to the BN97+98 dictionary, word d-trees give a small improvement in both test sets, whereas syllable d-trees show no improvement. As hypothesized, ROVER does increase robustness; when the word d-trees are evaluated independent of the ROVER combination, the word error rate is 21.7% in the uninterpolated case and 21.4% in the interpolated case.

Tables 6.11 and 6.12 show the word error rates for the focus conditions defined in the Broadcast News corpus, as well as separate error rates for female and male speech. The new static and dynamic pronunciation models never help in the planned speech condition (F0); in fact, performance degrades slightly in F0 for the 1996 test set (Hub4E-98_1) that matches the training set more closely. For studio spontaneous speech (F1), word trees double the static dictionary's performance increase over ABBOT96 (0.6% to 1.2%) for the 1996 test set, but for the 1998 test set, there is very little difference between the two dictionaries. For the other focus conditions, the dynamic word trees almost always seem to improve performance, the only exception being in the degraded acoustics condition (F4) for the first test set. The biggest absolute performance increases for the word trees were in

[27] The best ROVER parameters for syllable d-trees used the maximum confidence metric, but with alpha=0.5 and Null_conf=0.85.

[28] There are some differences between the baseline system used here and that reported in [Cook *et al.*, 1999]; in particular, that system used the SPRACH98 dictionary reported in Chapter 5 and included a context-dependent RNN as one of the acoustic models. In addition, there are many known improvements that have not been implemented in either this system or the SPRACH system, such as acoustic adaptation, which should improve results by 10-20% relative.

| | Word Error Rates (%) for Broadcast News | | |
|---|---|---|---|
| Dictionary | Hub4E-98_1 | Hub4E-98_2 | Hub4E-98 overall |
| ABBOT96 | 22.6 | 21.4 | 22.0 |
| Static BN97+98 | 22.2 | 20.7 | 21.4 |
| Dynamic word trees | 21.9 | 20.4 | 21.1 |
| Dynamic syllable trees | 22.2 | 20.6 | 21.4 |

Table 6.10: Final evaluation word error rates for the on the 1998 Broadcast News evaluation test sets.

the difficult FX condition.[29]

The most impressive combined static/dynamic performance, though, is for the telephone speech condition (F2): the automatically derived dictionaries were 12% better (relative) in this condition (corresponding to a 4.3% absolute decrease in word error across both test sets); even with the smaller test set size this is a significant difference ($p = 0.016$). This may be due to an interaction of the automatically derived pronunciation models with the acoustic model A-Model II: one of the three neural net models was trained on 8kHz (telephone) bandwidth speech using the Modulation Spectrogram Filtered (MSG) features. Since the pronunciation data for the dictionaries were generated by A-Model II, these new dictionaries may reflect the improved acoustic modeling for this focus condition.

There do not seem to be any significant patterns in word error rate due to gender. The word trees improve word error rate by 0.9% to 1.0% compared to the ABBOT96 dictionary in both female and one of the male test sets, the exception being Hub4E-98_1 male speech, for which only a 0.4% improvement is seen.

Dynamic rescoring with syllable trees was almost always worse than rescoring with word trees when compared across focus conditions. The syllable trees were also inconsistent with respect to the BN97+98 dictionary, sometimes performing better and sometimes worse.

## 6.6 Summary

The foundation of the work in this chapter was started at the 1996 Johns Hopkins Summer Workshop, where I collaborated with a research team that found that automatically learned pronunciation models improved performance on the Switchboard corpus by around 1% absolute. Almost all of the gain was due to building static pronunciation models directly from phone recognition; decision tree-smoothed static models were only 0.2% absolute better. Dynamic rescoring did *not* show any improvement.

Four basic questions motivated the research reported in this chapter. First was the question of whether dynamic rescoring would actually improve results at all. Subsequently, existence proofs for dynamic rescoring have been offered by Riley *et al.*'s [1998]

[29] The FX category contains all the speech that was difficult to fit into one of the other categories; usually this speech fits into more than one of the F0-F5 categories, *e.g.*, foreign accented speech over telephone lines.

| Dictionary | Overall WER (%) | Focus condition WER (%) | | | | | | | Gender WER (%) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F0 | F1 | F2 | F3 | F4 | F5 | FX | Female | Male |
| [Word Count] | [15651] | [2527] | [3879] | [513] | [527] | [4379] | [165] | [3661] | [6961] | [8662] |
| ABBOT96 | 22.6 | 12.0 | 25.2 | 35.7 | 36.8 | 16.5 | 29.1 | 30.2 | 23.4 | 21.6 |
| Static BN97+98 | 22.2 | 12.4 | 24.6 | 33.1 | 34.9 | 15.8 | 29.7 | 30.3 | 22.9 | 21.3 |
| Dynamic word trees | 21.9 | 12.5 | 24.0 | 32.6 | 34.2 | 16.2 | 27.3 | 29.3 | 22.5 | 21.2 |
| Dynamic syllable trees | 22.2 | 12.9 | 24.4 | 33.3 | 34.5 | 16.3 | 26.7 | 29.5 | 23.0 | 21.3 |

Table 6.11: Categorical word error rate for Hub4E-98_1 (10/96-11/96).

| Dictionary | Overall WER (%) | Focus condition WER (%) | | | | | | | Gender WER (%) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | F0 | F1 | F2 | F3 | F4 | F5 | FX | Female | Male |
| [Word Count] | [16784] | [7417] | [2367] | [582] | [858] | [4763] | [70] | [727] | [6199] | [10585] |
| ABBOT96 | 21.4 | 14.6 | 25.9 | 35.1 | 21.3 | 24.5 | 24.3 | 43.5 | 20.4 | 21.9 |
| Static BN97+98 | 20.7 | 14.6 | 25.3 | 31.3 | 21.0 | 23.5 | 20.0 | 41.8 | 19.8 | 21.3 |
| Dynamic word trees | 20.4 | 14.6 | 25.2 | 29.7 | 20.6 | 22.9 | 17.1 | 40.7 | 19.4 | 21.0 |
| Dynamic syllable trees | 20.6 | 14.6 | 25.8 | 29.2 | 20.7 | 23.0 | 21.4 | 40.7 | 19.4 | 21.2 |

Table 6.12: Categorical word error rate for Hub4E-98_2 (6/98).
Focus conditions: Planned Studio Speech (F0), Spontaneous Studio Speech (F1), Speech Over Telephone Channels (F2), Speech in the Presence of Background Music (F3), Speech Under Degraded Acoustic Conditions (F4), Speech from Non-Native Speakers (F5), All Other Speech (FX)

1997 JHU Workshop system that found improvements on Switchboard using dynamically rescored bigram lattices, and Finke and Waibel's [1997b] Switchboard system that dynamically changed phonological rule probabilities based on a hidden "speaking mode" variable. The work in this chapter adds to this evidence by showing that improvements over a well-tuned static dictionary are possible by dynamic rescoring of $n$-best lists using a word-based decision tree dictionary.

Another research question was whether building longer-term decision trees (to the extent of the syllable or word) would improve modeling. By making trees word- or syllable-specific, I hoped that the tree-building algorithm would spend less effort trying to model syllable or word-internal variation, and focus on effects at the model boundaries that were dependent on features other than the phonetic context. The comparison of syllable trees to phone trees bears this out: even when phone d-tree models are constructed with the knowledge of the possible pronunciation variants for a particular word (as was given to the syllable models), they are still worse than syllabic models in both test set probability and pronunciation coverage.

The models in this chapter also included extra features that were found to affect pronunciations in Chapter 3. In particular, various measures of speaking rate and word probability were given to the d-tree learning algorithm. Word duration was the most effective factor, followed by the localized speaking rate. It is clear that these features as a class are secondary to the phone, syllable, and word context; but part of the increase in the test set log probability can be attributed to the inclusion of these features. It appears, however, that their presence in the trees does not help ASR system accuracy for these experiments and methods.

The final question was whether techniques designed for the spontaneous speech of the Switchboard corpus would carry over to other corpora. The Broadcast News corpus is an excellent test case for this because the evaluation set is divided into different focus conditions, allowing comparison of recognition of planned speech to that of spontaneous speech. In the previous chapter, it was demonstrated that static pronunciation models improved recognition results in all focus conditions. In contrast, dynamic models seem to improve the non-F0 conditions (*i.e.*, focus conditions other than planned studio speech), particularly spontaneous speech and speech with background music. Unlike static dictionaries, performance of dynamic models are more dependent on the acoustic model used: rescoring with trees not matched to the original static dictionary (used to generate lattices) does not produce good results.

Several extensions that I thought would be useful did not improve decoding. As mentioned above, questions remain about the robustness of speaking rate and word predictability features, since trees using only segmental context outperformed trees using all features. Attempts to integrate the pronunciation model earlier in the decoding by rescoring lattices instead of $n$-best lists also proved to be ineffectual compared to just rescoring the top 100 hypotheses. I still believe, though, that both of these ideas should help recognition, particularly because of the pronunciation analyses in Chapter 3. Future research will be needed to refine the experiments performed here, as addressed in the next chapter. For instance, the confidence-based pronunciation evaluation that Gethin Williams and I developed for static dictionaries (see Section 5.3.5) may help determine when the extra-segmental features would be more robust.

Dynamic dictionary rescoring has provided a small decrease in word error rate on the Broadcast News corpus, in the range of 2% relative improvement over an improved static dictionary; the total improvement compared to the ABBOT96 dictionary was 4-5% relative depending on the acoustic models and test set used for evaluation. While I am slightly disappointed that the final improvement was not larger, historically, progress on the Broadcast News corpus has usually been made in small increments. These studies are intended to guide future modelers toward a better model of the dynamic variation of pronunciations seen in continuous speech.

# Chapter 7

# Conclusions

## 7.1 Summary

The field of ASR pronunciation modeling has been enjoying a renaissance in recent years, particularly with system developers tackling the difficult problem of recognizing spontaneous speech. Understanding phonetic variation has become a priority for many researchers; this topic has engendered a workshop [Strik *et al.*, 1998], several projects within workshops dedicated to the improved recognition of spontaneous speech [Ostendorf *et al.*, 1997; Weintraub *et al.*, 1997; Riley *et al.*, 1998], and a renewed interest in linguistic transcription of speech [Greenberg, 1997b]. This thesis contributes to the field in two ways: first, I have analyzed pronunciations in the Switchboard corpus, finding correlations between pronunciation differences, speaking rate, and word predictability; second, the latter half of the thesis shows how building context, in its various forms, into static and dynamic ASR pronunciation models can improve performance in speech recognizers.

### 7.1.1 Linguistic analyses

Chapter 3 was devoted to the study of pronunciation statistics in the Switchboard corpus, examining both general trends across an entire corpus and the statistics of individual linguistic segments. I investigated the effects of speaking rate and two measures of word predictability on pronunciations in the corpus.

Throughout the corpus as a whole, it became clear that variations in speaking rate and predictability correlate with pronunciation differences: the percentage of phones pronounced according to dictionary form decreases with increased speaking rate. Faster speaking rate and high unigram probability of the word also corresponded to an increase in the entropy of phone distributions. On the other hand, not all linguistic units were equally affected by these factors. Investigations of individual syllables and words sometimes revealed idiosyncratic patterns of reduction due to word predictability and speaking rate. This suggested that modeling syllables and words, rather than phones, may be appropriate in a dynamic pronunciation dictionary.

In another important finding, the investigated features did not independently influence pronunciation phenomena. Word frequency had an impact on how much pronunciation varied with variations in speaking rate: syllables selected from high-frequency words showed the largest pronunciation differences corresponding to differences in speaking rate. Statistics for the probability of canonical syllable pronunciations were also dependent on syllable structure and stress. These correlations indicate that these features should be included jointly within an ASR pronunciation model.

The phonetic transcription of the Switchboard corpus sometimes indicated severe phonological reductions, in which a word had little or no phonetic data associated with it — resulting in a phonetic segmentation with a completely "deleted" word, despite the fact that the word could be heard in the transcription. Cues from timing and feature spreading may enable listeners to comprehend the non-phonetic words. These examples, however, occurred in very restricted syntactic environments, suggesting that higher-level constraints facilitate identification of these words. Most of these examples could be modeled either with the use of $n$-gram statistics, or by modeling frequent collocations (word pairs). This suggested that including both $n$-gram statistics and neighboring word identities in ASR pronunciation models could improve recognizer performance.

### 7.1.2 Recognizer analyses

Chapter 4 presented an analysis of two speech recognizers, with the aim of determining the effect of speaking rate and word predictability on pronunciations in the context of ASR system. One recognizer was a Gaussian HMM system trained on the Switchboard corpus, the other a Hybrid Neural Network/HMM system trained to recognize Broadcast News utterances. For both systems, I studied the matching of the recognizer pronunciation model to a phonetic transcription, contrasting recognizer performance in conditions for which the transcription agreed with the recognizer dictionary against conditions for which the model mismatched the transcription.

Switchboard recognizer performance decreases in cases where the transcription does not match the canonical (dictionary) pronunciation. This can be devastating for overall performance because words in this corpus are pronounced canonically only 33% of the time. When errors are categorized by word predictability, the recognizer performs more poorly for very unlikely words than for likely words. However, errors attributable to the pronunciation model occur only when words are likely (log unigram probability $>$ -4 or log trigram probability $>$ -3); for unlikely words, canonically and non-canonically pronounced words are recognized with similar error rates.

