PRONUNCIATION MODELING IN SPEECH SYNTHESIS

Corey Andrew Miller

A DISSERTATION

in

Linguistics

Presented to the Faculties of the University of Pennsylvania in Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy

1998

_____
Mark Liberman
Supervisor of Dissertation

_____
George Cardona
Graduate Group Chairperson

DEDICATION

To Jonathan Connett.

ACKNOWLEDGMENTS

I am very pleased to have had the encouragement and support of a committee of three linguists for whom I have the greatest respect and admiration: Mark Liberman, William Labov and Eugene Buckley. Each of them made my transition back to Penn pleasant after what seemed like a long absence. It was a great pleasure to have Mark Randolph both as an external reader and as a colleague at Motorola. Mark's work at MIT a decade ago has served as an inspiration to me. Orhan Karaali made this dissertation possible in this millennium. As my manager for over two years at Motorola, Orhan insisted on making my dissertation a priority at work. Harry Bliss provided his voice to this project and our whole group is very grateful for his patience and cooperation. My colleagues at Motorola listened to my ideas and provided technical and theoretical assistance at every turn: Noel Massey, Jerry Corrigan, William Thompson, Andrew Mackie, Erica Zeinfeld, Otto Schnurr, Lea Adams, Michael Murdock, Joseph Goldberg and Lynette Melnar. My entire management was supportive in this effort: Ira Gerson, Kevin Kloker and James Mikulski. D. J. Stockley made the pursuit of intellectual property a positive learning experience. My colleagues at my former workplace, Franklin Electronic Publishers, provided a great environment as I wet my feet in the industrial world: Will Dowling, David Justice and Mike Wolff. I thank Matt Lennig and Kristin Precoda for introducing me to speech technology and software development at BNR. My linguist friends and colleagues at Penn and others schools provided support and encouragement while I lived in Philadelphia and afterwards: Christine Zeller, Tom Veatch, Peter Slomanson, Fabiola Varela-García, Mary O'Malley, Bill Reynolds, Hadass Sheffer, Stephanie Strassel, Nadia

Biassou, Lisa Lavoie, Hikyoung Lee, Alex Dimitriadis, and Jason Eisner. I thank Brian Sietsema, Pat Keating, Sean Fulop and Howard Nusbaum for providing valuable advice and information. My mother always believed that this dissertation would become a reality, and to her I