Speaking rate also plays a significant role in determining recognizer errors. The accuracy of the Switchboard recognizer decreases by 14% absolute when performance for the fastest speech is compared against performance for slow speech; this trend is consistent with findings in other corpora [Mirghafori *et al.*, 1995; Siegler and Stern, 1995; Fisher, 1996a]. Some of this increase in error is attributable to the pronunciation model; the difference in error rates between canonically and non-canonically pronounced words for fast speech is twice that of slow speech.

An analysis of the SPRACH Broadcast News recognizer found that many of the pronunciation model error patterns were similar to that of the Switchboard recognizer, despite the fact that an automatic phone transcription was used to determine ground truth instead of hand transcriptions. This bodes well for automatic pronunciation learning systems, since lessons learned from analysis of linguistically based transcriptions may help improve automatically derived models. Speaking rate affected recognizer performance in this corpus as well, although increased recognition errors were seen in both the slow and fast extremes, in contrast to the Switchboard system, which performed well on slower speech. The other major difference between the Switchboard and Broadcast News recognizers was the percentage of canonically pronounced words depending on word frequency; in Switchboard, more frequent words were less likely to be canonical, whereas the opposite was true for Broadcast News. This reflected the bias of the training set on acoustic models: the likely words were seen more frequently in the training set, so in the automatic transcription of Broadcast News the acoustic models were more likely to produce canonical dictionary transcriptions when these words were presented in the test set. Since the Switchboard corpus was phonetically transcribed by linguists, the transcription of that corpus did not reflect this bias.

Evaluating recognizers on the Broadcast News corpus allowed for a comparison of planned and spontaneous speech. When error rates for non-canonical pronunciations were compared to error rates for canonical pronunciations, the decrease in accuracy due to non-canonical pronunciation was much larger in the spontaneous speaking style. Acoustically noisy conditions, such as telephone speech, also induced the acoustic models to produce a more non-canonical phonetic transcript, contributing to pronunciation model error in this corpus.

Thus, pronunciation model mismatches with transcribed data are a significant source of errors in speech recognition systems. This corroborates the findings of McAllaster *et al.* [1998]; by simulating acoustic data from both dictionary alignments and phonetic transcriptions of the Switchboard corpus, they found that the pronunciation dictionary was a significant source of errors in their spontaneous speech system. By including speaking rate and word predictability in my analysis, I was able to pinpoint some of the conditions under which the pronunciation model performed poorly.

### 7.1.3 Building better dictionaries

Chapters 5 and 6 were concerned with improving the dictionaries in our Broadcast News recognizer, taking into account some of the lessons learned in Chapters 3 and 4. In the first set of experiments, the baseline static dictionary was improved by automatic pronunciation learning techniques. The second set of experiments focused on dynamically selecting the appropriate pronunciation model based on an extended concept of context, including segmental context, speaking rate, word duration, and word predictability.

For static dictionaries, instead of learning pronunciations directly from phone recognition, it was beneficial to constrain recognition using decision trees, by means of smoothed phone recognition. Decoding with the constrained dictionary was slightly better than with the plain phone recognition dictionary in terms of word error rate; perhaps more importantly, the new dictionary did not increase decoding time as much. Even with this speedup, though, the decode time was 3.3 times as long as decoding with the baseline dictionary, so pruning techniques based on word frequency were used to select the appropriate pronunciations, resulting in slightly better accuracy and a large improvement in decoding time (146% better than the unpruned dictionary). Confidence measures also proved useful for model selection, providing an additional gain in accuracy and speed, as well as for the evaluation of new baseforms generated by letter-to-phone decision trees.

Using the pronunciation transcripts generated by smoothed phone recognition, I trained decision trees to predict the pronunciations of individual syllables and words, based on the extended context. The motivation for constructing these longer-term models (as opposed to constructing phone models) was to allow the decision tree learning algorithm to focus on finding pronunciation variations at syllable or word boundaries, rather than the less frequent syllable- or word-internal variations. This was an effective strategy, as syllable models achieved a better test set log probability and test set pronunciation coverage than comparable phone models. Both syllable and word trees showed large increases in the log probability of pronunciations in the test set; most of the gain can be attributed to word, syllable, and phone context, but extra-segmental features did also provide some benefit.

When word and syllable trees were employed in $n$-best rescoring, the results were equivocal. Initial experiments were somewhat promising; syllable d-trees trained on 100 hours of Broadcast News data were slightly better than the baseline (0.4% absolute), although this gain was due almost completely to improvements in the spontaneous focus condition (1.4% absolute) and speech with background music condition (0.9% absolute). Word d-trees did not perform as well in this experiment, decreasing word error rate by only 0.2%. The tables were turned, however, when the full 200 hour training set was used — word trees gained 0.3-0.6% over the baseline (depending on the test set) compared to 0-0.5% gain for the syllable trees. The disappointing part of this result was that the trees performed best when the extra-segmental features were removed from consideration during training, perhaps indicating that these features are not currently robust enough for accurate prediction of pronunciation probabilities in an automatic learning system. The dynamic rescoring results are disappointing from another perspective: experiments by Saraclar [1997] and McAllaster and Gillick [1999] indicate that a large reduction in word error (on the order of 50%) is possible if the correct pronunciation models are employed at the right time. In the next section, I will examine possible directions for future work that may come closer to achieving this goal.

Because the test sets for the Broadcast News corpus were subdivided into different focus conditions, I was able to evaluate the performance of the new static and dynamic pronunciation models on different types of speech. In the final test set, both static and dynamic dictionaries improved recognition only in the non-planned speech conditions, including a 12% relative error reduction in the telephone speech focus condition. This last result indicates that the automatically derived pronunciation models are probably capturing acoustic model variability due to channel differences; it may also indicate improved acoustic modeling for that condition, since one of the MLP acoustic models was trained with 8kHz band-limited data to better model speech from this focus condition. Dynamic rescoring of word d-trees showed the largest improvement over static dictionaries in the FX focus condition of the Broadcast News test set, which contained the difficult-to-classify (and also difficult-to-recognize) speech, improving recognition in this class by 1% absolute over the static dictionary.

## 7.2 Discussion and future directions

The disparity between improved performance of decision tree classifiers and the lack of large improvements when these models are employed in dynamic rescoring of $n$-best lists is puzzling. One would believe that improving the log probability of pronunciations in a test set would decrease the word error rate by a larger amount. These results also seem at odds with those of both Saraclar [1997] and McAllaster and Gillick [1999]; through best-case-scenario studies they suggest that improved pronunciation models *should* bring much lower word error rates.

This disparity in results is not without parallel in the speech recognition literature, however. Improvements in language models (the probability of a word sequence $P(M)$) are often reported in terms of reduction of perplexity, a measure related to the average log probability metric I used to evaluate pronunciation models. Sometimes, even with large decreases in perplexity, improved language models do not change the word error rate much.[1] The decoupling between perplexity and word error rate has been well documented by ASR researchers; it appears that this phenomenon applies to pronunciation modeling as well as to language modeling.

This parallel behavior between pronunciation and language models may allow some insight into the reasons behind the disassociation of these metrics. Pronunciation models and language models are dissimilar from acoustic models in that they are in some sense *a priori* models of language: codified top-down knowledge systems that say "under these conditions, this model is expected to apply." Because of the way most speech recognizers decompose the probability model ($\text{argmax}\, P(M|X) = \text{argmax}\, P(X|M)P(M)$), these models do not depend on the acoustic data; thus, they are lacking in bottom-up language constraints. Perhaps this lack of dependence of *a priori models* (language and pronunciation models) on acoustic data contributes to the disparity between perplexity (which is independent of acoustics) and word error rate (which depends on the acoustics). It is not clear how the dependency on acoustics can be implemented in a language model without radically changing the models.[2] An argument against this position is that the word and

[1] An informal survey of language modeling papers in some of the most recent speech recognition conferences (ICASSP 1997, 1998, and 1999) reveals that in some studies, improved perplexity accompanied improved word error rate. In others, the word error rate was almost unchanged. A third set of studies did not report the word error rate.

[2] The REMAP framework of Konig *et al.* [1995] decompose the probability model by setting $P(M|X) =$

syllable d-trees in this thesis *did* have some dependence on local acoustic measures (*e.g.*, the mrate speaking rate estimate), yet a difference in perplexity and word error results persisted; it could be that there was not enough dependence on acoustics in this model to matter in evaluation, or that the above hypothesis that acoustic data makes the difference between perplexity and word error rate is incorrect. Since all of the models within a speech recognizer must interact, it may be the case that we do not understand how to integrate the information they provide correctly; also, improvements in the pronunciation model may be masked by errors made in other components of the system.

Further insight may be gained by comparing the work in Chapter 6 to other dynamic pronunciation models in the literature. Both Weintraub *et al.* [1997] and Riley *et al.* [1998] constructed stream-transformation models to predict pronunciations on a phone-by-phone basis for the Switchboard corpus; the primary difference between the two experiments was that Weintraub *et al.* used phone recognition to determine the phonetic training data, whereas Riley *et al.* bootstrapped from the hand-transcribed labels provided by the Switchboard Transcription Project.[3] Besides the orientation toward phones rather than syllables and words, the main difference between these studies and the models in Chapter 6 was the inclusion in my study of extra-segmental factors in the word and syllable models, although these factors were not used in the final results (see pp. 137–142). For both Weintraub *et al.* and Riley *et al.* [1998], the addition of dynamic rescoring with d-tree pronunciation models in essence changed only pronunciations at word boundaries due to the context of the surrounding words, as word-internal pronunciation variation could be captured by compilation into a static dictionary (as in Chapter 5). Both studies found that dynamic rescoring did not improve word accuracy over static dictionaries;[4] Weintraub *et al.* found that dynamic rescoring degraded results (-0.3%), whereas Riley *et al.* saw a very small increase in accuracy (0.05%) when cross-word models were used. This suggests that the small improvement in the word trees in my experiments may be due to the cross-word models acting as multi-word models, which Riley *et al.* found to be effective in the Switchboard domain.

Another successful approach to dynamic pronunciation modeling has been that of Finke and Waibel [1997b], who implemented a system that incorporated variables correlated with speaking mode. The backbone of their system was a set of probabilistic phonological rules, similar to the rules used by Tajchman *et al.* [1995b],[5] but expanded to include common multi-word phrase rules (*e.g.*, *going to* ⇒ *gonna*). The main difference between the systems was that Tajchman *et al.* computed the rule probabilities unconditionally, whereas Finke and Waibel estimated rule probabilities by decision trees; probability distributions depended on both the word context features (including word identity and unigram probability) and extra-segmental features such as speaking rate and word duration. Including word context dependencies in phonological rule probability estimation improved error rate by 1.6% on the Switchboard corpus (28.7% to 27.1%); adding conditioning based on extra-segmental

$\sum_Q P(Q|X)P(M|Q,X)$. This decomposition does allow the language and pronunciation models to depend directly on the acoustics.

[3] There were other differences between the models, *e.g.*, the encoding scheme for phonemes in the system.

[4] Both studies *did* achieve improvements using an improved static dictionary, obtaining roughly a 1% absolute improvement.

[5] See Chapter 2, page 22 for the list of rules used by Tajchman *et al.*

features improved error rate by another 0.4%.[6]

The primary difference between the Finke and Waibel system and the dynamic trees in Chapter 6, therefore, is what decision trees are estimating (probabilities of phonological rules, or probabilities of word or syllable pronunciations). Finke and Waibel's [1997b] d-tree phonological rule system limits the possible pronunciation variations allowed in the model using pre-compiled linguistic knowledge of variations; thus, it may not capture some of the acoustic model variation that is modeled by automatically induced d-tree rules. What the phonological rule system gains is generality across words — rules can be word-dependent, but can also apply across all words. Estimating rule probabilities in this way pools the training data across many words, which can improve probability estimates.

That the phonological rule system works well for improving pronunciation models suggests that incorporating top-down linguistic variations, rather than learning them automatically, may provide better results, at least for spontaneous speech. Since automatically derived models provided the best performance in acoustically noisy conditions, another possibility is to implement a hybrid strategy: use phonological rules for spontaneous speech, and automatically derived models for noisy speech. If the focus condition of the segment being decoded can be estimated from the acoustic signal (*e.g.*, determining telephone speech by the bandwidth of the signal), then one may be able to switch between these models, or merge them in some fashion.

It is encouraging that Finke and Waibel's system was able to use extra-segmental variables to improve pronunciation models. Their system included slightly different features than did my system, such as deviation from the mean word duration and fundamental frequency; these could be integrated into the syllable and word trees. Other features that might merit inclusion are the spectral energy (both local and global), as well as estimates of acoustic stress [Silipo and Greenberg, 1999]. Some of the context features in the syllable and word d-trees may also not be robust enough for use in a pronunciation model; future studies should examine the utility of different variables.

One of the main premises of this thesis was modeling the variation of phone pronunciations within syllable and word models. The disconnection between perplexity-based evaluations of pronunciation models and speech recognition results may suggest that the phone is too coarse of a unit to be modeling this type of variation. Recent work by Finke *et al.* [1999] and Kirchhoff [1998] suggests that using phonetic features may allow for finer-grained control of models of pronunciations. In the model of Saraclar *et al.* [1999], pronunciation variations are described at the HMM state level (through the sharing of Gaussian distributions) rather than at the phone level. These proposals blur the line somewhat between acoustic models and pronunciation models; perhaps the integration of these two models may be necessary to better capture pronunciation variation.

A lesson learned from both the linguistic investigations of Chapter 3 and the

[6] Finke and Waibel compared a system that used rule probabilities conditioned on the words ($P(r|w)$) against a system that used rules without probabilities. One wonders how much of the 1.6% improvement is due to inclusion of word context, and how much is due to having probabilities on the rules. Unfortunately, they did not report on a system that just used rules with unconditional probabilities ($P(r)$), although it appears that classification of pronunciations was improved only slightly by the inclusion of word context [Finke and Waibel, 1997b, figure 1].

decision tree examples presented in Chapter 6 was that there is a significant interaction among extra-segmental factors. In addition, some of the real-valued features (as opposed to categorical features) varied smoothly with pronunciation variation — that is, there were sometimes no natural partition points for determining pronunciation distributions. This suggests that decision trees, which recursively partition the training set, may not be the best classifier for this task. Instead, neural net classifiers (as used by Fukada *et al.* [1999]) may allow for more natural integration of real-valued attributes into the model.

The experiment that integrated acoustic confidence measures in the static dictionary construction (Chapter 5) demonstrated that changing the baseform selection scheme can improve results. A future direction for decision tree pronunciation models is to select the pronunciations for each word based on confidence scores; this idea could be extended by selecting d-tree partitions based on overall confidence scores rather than entropy-based partitioning. Other possible techniques for model selection would be to maximize a phonetic distance between baseforms to increase the span of pronunciations for each word [Holter and Svendsen, 1998], or to maximize discriminability between pronunciations of different words, in hopes of reducing confusability among words.

The results in this thesis have been achieved using mostly context-independent acoustic models. Dynamic rescoring using d-tree pronunciation models should also be tried at some point with context-dependent acoustic models, to see if they would still give an improvement over baseline models. The static dictionary created in this thesis has been used with a Context-Dependent Neural Network (CDNN) acoustic model as part of the SPRACH system, although a comparison between the old ABBOT96 dictionary and the new SPRACH dictionary was not carried out with the context-dependent models. It has been noted by some researchers that context-independent systems often show more improvement with better pronunciation modeling than do context-dependent systems, although these observations were made mostly about Gaussian HMM recognizers. Since neural net acoustic models implement context-dependency in a different way than "traditional" systems [Bourlard *et al.*, 1992; Franco *et al.*, 1992; Kershaw *et al.*, 1996], it is not clear that this postulate will hold true for CDNN models.

## 7.3 Final thoughts

In this work, I have improved pronunciation models for automatic speech recognition. This thesis contributes to the linguistic literature an analysis of pronunciations within a large corpus of spontaneous speech; in particular, word predictability and speaking rate were shown to be good predictors of pronunciation variability, confirming and extending other studies of this phenomena. Within the speech recognition realm, this thesis contributes a detailed study of the effects of pronunciation model errors on recognition performance, and an examination of automatic pronunciation learning techniques that improve performance a relative 4-5% for a large speech corpus. It also raises the question of how to evaluate pronunciation models: standard techniques for judging model quality do not necessarily translate into improvements in word error rate.

As people increasingly use speech as a natural interface to computers, the speech recognition community will be faced with the challenge of recognizing more and more natural, spontaneous speech in a wide array of acoustic conditions. These factors increase the variability of the ways that people speak, as well as the ways that recognizer acoustic models respond to speech input. The work in this dissertation has taken a small step toward the goal of creating a better model for this variability.

# Appendix A

# Phone symbols used in ASR

The following are the set of symbols used in this thesis. The ASCII version of the set is based on the symbol set of TIMIT [Garofolo *et al.*, 1993], a hand-labeled read-speech database commonly used in the ASR community. The IPA equivalent of each phone is also listed.

| ASR Phone Symbols | | | | | |
|---|---|---|---|---|---|
| TIMITset | IPA | Example | TIMITset | IPA | Example |
| pcl | p$^{o}$ | (p closure) | bcl | b$^{o}$ | (b closure) |
| tcl | t$^{o}$ | (t closure) | dcl | d$^{o}$ | (d closure) |
| kcl | k$^{o}$ | (k closure) | gcl | g$^{o}$ | (g closure) |
| p | p | **p**ea | b | b | **b**ee |
| t | t | **t**ea | d | d | **d**ay |
| k | k | **k**ey | g | g | **g**ay |
| q | ʔ | ba**t** | dx | ɾ | dir**t**y |
| ch | t͡ʃ | **ch**oke | jh | d͡ʒ | **j**oke |
| f | f | **f**ish | v | v | **v**ote |
| th | θ | **th**in | dh | ð | **th**en |
| s | s | **s**ound | z | z | **z**oo |
| sh | ʃ | **sh**out | zh | ʒ | a**z**ure |
| m | m | **m**oon | n | n | **n**oon |
| em | m̩ | botto**m** | en | n̩ | butto**n** |
| ng | ŋ | si**ng** | eng | ŋ̩ | Washi**ng**ton |
| nx | ɾ̃ | wi**nn**er | el | l̩ | bott**le** |
| l | l | **l**ike | r | r | **r**ight |
| w | w | **w**ire | y | j | **y**es |
| hh | h | **h**ay | hv | ɦ | a**h**ead |
| er | ɝ | b**ir**d | axr | ɚ | butt**er** |
| iy | i | b**ee**t | ih | ι | b**i**t |
| ey | e | b**ai**t | eh | ɛ | b**e**t |
| ae | æ | b**a**t | aa | ɑ | f**a**ther |
| ao | ɔ | b**ou**ght | ah | ʌ | b**u**t |
| ow | o | b**oa**t | uh | ϖ | b**oo**k |
| uw | u | b**oo**t | ux | ü | t**oo**t |
| aw | ɑ$^{w}$ | ab**ou**t | ay | ɑ$^{y}$ | b**i**te |
| oy | ɔ$^{y}$ | b**oy** | ax-h | ə̥ | s**u**spect |
| ax | ə | **a**bout | ix | ɨ | deb**i**t |
| epi | | (epenthetic sil.) | pau | | (pause) |
| h# | | (silence) | | | |

# Appendix B

# Syllable and Word D-tree Examples

Chapter 6 describes the process of building word and syllable decision trees for the Broadcast News corpus, and presents a few example trees. In this appendix, I present automatically-derived trees that illustrate various linguistic phenomena; these trees were not included in the main chapter because of space considerations. The first section describes word-based d-trees; syllable trees are presented in the second section. Table B.1 lists the values of the categorical features used in the d-trees, as described in Section 5.3.1; Table B.2 lists some abbreviations found in these examples.

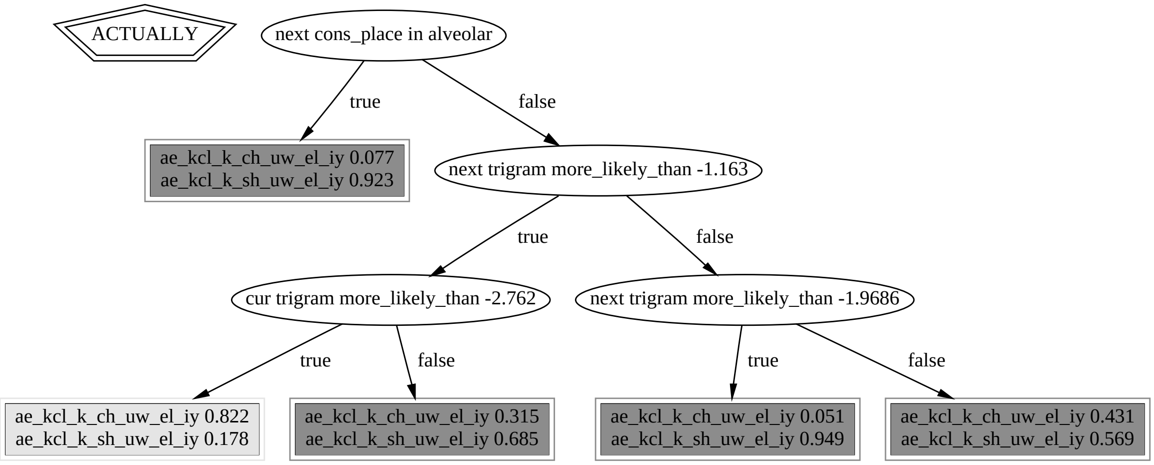
| | |
|---|---|
| Stress | unstressed, primary, secondary, unknown |
| Consonant manner of adjacent phone | voiceless stop, voiced stop, silence, approximant, syllabic, voiceless fricative, voiced fricative, nasal, n/a (=vowel/pause) |
| Consonant place of adjacent phone | labial, dental, alveolar, post-alveolar, palatal, velar, glottal, n/a (=vowel/pause) |
| Vowel manner of adjacent phone | monophthong, w-diphthong, y-diphthong, n/a (=consonant/pause) |
| Vowel place of adjacent phone | cross between height (high, mid-high, mid-low, low) and frontness (front, mid, back), or n/a (=consonant/pause) |
| Phone ID | aa, ae, ah, ao, aw, ax, axr, ay, b, bcl, ch, d, dcl, dh, dx, eh, el, em, en, er, ey, f, g, gcl, h#, hh, hv, ih, ix, iy, jh, k, kcl, l, m, n, ng, ow, oy, p, pcl, r, s, sh, t, tcl, th, uh, uw, v, w, y, z, zh |

Table B.1: Values of some categorical attributes in decision trees

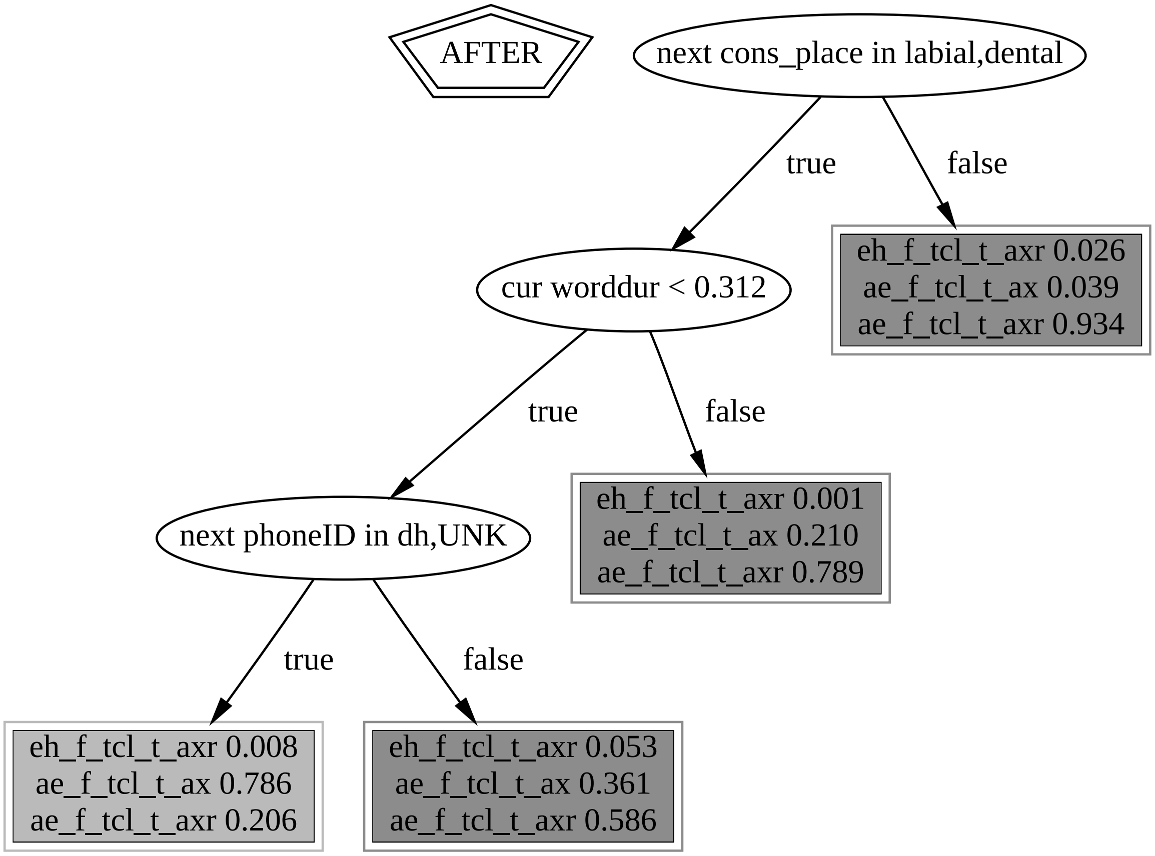
| | | |
|---|---|---|
| cur | = | current |
| len2prevpause | = | time since the previous hypothesized pause |
| localrate | = | speaking rate from duration of three words in the current hypothesis being rescored |
| mrate | = | interpausal speaking rate from the mrate measure |
| phoneID | = | identity of the phone |
| prev | = | previous |
| ratediff | = | recograte-mrate |
| rateratio | = | mrate/recograte |
| recograte | = | interpausal speaking rate from best hypothesis in first pass recognition |
| sylID | = | identity of the syllable |
| trigram | = | trigram probability of the word |
| wordID | = | identity of the word |
| worddur | = | duration of the word |

Table B.2: Some abbreviations found in the tree examples in this appendix.

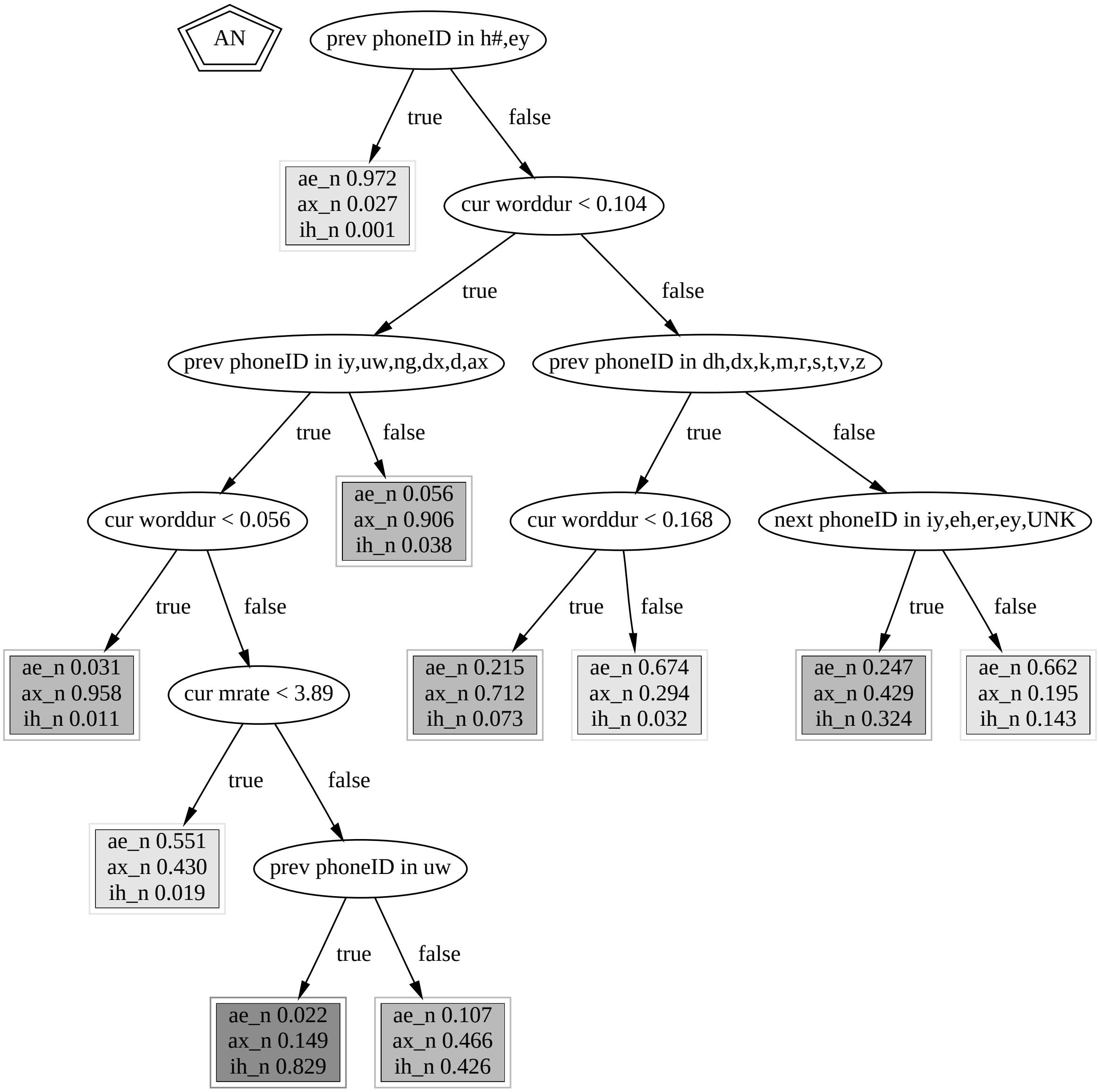
## B.1 Word trees


ACTUALLY
next cons_place in alveolar
true
false
ae_kcl_k_ch_uw_el_iy 0.077
ae_kcl_k_sh_uw_el_iy 0.923
next trigram more_likely_than -1.163
true
false
cur trigram more_likely_than -2.762
next trigram more_likely_than -1.9686
true
false
true
false
ae_kcl_k_ch_uw_el_iy 0.822
ae_kcl_k_sh_uw_el_iy 0.178
ae_kcl_k_ch_uw_el_iy 0.315
ae_kcl_k_sh_uw_el_iy 0.685
ae_kcl_k_ch_uw_el_iy 0.051
ae_kcl_k_sh_uw_el_iy 0.949
ae_kcl_k_ch_uw_el_iy 0.431
ae_kcl_k_sh_uw_el_iy 0.569

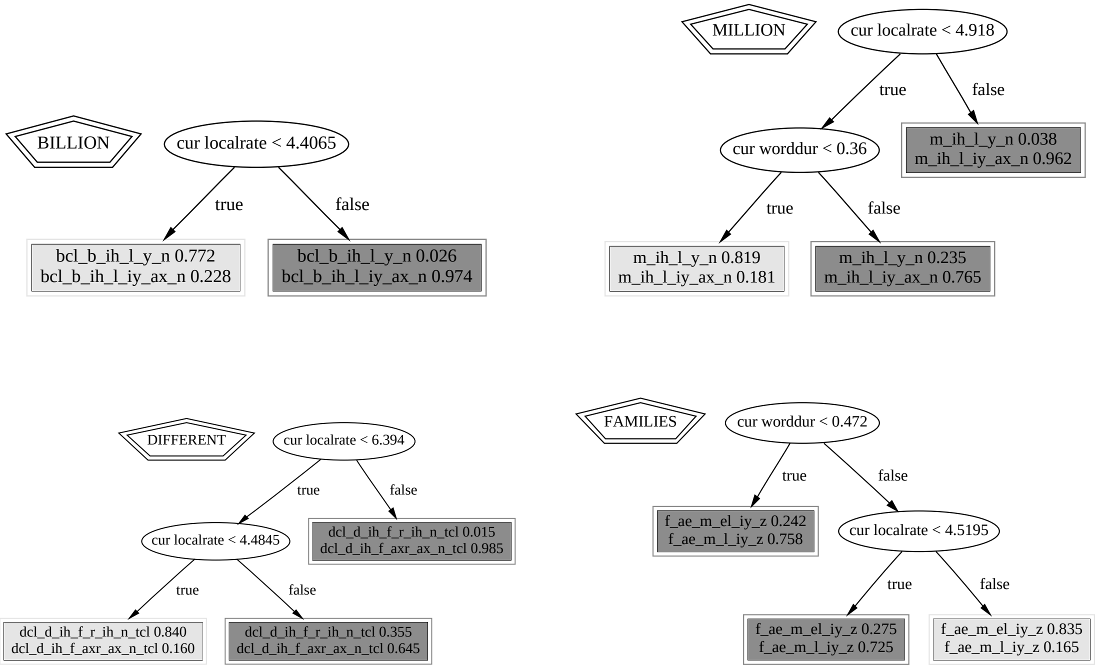
Example 1: The fact that a medial consonant in the word *actually* can be affected by context is unusual; typically one would expect pronunciations that vary medially to not depend on the segmental identity of a neighboring word. It is unclear why having a following alveolar consonant makes the `[sh]` pronunciation so likely.



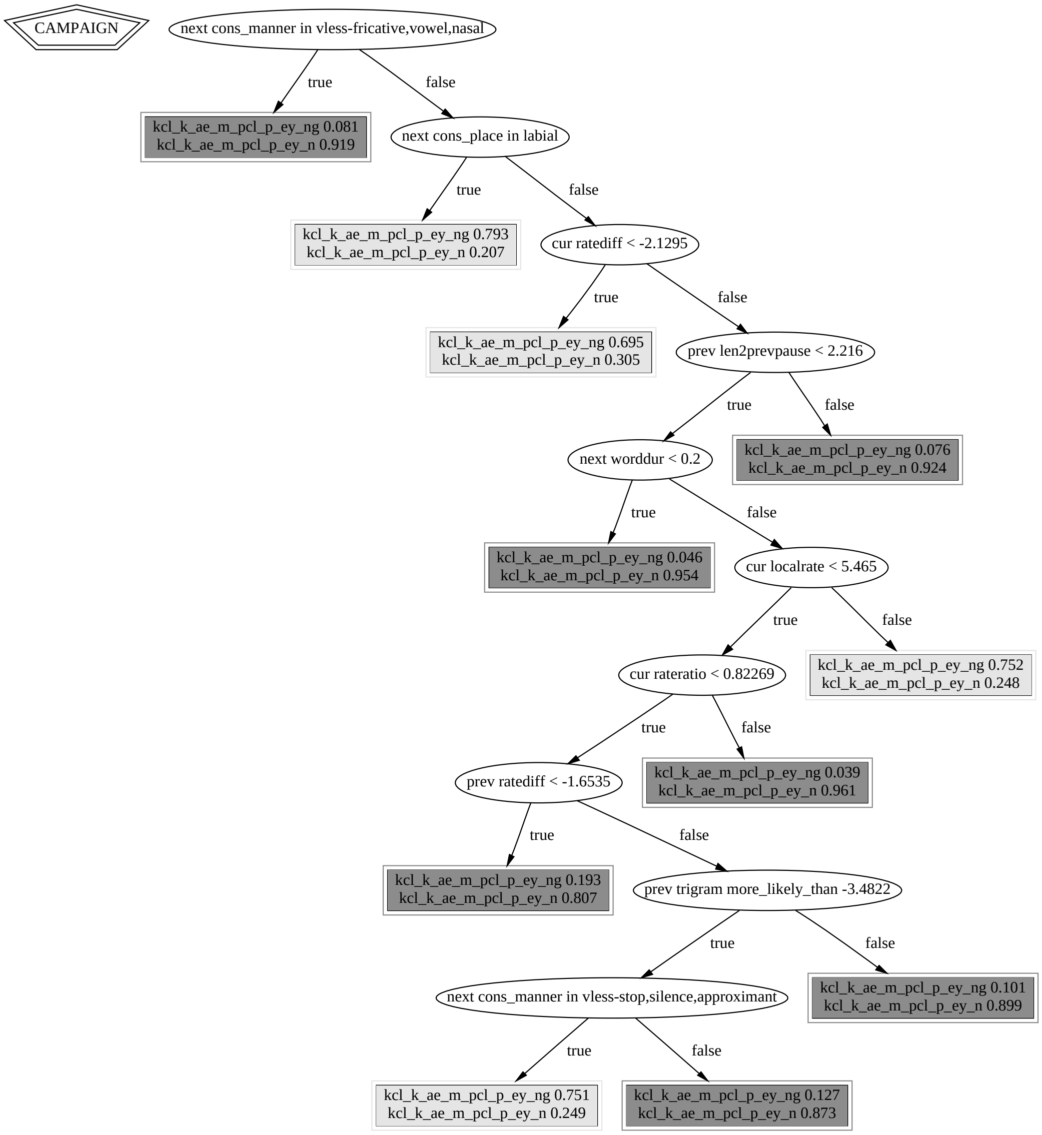
Example 2: The major pronunciation variation for the word *after* is the loss of the final r-coloration on the schwa, which seems to occur most frequently when the following word starts with a frontal consonant. The sub-selection for the phone `[dh]` may be a frequency effect: very few words actually start with this phone, but they are all frequent words (*the, this, they, that, their, them, those, they're, and these*). The unknown category in this case covers the phones `[p]`,`[m]`,`[n]`, and `[v]`, but the `[dh]` phone dominates these others, with about 95% of the training data.



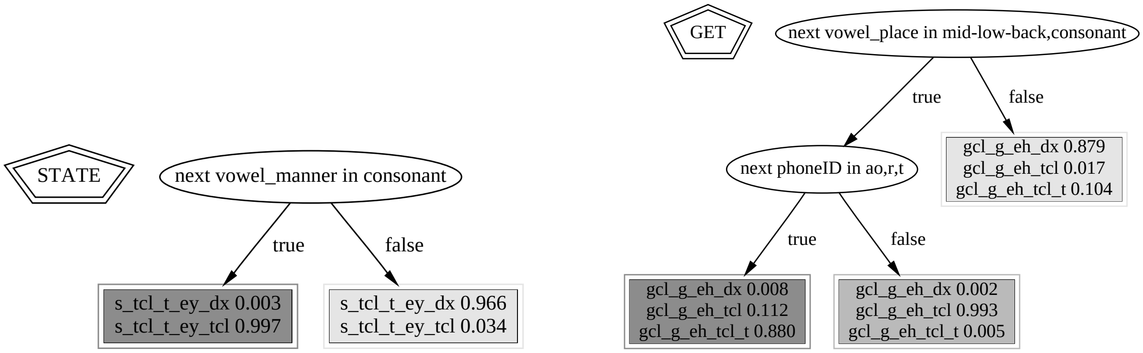
Example 3: For the word *an*, the major variation is in the quality of the vowel; there is one full pronunciation (`[ae_n]`), and two reduced pronunciations (`[ax_n]` and `[ih_n]`). Two major factors affect pronunciation choice: duration determines the choice of reduced/unreduced (longer examples are unreduced), and the adjacent phones determine the vowel height of reduced vowels. For example, in the lowest node on the tree, if the previous phone is `[uw]` (a high vowel), then the pronunciation `[ih_n]` is much more likely than `[ax_n]`.



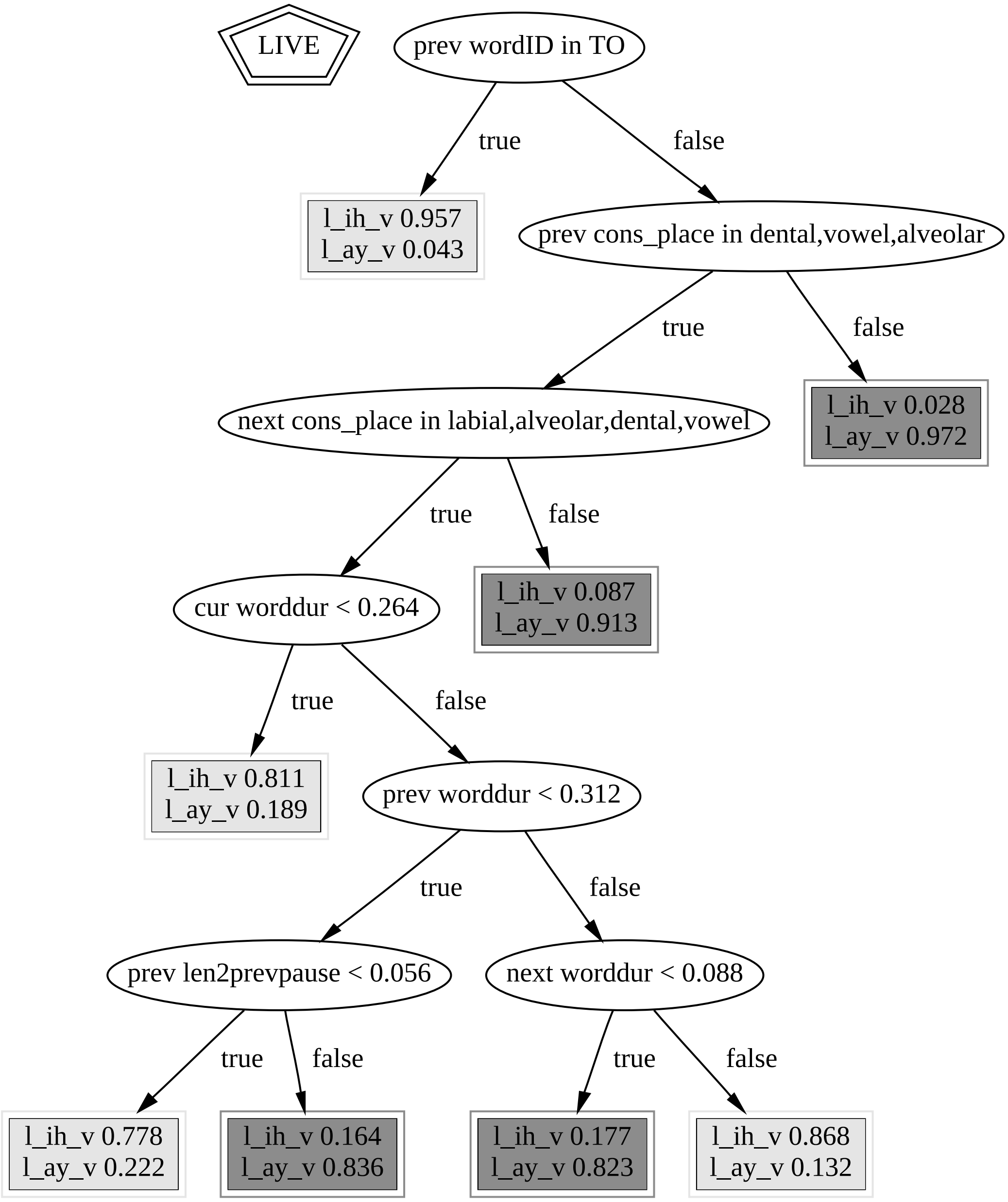
Example 4: These trees show the importance of how syllables are counted in determining rate. At first glance, the observation that the unreduced version is actually correlated with higher local speaking rate is counterintuitive. The localrate measure is computed by syllabifying the pronunciations in the first-pass hypothesis, so if more syllables are present (as in the unreduced variants) then there are more syllables per unit time, and therefore higher rates. It may have been a mistake to include the rate based on the first-pass hypothesis, since this will tend to favor the pronunciation selected by first-pass recognition. However, these trees do indicate that an speaking rate calculation that determines the actual number of syllables present (such as mrate), rather than the canonical number of syllables (given by recograte) may be very useful.


CAMPAIGN
next cons_manner in vless-fricative,vowel,nasal
true
false
kcl_k_ae_m_pcl_p_ey_ng 0.081
kcl_k_ae_m_pcl_p_ey_n 0.919
next cons_place in labial
true
false
kcl_k_ae_m_pcl_p_ey_ng 0.793
kcl_k_ae_m_pcl_p_ey_n 0.207
cur ratediff < -2.1295
true
false
kcl_k_ae_m_pcl_p_ey_ng 0.695
kcl_k_ae_m_pcl_p_ey_n 0.305
prev len2prevpause < 2.216
true
false
next worddur < 0.2
kcl_k_ae_m_pcl_p_ey_ng 0.076
kcl_k_ae_m_pcl_p_ey_n 0.924
true
false
kcl_k_ae_m_pcl_p_ey_ng 0.046
kcl_k_ae_m_pcl_p_ey_n 0.954
cur localrate < 5.465
true
false
cur rateratio < 0.82269
kcl_k_ae_m_pcl_p_ey_ng 0.752
kcl_k_ae_m_pcl_p_ey_n 0.248
true
false
prev ratediff < -1.6535
kcl_k_ae_m_pcl_p_ey_ng 0.039
kcl_k_ae_m_pcl_p_ey_n 0.961
true
false
kcl_k_ae_m_pcl_p_ey_ng 0.193
kcl_k_ae_m_pcl_p_ey_n 0.807
prev trigram more_likely_than -3.4822
true
false
next cons_manner in vless-stop,silence,approximant
kcl_k_ae_m_pcl_p_ey_ng 0.101
kcl_k_ae_m_pcl_p_ey_n 0.899
true
false
kcl_k_ae_m_pcl_p_ey_ng 0.751
kcl_k_ae_m_pcl_p_ey_n 0.249
kcl_k_ae_m_pcl_p_ey_ng 0.127
kcl_k_ae_m_pcl_p_ey_n 0.873

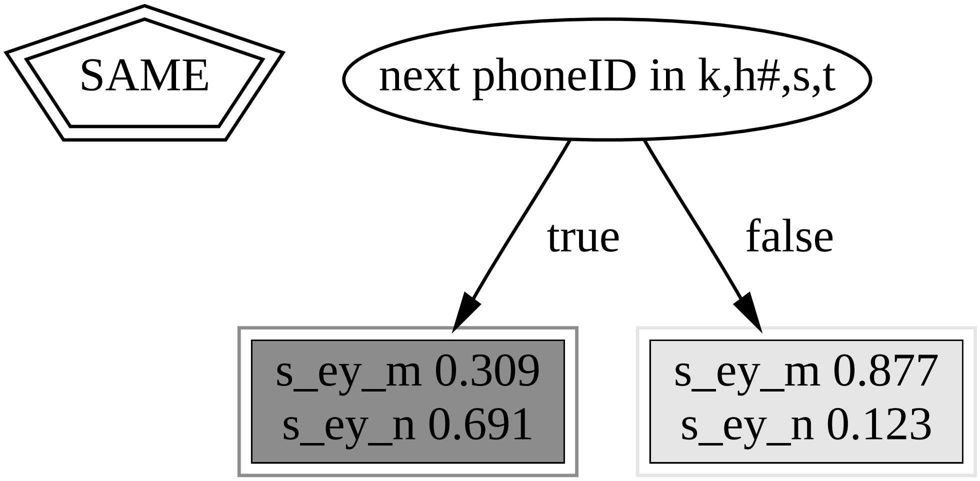
Example 5: The phones `[n]` and `[ng]` can sometimes be allophonic, as in this example tree for the word *campaign*. This variation may just be due to the nature of statistical modeling within the speech recognizer, but it is also possible that it is capturing a physiological and phonological effect. The tongue position after the penultimate phone `[ey]` is high in the mouth with the tip of the tongue lowered; substantial tongue movement would be required to reach the appropriate position for `[n]`, whereas a small tongue movement to a velar closure would produce `[ng]`. The speaker may be economizing effort in some situations, such as when the next word begins with a labial stop, as in *campaign promises*.



Example 6: These two trees (for *state* and *get*) display cross-word flapping phenomena; the realization of the final `[t]` depends on whether the next word starts with a vowel or not.



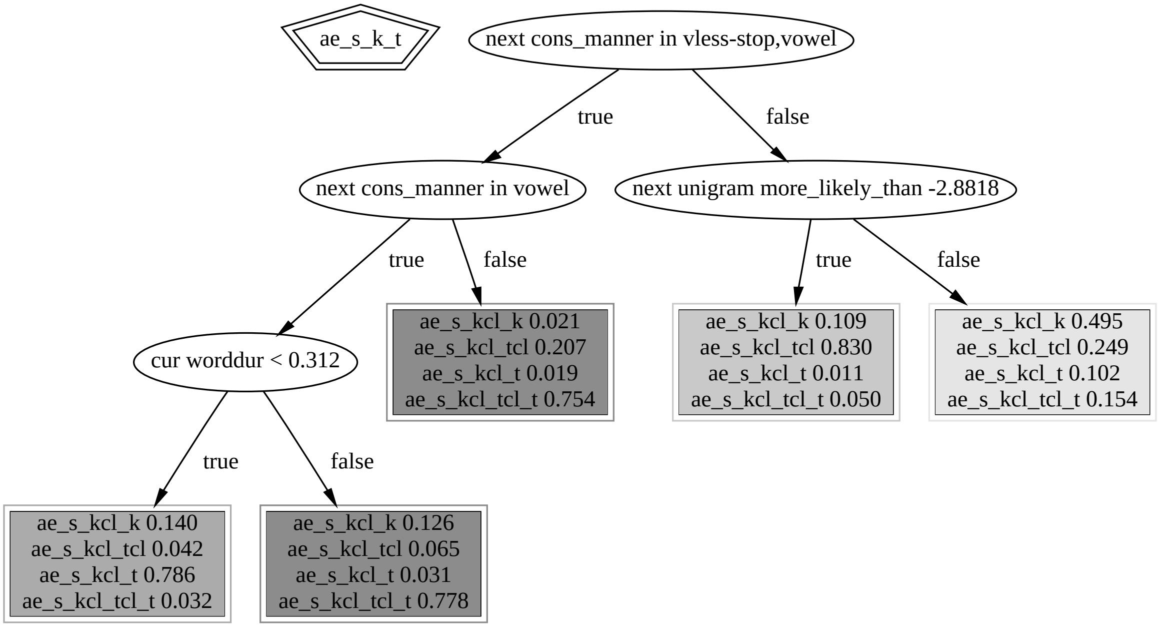
Example 7: This is an example where syntactic information would be a boon for disambiguating pronunciations. The verb *live* (`[l_ih_v]`) is pronounced differently than the adjective *live* (`[l_ay_v]`); the top node in the tree defines a syntactic environment in which only the verb form can occur (after the word *to*). If the d-tree learning process had notions of syntax, it is clear that the resulting tree would be more compact: modeling power would not be wasted trying to discriminate between these two parts of speech.



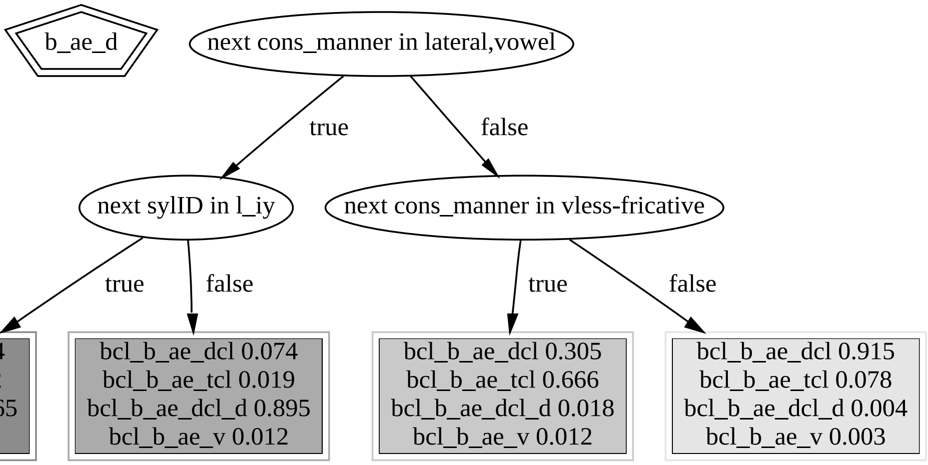
Example 8: This tree exemplifies the process of place assimilation, at least to some degree. The process of [m] being pronounced as [n] before [s] and [t] makes linguistic sense, since these latter phones all are alveolar or post-alveolar. However, it is not clear why [k] and [h#] (pause) are grouped in the same splitting condition.

## B.2 Syllable trees

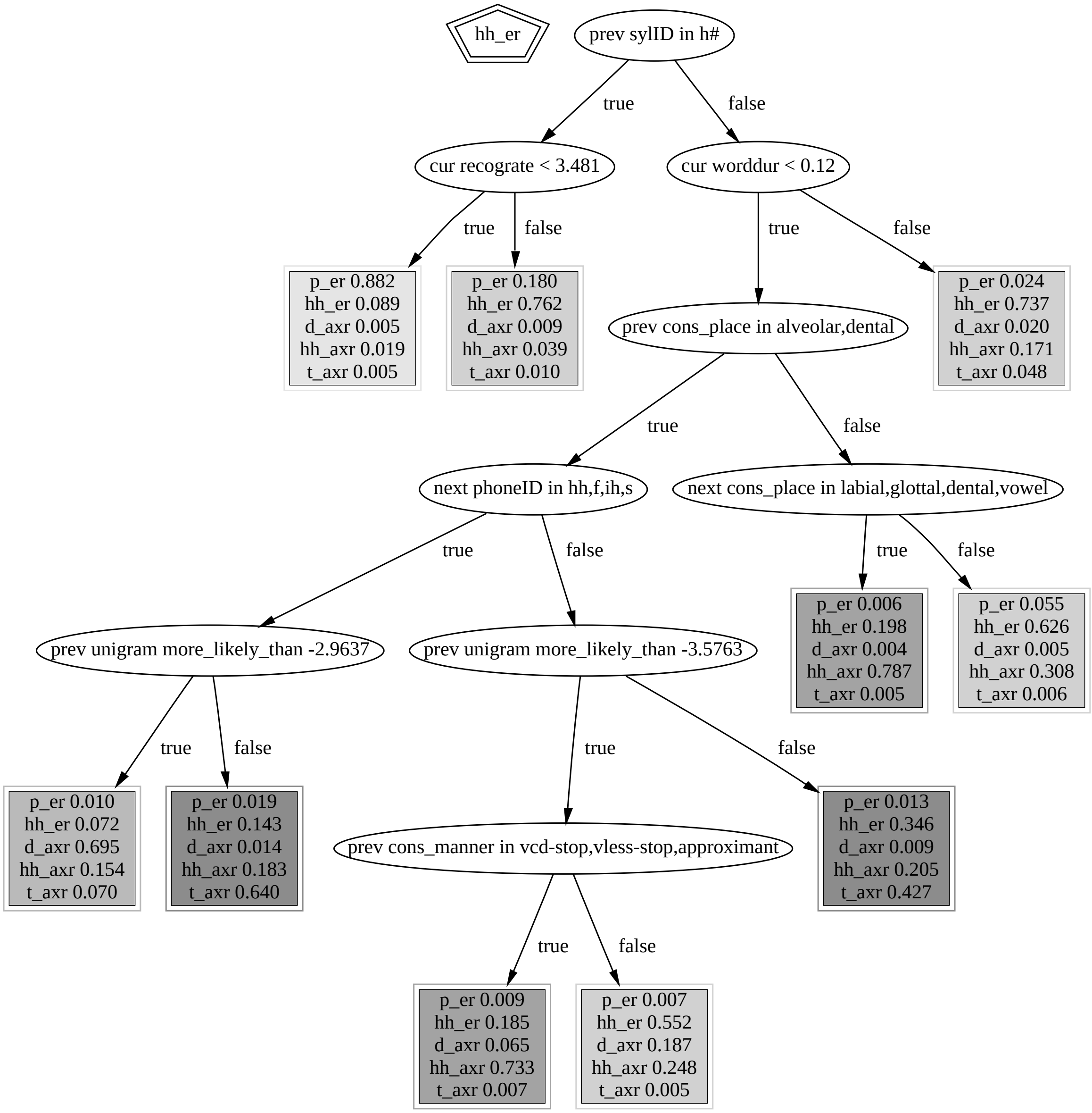
The following describes trees for a few individual syllables. Many of the trees show linguistic effects similar to those seen in the word trees, so fewer trees are displayed in this section.



Example 9: This tree corresponds only to the word *asked.* The stop burst for the final [t] is frequently omitted when the next word starts with a consonant (particularly if the next word is very likely), whereas if vowels follow the [t] burst is likely to occur.

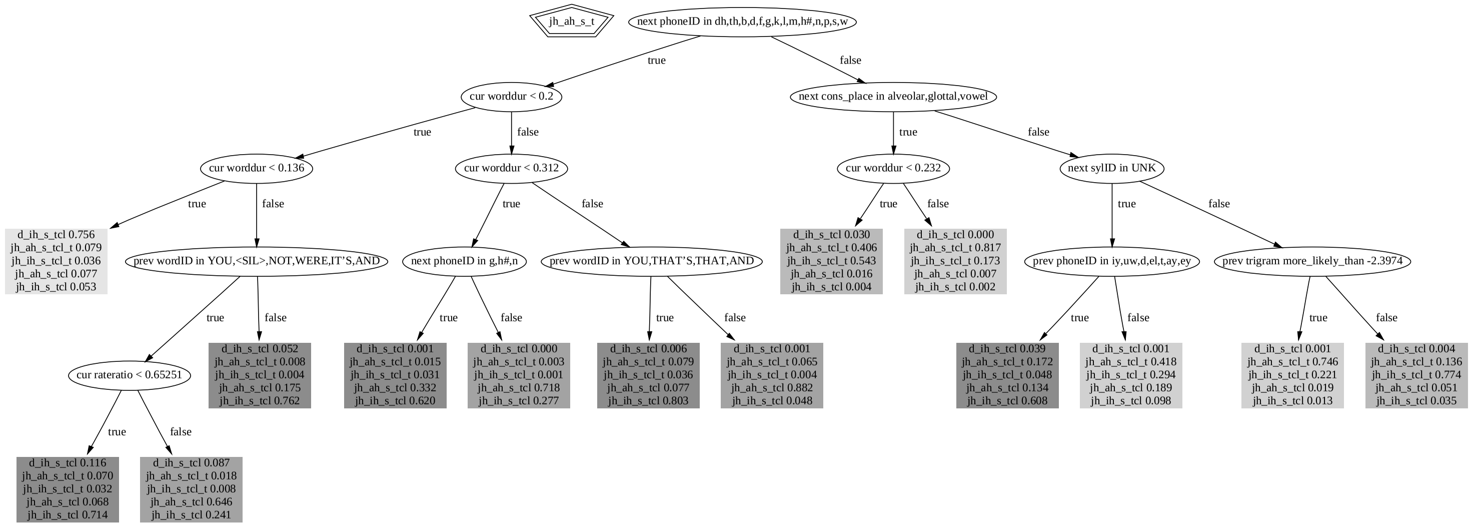


Example 10: This tree represents the common syllable in the words *Ahmedabad, bad, Badelain, badlands, badly, badminton, Carlsbad,* and *Sinbad.* There are two points to be made here: the right side of the tree shows devoicing of the final `[d]` before voiceless fricatives, and for some reason, the preferred pronunciation of *badly* is `[b_ae_v_l_iy]`.



Example 11: This tree represents a syllable in 89 different words. Two different linguistic effects are at play in this syllable. Since [hh] is an acoustically weak phone, the recognizer has a difficult time recognizing it— the features of previous phones may affect machine perception of the onset to this syllable. For instance, at the starts of phrases, [p] can be substituted for [hh], particularly in slow speech. If the previous word ends in an alveolar or dental phone, then [hh] may be realized as [t] or [d] more frequently. The second variability in pronunciation is the alternation between [er] and unstressed [axr]. Near the top of the tree (in the second rank on the right), we see that longer words[1] tend to have the full vowel; below that point in the tree it becomes difficult to disentangle the contextual dependencies.

[1] Longer words in this case will probably include all multisyllabic words and longer instances of *her*, since the division point at 0.12 seconds is roughly half the average syllable length.



Example 12: This is the syllable tree for `[jh_ah_s_t]`, which represents the common syllable in *adjust*, *adjustment*, *adjustments*, *just*, *justly*, *readjust*, *readjustment*, *unjust*, and *unjustly*. Most of the trees selected for these examples were rather small, which made them good examples for illustrating various linguistic phenomena. However, quite a few trees were as complex as this one, and were consequently difficult to analyze.

# Bibliography

M. Adamson and R. Damper. A recurrent network that learns to pronounce English text. In *Proceedings of the 4th Int'l Conference on Spoken Language Processing (ICSLP-96)*, Philadelphia, PA, October 1996.

G. Adda, L. Lamel, M. Adda-Decker, and J. Gauvain. Language and lexical modeling in the LIMSI Nov96 Hub4 system. In *Proceedings of the DARPA Speech Recognition Workshop*, February 1997.

J. Allen, M. Hunnicutt, and D. Klatt. *From Text to Speech: The MITalk System.* Cambridge University Press, Cambridge, 1987.

G. M. Ayers. Discourse functions of pitch range in spontaneous and read speech. In *Working Papers in Linguistics No. 44*, pages 1–49. Ohio State University, 1994.

L. R. Bahl, J. K. Baker, P. S. Cohen, F. Jelinek, B. L. Lewis, and R. L. Mercer. Recognition of a continuously read natural corpus. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-78)*, pages 422–424, Tulsa, 1978.

L. R. Bahl, S. Das, P. V. deSouza, M. Epstein, R. L. Mercer, B. Merialdo, D. Nahamoo, M. A. Picheny, and J. Powell. Automatic phonetic baseform determination. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-91)*, volume 1, pages 173–176, 1981.

A. Bell, D. Jurafsky, E. Fosler-Lussier, C. Girand, and D. Gildea. Forms of English function words— effects of disfluencies, turn position, age and sex, and predictability. In *Proc. of the XIVth International Congress of Phonetic Sciences (ICPhS)*, San Francisco, August 1999.

K. Beulen, S. Ortmanns, A. Eiden, S. Martin, L. Welling, J. Overmann, and H. Ney. Pronunciation modelling in the RWTH large vocabulary speech recognizer. In H. Strik, J. Kessens, and M. Wester, editors, *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 13–16, Kerkrade, Netherlands, April 1998.

S. Bird and T. Ellison. One level phonology: Autosegmental representations and rules as finite automata. *Computational Linguistics*, 20:55–90, 1994.

H. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Academic Publishers, 1993.

H. Bourlard, N. Morgan, C. Wooters, and S. Renals. CDNN: A context dependent neural network for continuous speech recognition. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-92)*, pages II.349–352, San Francisco, California, March 1992.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees.* Wadsworth, Belmont, 1984.

W. Buntine. Tree classification software. In *Technology 2002: The Third National Technology Transfer Conference and Exposition*, Baltimore, December 1992. IND Software available from NASA at `http://ic-www.arc.nasa.gov/ic/projects/bayes-group/ind/IND-program.html`.

J. Bybee. The phonology of the lexicon: evidence from lexical diffusion. In M. Barlow and S. Kemmer, editors, *Usage-based models of Language.* CSLI, Stanford, in press.

N. Campbell. Segmental elasticity and timing in japanese speech. In Y. Tohkura, E. Vatikiotis-Bateson, and Y. Sagisaka, editors, *Speech Perception, Production, and Linguistic Structure*, pages 403–418. Ohmsha/IOS Press, Tokyo, 1992.

W. N. Campbell. Syllable-level duration determination. In *1st European Conference on Speech Communication and Technology (Eurospeech-89)*, volume 2, pages 698–701, Paris, 1989.

C. Cardinal, R. Coulston, and C. Richey. People say the darnedest things. In *3rd Annual Mini-symposium of the UC Berkeley Phonology Lab and Society of Linguistics Undergraduates*, Berkeley, CA, April 1997.

R. Carlson and B. Granström. Modelling duration for different text materials. In *1st European Conference on Speech Communication and Technology (Eurospeech-89)*, volume 2, pages 328–331, Paris, 1989.

F. Chen. Identification of contextual factors for pronounciation networks. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-90)*, pages 753–756, 1990.

N. Chomsky and M. Halle. *The Sound Pattern of English.* Harper and Row, New York, 1968.

K. W. Church. *Phonological Parsing in Speech Recognition.* Kluwer Academic Publishers, Boston, 1987.

M. H. Cohen. *Phonological Structures for Speech Recognition.* Ph.D. thesis, University of California, Berkeley, 1989.

P. S. Cohen and R. L. Mercer. The phonological component of an automatic speech-recognition system. In R. Reddy, editor, *Speech Recognition.* Academic Press, 1975.

G. Cook, J. Christie, D. Ellis, E. Fosler-Lussier, Y. Gotoh, B. Kingsbury, N. Morgan, S. Renals, T. Robinson, and G. Williams. The SPRACH system for the transcription of broadcast news. In *DARPA Broadcast News Workshop*, Herndon, Virginia, February 1999.

G. Cook, D. Kershaw, J. Christie, and A. Robinson. Transcription of broadcast television and radio news: The 1996 ABBOT system. In *DARPA Speech Recognition Workshop*, Chantilly, Virginia, February 1997.

W. Cooper, C. Soares, A. Ham, and K. Damon. The influence of inter and intra–speaker tempo on fundamental frequency and palatalization. *J. Acoustical Society of America*, 73(5):1723–1730, 1983.

W. E. Cooper and M. Danly. Segmental and temproal aspects of utterance-final lengthening. *Phonetica*, 38:106–115, 1981.

N. Cremelie and J.-P. Martens. On the use of pronunciation rules for improved word recognition. In *4th European Conference on Speech Communication and Technology (Eurospeech-95)*, pages 1747–1750, Madrid, 1995.

N. Cremelie and J.-P. Martens. In search of pronunciation rules. In H. Strik, J. Kessens, and M. Wester, editors, *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 23–27, Kerkrade, Netherlands, April 1998.

T. Crystal and A. House. Segmental durations in connected–speech signals: Current results. *J. Acoustical Society of America*, 83(4):1533–1573, 1988.

N. Daly and V. Zue. Statistical and linguistic analyses of $f_0$ in read and spontaneous speech. In *Proceedings of the 2nd Int'l Conference on Spoken Language Processing (ICSLP-92)*, pages 763–766, Banff, 1992.

R. De Mori, C. Snow, and M. Galler. On the use of stochastic inference networks for representing multiple word pronunciations. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-95)*, pages 568–571, Detroit, Michigan, 1995.

Defense Advanced Research Projects Agency. *DARPA Broadcast News Workshop*, Herndon, Virginia, March 1999.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum-likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Ser. B.*, 39, 1977.

N. Deshmukh, J. Ngan, J. Hamaker, and J. Picone. An advanced system to generate pronunciations of proper nouns. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-97)*, Munich, Germany, 1997.

E. Eide. Automatic modeling of pronunciation variations. In *DARPA Broadcast News Workshop*, Herndon, Virginia, March 1999.

J. Eisner. Efficient generation in primitive optimality theory. In *Proceedings of the 35th Meeting of the Association for Computational Linguistics*, Madrid, 1997.

T. Ellison. Phonological derivation in optimality theory. In *COLING-94*, 1994.

G. Fant, A. Kruckenberg, and L. Nord. Prediction of syllable duration, speech rate, and tempo. In *Proceedings of the 2nd Int'l Conference on Spoken Language Processing (ICSLP-92)*, pages 667–670, Banff, 1992.

M. Finke. The JanusRTK Switchboard/CallHome system: Pronunciation modeling. In *Proceedings of the LVCSR Hub 5 Workshop*, 1996.

M. Finke, J. Fritsch, and D. Koll. Modeling and efficient decoding of large vocabulary conversational speech. In *Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation*, June 1999.

M. Finke and A. Waibel. Flexible transcription alignment. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 34–40, Santa Barabara, CA, December 1997a.

M. Finke and A. Waibel. Speaking mode dependent pronunciation modeling in large vocabulary conversational speech recognition. In *5th European Conference on Speech Communication and Technology (Eurospeech-97)*, 1997b.

J. Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (ROVER). In *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997.

W. Fisher. Factors affecting recognition error rate. In *DARPA Speech Recognition Workshop*, Chantilly, VA, February 1996a.

W. Fisher. *The tsylb2 Program: Algorithm Description.* NIST, 1996b. Part of the tsylb2-1.1 software package.

W. M. Fisher. A statistical text-to-phone function using ngrams and rules. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-99)*, Phoenix, March 1999.

E. Fosler. Automatic learning of a model for word pronunciations: Status report. In *Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation.* NIST, May 13–15 1997.

E. Fosler, M. Weintraub, S. Wegmann, Y.-H. Kao, S. Khudanpur, C. Galles, and M. Saraclar. Automatic learning of word pronunciation from data. In *Proceedings of the 4th Int'l Conference on Spoken Language Processing (ICSLP-96)*, October 1996.

E. Fosler-Lussier, S. Greenberg, and N. Morgan. Incorporating contextual phonetics into automatic speech recognition. In *Proc. of the XIVth International Congress of Phonetic Sciences (ICPhS)*, San Francisco, August 1999.

E. Fosler-Lussier and N. Morgan. Effects of speaking rate and word frequency on conversational pronunciations. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 35–40, Kerkrade, Netherlands, April 1998.

E. Fosler-Lussier and N. Morgan. Effects of speaking rate and word frequency on conversational pronunciations. *Speech Communication*, in press.

E. Fosler-Lussier and G. Williams. Not just what, but also when: Guided automatic pronunciation modeling for broadcast news. In *DARPA Broadcast News Workshop*, Herndon, Virginia, March 1999.

E. Fosler-Lussier. Multi-level decision trees for static and dynamic pronunciation models. In *6th European Conference on Speech Communication and Technology (Eurospeech-99)*, Budapest, Hungary, September 1999.

C. A. Fowler and J. Housum. Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language*, 26: 489–504, 1987.

H. Franco, M. Cohen, N. Morgan, D. Rumelhart, and V. Abrash. Hybrid neural network/hidden Markov model continuous-speech recognition. In *Proceedings of the 2nd Int'l Conference on Spoken Language Processing (ICSLP-92)*, pages 915–918, Banff, 1992.

J. Friedman. Computer exploration of fast-speech rules. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):100–103, February 1975.

T. Fukada, T. Yoshimura, and Y. Sagisaka. Automatic generation of multiple pronunciations based on neural networks. *Speech Communication*, 27:63–73, 1999.

W. Ganong. Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Performance and Perception*, 6:110–125, 1980.

J. Garofolo, L. Lamel, W. Fisher, J. Fiscus, D. Pallett, and N. Dahlgren. DARPA TIMIT acoustic-phonetic continuous speech corpus. Technical Report NISTIR 4930, National Institute of Standards and Technology, Gaithersburg, MD, February 1993. Speech Data published on CD-ROM: NIST Speech Disc 1-1.1, October 1990.

D. Gildea and D. Jurafsky. Learning bials and phonological-rule induction. *Computational Linguistics*, 22(4):497–530, December 1996.

I. J. Good. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40:16–264, 1953.

S. Greenberg. In praise of imprecision: Insights into spoken language gleaned from phonetic transcription of the Switchboard corpus. Invited presentation at the Workshop on Innovative Techniques in Large Vocabulary Speech Recognition, Johns Hopkins University, 14 August 1996.

S. Greenberg. On the origins of speech intelligibility in the real world. In *Proceedings of the ESCA Workshop on Robust Speech Recognition for Unknown Communication Channels*, April 1997a.

S. Greenberg. WS96 project report: The Switchboard transcription project. In F. Jelinek, editor, *1996 LVCSR Summer Research Workshop Technical Reports*, chapter 6. Center for Language and Speech Processing, Johns Hopkins University, April 25 1997b.

S. Greenberg. Speaking in shorthand — a syllable-centric perspective for understanding pronunciation variation. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 47–56, Kerkrade, Netherlands, April 1998.

M. Gregory, W. Raymond, A. Bell, E. Fosler-Lussier, and D. Jurafsky. The effects of collocational strength and contextual predictability in lexical production. In *Proceedings of the Chicago Linguistic Society*, 1999.

A. Henderson, F. Goldman-Eisler, and A. Skarbek. Sequential temporal patterns in spontaneous speech. *Language and Speech*, 9:207–216, 1966.

H. Hermansky. Perceptual linear predictive (PLP) analysis of speech. *J. Acoustical Society of America*, 87(4), April 1990.

J. L. Hieronymus, D. McKelvie, and F. R. McInnes. Use of acoustic sentence level and lexical stress in HSMM speech recognition. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-92)*, volume 1, pages 225–227, San Francisco, 1992.

T. Holter and T. Svendsen. Maximum likelihood modelling of pronunciation variation. In H. Strik, J. Kessens, and M. Wester, editors, *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 63–66, Kerkrade, Netherlands, April 1998.

J. Hooper. Word frequency in lexical diffusion and the source of morphophonological change. In W. Christie, editor, *Current Progress in Historical Linguistics*, pages 96–105. North Holland, Amsterdam, 1976.

J. J. Humphries. *Accent Modelling and Adaptation in Automatic Speech Recognition.* Ph.D. thesis, Trinity Hall, University of Cambridge, Camridge, England, October 1997.

T. Imai, A. Ando, and E. Miyasaka. A new method for automatic generation of speaker-dependent phonological rules. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-95)*, volume 6, pages 864–867, 1995.

F. Jelinek. *Statistical Methods for Speech Processing.* Language, Speech and Communcation Series. MIT Press, Cambridge, MA, 1997.

F. Jelinek and R. Mercer. Interpolated estimations of Markov source parameters from sparse data. In *Pattern Recognition in Practice*, pages 381–397, Amsterdam, May 1980.

J. Jiang, H.-W. Hon, and X. Huang. Improvements on a trainable letter-to-sound converter. In *5th European Conference on Speech Communication and Technology (Eurospeech-97)*, Rhodes, Greece, 1997.

C. Johnson. *Formal Aspects of Phonological Description.* Mouton, The Hague, 1972. Monographs on Linguistic Analysis No. 3.

M. Jordan and R. Jacobs. Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, 6:181–214, 1994.

D. Jurafsky, A. Bell, E. Fosler-Lussier, C. Girand, and W. Raymond. Reduction of English function words in Switchboard. In *Proceedings of the 5th Int'l Conference on Spoken Language Processing (ICSLP-98)*, Sydney, Australia, December 1998.

D. Jurafsky and J. Martin. *Speech and Language Processing.* Prentice-Hall, Englewood Cliffs, NJ, 2000. To appear.

D. Kahn. *Syllable-Based Generalizations in English Phonology.* Garland Publishing, New York and London, 1980.

E. Kaisse. *Connected Speech: the Interaction of Syntax and Phonology.* Academic Press, 1985.

R. Kaplan and M. Kay. Phonological rules and finite-state transducers. Paper presented at the annual meeting of the Linguistics Society of America, New York. See also [Kaplan and Kay, 1994]., 1981.

R. Kaplan and M. Kay. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378, 1994.

L. Karttunen. Finite state constraints. In J. Goldsmith, editor, *The Last Phonological Rule*, chapter 6, pages 173–194. University of Chicago Press, Chicago, 1993.

S. M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recogniser. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 35(3):400–401, 1987.

P. Keating. Word-level phonetic variation in large speech corpora. To appear in an issue of *ZAS Working Papers in Linguistics,* ed. Berndt Pompino-Marschal. Available as `http://www.humnet.ucla.edu/humnet/linguistics/people/keating/berlin1.pdf`., 1997.

M. Kenestowicz and C. Kisseberth. *Generative Phonology.* Harcourt Brace Jovanovich, Orlando, 1979.

D. Kershaw, M. Hochberg, and A. Robinson. Context dependent classes in a hybrid recurrent network-HMM speech recognition system. In *Neural Information and Processing Systems 8*, 1996.

B. E. D. Kingsbury. *Perceptually-inspired signal processing strategies for robust speech recognition in reverberant environments.* Ph.D. thesis, University of California, Berkeley, California, 1998.

K. Kirchhoff. Combining articulatory and acoustic information for speech recognition in noisy and reverberant environments. In *Proceedings of the 5th Int'l Conference on Spoken Language Processing (ICSLP-98)*, Sydney, December 1998.

S. Kitazawa, H. Ichikawa, S. Kobayashi, and Y. Nishinuma. Extraction and representation of rhythmic components of spontaneous speech. In *5th European Conference on Speech Communication and Technology (Eurospeech-97)*, pages 641–644, Rhodes, Greece, 1997.

D. H. Klatt. Synthesis by rule of segmental durations in English sentences. In B. Lindblom and S. Ohman, editors, *Frontiers of Speech Communication Research.* Academic Press, London, 1979.

Y. Konig, H. Bourlard, and N. Morgan. Remap: Recursive estimation and maximization of a posteriori probabilities - application to transition-based connectionist speech recognition. In *Advances in Neural Information Processing Systems (NIPS 8)*, 1995.

F. Koopmans-van Beinum. Spectro-temporal reduction and expansion in spontaneous speech and read text: the role of focus words. In *Proceedings of the 1st Int'l Conference on Spoken Language Processing (ICSLP-90)*, Kobe, Japan, 1990.

K. Koskenniemi. Two-level morphology: A general computational model of word-form recognition and production. Technical Report Publication No. 11, Department of General Linguistics, University of Helsinki, 1983.

R. Kuhn, J.-C. Junqua, and P. Martzen. Rescoring multiple pronunciations generated from spelled words. In *Proceedings of the 5th Int'l Conference on Spoken Language Processing (ICSLP-98)*, Sydney, Austrailia, 1998.

G. Laan. The contribution of intonation, segmental durations, and spectral features to the perception of a spontaneous and a read speaking style. *Speech Communication*, 22:43–65, 1997.

W. Labov. *Sociolinguistic Patterns.* U. of Pennsylvania Press, Philadelphia, Pennsylvania, 1972.

W. Labov. *Principles of linguistic change: internal factors.* Basil Blackwell, Oxford, 1994.

L. Lamel and G. Adda. On designing pronunciation lexicons for large vocabulary, continuous speech recognition. In *Proceedings of the 4th Int'l Conference on Spoken Language Processing (ICSLP-96)*, 1996.

K. F. Lee. *Automatic Speech Recognition: The Development of the SPHINX System.* Kluwer Academic Publishers, 1989.

W. J. M. Levelt. *Speaking: From Intention to Articulation.* ACL-MIT Press Series in Natural-Language Processing. MIT Press, Cambridge, Massachusetts, 1989.

H. Levin, C. Schaffer, and C. Snow. The prosodic and paralinguistic features of reading and telling stories. *Language and Speech*, 25(1):43–54, 1982.

P. Lieberman. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6(3):172–187, 1963.

M. Lincoln, S. Cox, and S. Ringland. A comparison of two unsupervised approaches to accent identification. In *Proceedings of the 5th Int'l Conference on Spoken Language Processing (ICSLP-98)*, Sydney, December 1998.

Linguistic Data Consortium (LDC). The PRONLEX pronunciation dictionary. Available from the LDC, ldc@unagi.cis.upenn.edu, 1996. Part of the COMLEX distribution.

B. Lowerre and R. Reddy. The HARPY speech recognition system. In W. A. Lea, editor, *Trends in Speech Recognition*, chapter 15, pages 340–360. Prentice Hall, 1980.

J. Lucassen and R. Mercer. An information theoretic approach to the automatic determination of phonemic baseforms. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-84)*, 1984.

K. Ma, G. Zavaliagkos, and R. Iyer. Pronunciaion modeling for large vocabulary conversational speech recognition. In *Proceedings of the 5th Int'l Conference on Spoken Language Processing (ICSLP-98)*, Sydney, Australia, December 1998.

D. McAllaster and L. Gillick. Fabricating conversational speech data with acoustic models: a program to examine model-data mismatch. In *Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation*, 1999.

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 the 5th Int'l Conference on Spoken Language Processing (ICSLP-98)*, pages 1847–1850, Sydney, Australia, December 1998.

M. McCandless and J. Glass. Empirical acquisition of word and phrase classes in the ATIS domain. In *3rd European Conference on Speech Communication and Technology (Eurospeech-93)*, volume 2, pages 981–984, Berlin, Germany, 1993.

C. A. Miller. *Pronunciation Modeling in Speech Synthesis.* Ph.D. thesis, University of Pennsylvania, Philadelphia, June 1998. Also available as Institute for Research in Cognitive Science Report 98–09.

J. Miller and T. Baer. Some effect of speaking rate on the production of /b/ and /w/. *J. Acoustical Society of America*, 73(5):1751–1755, 1983.

J. Miller and F. Grosjean. How the components of speaking rate influence perception of phonetic segments. *Journal of Experimental Psychology: Human Performance and Perception*, 7(1):208–215, 1981.

J. Miller and A. M. Liberman. Some effects of later occurring information on perception of stop consonant and semivowel. *Perceptial Pschophysics*, 25:457–465, 1979.

N. Mirghafori, E. Fosler, and N. Morgan. Fast speakers in large vocabulary continuous speech recognition: Analysis & antidotes. In *4th European Conference on Speech Communication and Technology (Eurospeech-95)*, 1995.

N. Mirghafori, E. Fosler, and N. Morgan. Towards robustness to fast speech in ASR. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-96)*, pages I335–338, Atlanta, Georgia, May 1996.

M. Mohri, M. Riley, D. Hindle, A. Ljolje, and F. Pereira. Full expansion of context-dependent networks in large vocabulary speech recognition. In *Proceedings of the 5th Int'l Conference on Spoken Language Processing (ICSLP-98)*, Sydney, Austrailia, December 1998.

N. Morgan and E. Fosler-Lussier. Combining multiple estimators of speaking rate. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-98)*, Seattle, WA, May 1998.

J. Ngan, A. Ganapathiraju, and J. Picone. Improved surname pronunciations using decision trees. In *Proceedings of the 5th Int'l Conference on Spoken Language Processing (ICSLP-98)*, Sydney, Austrailia, 1998.

NIST. Switchboard corpus: Recorded telephone conversations. National Institute of Standards and Technology Speech Disc 9-1 to 9-25, October 1992.

NIST. 1996 Broadcast News speech corpus. CSR-V, Hub 4, Produced by the Lingustic Data Consortium, Catalog No. LDC97S44., 1996.

NIST. 1997 Broadcast News speech corpus. CSR-V, Hub 4, Produced by the Lingustic Data Consortium, Catalog No. LDC98S71., 1997.

H. Nock and S. Young. Detecting and correcting poor pronunciations for multiword units. In H. Strik, J. Kessens, and M. Wester, editors, *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 85–90, Kerkrade, Netherlands, April 1998.

J. J. Odell. The use of decision trees with context sensitive phoneme modeling. Master's thesis, Cambridge University, Cambridge, England, 1992.

J. J. Ohala. Around flat. In V. A. Fromkin, editor, *Phonetic Linguistics: Essays in honor of Peter Ladefoged*, chapter 15, pages 223–241. Academic Press, Orlando, 1985.

J. J. Ohala. There is no interface between phonetics and phonology: A personal view. *Journal of Phonetics*, 18:153–171, 1990.

R. N. Ohde and D. J. Sharf. *Phonetic Analysis of Normal and Abnormal Speech.* MacMillan Publishing Company, New York, 1992.

J. Oncina, P. García, and E. Vidal. Learning subsequential transducers for pattern recognition tasks. *IEEE Transcations on Pattern Analysis and Machine Intelligence*, 15:448–458, May 1993.

B. Oshika, V. Zue, R. V. Weeks, H. Neu, and J. Aurbach. The role of phonological rules in speech understanding research. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP-23(1):104–112, February 1975.

M. Ostendorf, B. Byrne, M. Bacchiani, M. Finke, A. Gunawardana, K. Ross, S. Roweis, E. Shriberg, D. Talkin, A. Waibel, B. Wheatley, and T. Zeppenfeld. Modeling systematic variations in pronunciation via a language-dependent hidden speaking mode. In F. Jelinek, editor, *1996 LVCSR Summer Research Workshop Technical Reports*, chapter 4. Center for Language and Speech Processing, Johns Hopkins University, April 1997.

D. Pallett, J. Fiscus, G. Garofolo, A. Martin, and M. Przybocki. 1998 Broadcast News benchmark test results: English and non-English word error rate performance measures. In *DARPA Broadcast News Workshop*. Defense Advanced Research Projects Agency, 1999.

S. Parfitt and R. Sharman. A bi-directional model of English pronunciation. In *2nd European Conference on Speech Communication and Technology (Eurospeech-91)*, Genova, Italy, 1991.

D. Pisoni, T. Carrell, and S. Gano. Perception of the duration of rapid spectrum changes in speech and nonspeech signals. *Perception and Psychophysics*, 34(4):314–322, 1983.

P. Placeway, S. Chen, M. Eskenazi, U. Jain, V. Parikh, B. Raj, M. Ravishankar, R. Rosenfeld, K. Seymore, M. Siegler, R. Stern, and E. Thayer. The 1996 Hub-4 Sphinx-3 system. In *DARPA Speech Recognition Workshop*, Chantilly, VA, February 1997.

I. Pollack and J. M. Pickett. Intelligibility of excerpts from fluent speech: Auditory vs. structural context. *Journal of Verbal Learning and Verbal Behavior*, 3:79–84, 1964.

R. F. Port. The influence of tempo on stop closure donation as a cue for voicing and place. *Journal of Phonetics*, 7:45–56, 1979.

P. Price, W. Fisher, J. Bernstein, and D. Pallet. The DARPA 1000-word Resource Management database for continuous speech recognition. In *Proceedings IEEE Intl. Conf. on Acoustics, Speech, and Signal Processing*, volume 1, pages 651–654, New York, 1988. IEEE.

A. Prince and P. Smolensky. Optimality theory: constraint interaction in generative grammar. Unpublished ms., Technical Reports of the Rutgers University Center for Cognitive Science, 1993.

L. Rabiner. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), February 1989.

M. A. Randolph. A data-driven method for discovering and predicting allophonic variation. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-90)*, volume 2, pages 1177–1180, Albuquerque, New Mexico, 1990.

S. Renals and M. Hochberg. Decoder technology for connectionist large vocabulary speech recognition. Technical report, Sheffield University Department of Computer Science, September 1995a.

S. Renals and M. Hochberg. Efficient search using posterior phone probability estimators. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-95)*, pages 596–599, Detroit, MI, 1995b.

K. Ries, F. Buø, and Y.-Y. Wang. Towards better language modeling in spontaneous speech. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-95)*, Yokohama, Japan, 1995.

M. Riley. A statistical model for generating pronunciation networks. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-91)*, pages 737–740, 1991.

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. In *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 109–116, Kerkrade, Netherlands, April 1998.

M. Riley, F. Pereira, and M. Mohri. Transducer composition for context-dependent network expansion. In *5th European Conference on Speech Communication and Technology (Eurospeech-97)*, Rhodes, Greece, 1997.

M. Riley and A. Ljolje. Automatic generation of detailed pronunciation lexicons. In *Automatic Speech and Speaker Recognition: Advanced Topics*. Kluwer Academic Publishers, 1995.

T. Robinson and J. Christie. Time-first search for large vocabulary speech recognition. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-98)*, pages 829–832, Seattle, WA, May 1998.

S. Sagayama. Phoneme environment clustering for speech recognition. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-89)*, pages 397–400, 1989.

M. Saraclar. Automatic learning of a model for word pronunciations: Status report. In *Conversational Speech Recognition Workshop: DARPA Hub-5E Evaluation*, Baltimore, MD, May 1997.

M. Saraclar, H. Nock, and S. Khudanpur. Pronunciation modeling by sharing Gaussian densities across phonetic models. In *6th European Conference on Speech Communication and Technology (Eurospeech-99)*, Budapest, 1999.

F. Schiel. A new approach to speaker adaptation by modelling pronunciation in automatic speech recognition. *Speech Communication*, 13:281–286, 1993.

F. Schiel, A. Kipp, and H. G. Tillmann. Statistical modelling of pronunciation: it's not the model, it's the data. In H. Strik, J. Kessens, and M. Wester, editors, *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, pages 131–136, Kerkrade, Netherlands, April 1998.

P. Schmid, R. Cole, and M. Fanty. Automatically generated word pronunciations from phoneme classifier output. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-93)*, volume 2, pages 223–226, 1987.

T. Sejnowski and C. Rosenberg. Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168, 1987.

M. A. Siegler and R. M. Stern. On the effects of speech rate in large vocabulary speech recognition systems. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-95)*, 1995.

R. Silipo and S. Greenberg. Automatic transcription of prosodic stress for spontaneous English discourse. In *Proc. of the XIVth International Congress of Phonetic Sciences (ICPhS)*, San Francisco, August 1999.

T. Sloboda and A. Waibel. Dictionary learning for spontaneous speech recognition. In *Proceedings of the 4th Int'l Conference on Spoken Language Processing (ICSLP-96)*, 1996.

R. Sproat and M. Riley. Compilation of weighted finite-state transducers from decision trees. In *Proceedings of the 34th Meeting of the Association for Computational Linguistics*, pages 215–222, Santa Cruz, CA, 1996.

H. Strik, J. Kessens, and M. Wester, editors. *ESCA Tutorial and Research Workshop on Modeling Pronunciation Variation for Automatic Speech Recognition*, Kerkrade, Netherlands, April 1998.

Q. Summerfield. Articulatory rate and perceptual constancy in phonetic perception. *Journal of Experimental Psychology: Human Performance and Perception*, 7:1074–1095, 1981.

G. Tajchman. ICSI phonetic feature matrix. Personal Communication, 1994.

G. Tajchman, E. Fosler, and D. Jurafsky. Building multiple pronunciation models for novel words using exploratory computational phonology. In *4th European Conference on Speech Communication and Technology (Eurospeech-95)*, Madrid, Spain, September 1995a.

G. Tajchman, D. Jurafsky, and E. Fosler. Learning phonological rule probabilities from speech corpora with exploratory computational phonology. In *Proceedings of the 33rd Meeting of the Association for Computational Linguistics*, 1995b.

B. Tesar. *Computational Optimality Theory.* Ph.D. thesis, University of Colorado, Boulder, 1995.

J. P. Verhasselt and J.-P. Martens. A fast and reliable rate of speech detector. In *Proceedings of the 4th Int'l Conference on Spoken Language Processing (ICSLP-96)*, pages 2258–2261, Philadelphia, PA, October 1996.

M. Weintraub, E. Fosler, C. Galles, Y.-H. Kao, S. Khudanpur, M. Saraclar, and S. Wegmann. WS96 project report: Automatic learning of word pronunciation from data. In F. Jelinek, editor, *1996 LVCSR Summer Research Workshop Technical Reports*, chapter 3. Center for Language and Speech Processing, Johns Hopkins University, April 1997.

M. Weintraub, H. Murveit, M. Cohen, P. Price, J. Bernstein, G. Baldwin, and D. Bell. Linguistic constraints in hidden Markov model based speech recognition. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-89)*, pages 651–654, 1988.

C.-M. Westendorf and J. Jelitto. Learning pronunciation dictionary from speech data. In *Proceedings of the 4th Int'l Conference on Spoken Language Processing (ICSLP-96)*, 1996.

M. Wester, J. Kessens, and H. Strik. Modeling pronunciation variation for a Dutch CSR: Testing three methods. In *Proceedings of the 5th Int'l Conference on Spoken Language Processing (ICSLP-98)*, Sydney, Australia, December 1998.

D. A. G. Williams. *Knowing what you don't know: roles for confidence measures in automatic speech recognition.* Ph.D. thesis, University of Sheffield, Sheffield, England, 1999.

M. M. Withgott and F. R. Chen. *Computational Models of American Speech.* Center for the Study of Language and Information, Stanford, CA, 1993.

J. J. Wolf and W. A. Woods. The HWIM speech understanding system. In W. A. Lea, editor, *Trends in Speech Recognition*, chapter 14, pages 316–339. Prentice Hall, 1980.

C. Wooters and A. Stolcke. Multiple-pronunciation lexical modeling in a speaker independent speech understanding system. In *Proceedings of the 3rd Int'l Conference on Spoken Language Processing (ICSLP-94)*, 1994.

C. C. Wooters. *Lexical modeling in a speaker independent speech understanding system.* Ph.D. thesis, University of California, Berkeley, 1993. International Computer Science Institute Technical Report TR-93-068.

S. Young, J. Jansen, J. Odell, D. Ollasen, and P. Woodland. *The HTK Book (Version 2.0).* Entropic Cambridge Research Laboratory, 1995.

S. J. Young, J. J. Odell, and P. C. Woodland. Tree-based state tying for high accuracy acoustic modelling. In *Proceedings IEEE Int'l Conference on Acoustics, Speech, & Signal Processing (ICASSP-94)*, pages 307–312, 1994.

A. Zwicky. Auxiliary Reduction in English. *Linguistic Inquiry*, 1(3):323–336, July 1970.

A. Zwicky. Note on a phonological hierarchy in English. In R. Stockwell and R. Macaulay, editors, *Linguistic Change and Generative Theory.* Indiana University Press, 1972a.

A. Zwicky. On Casual Speech. In *Eighth Regional Meeting of the Chicago Linguistic Society*, pages 607–615, April 1972b.

# Citation Index

# Index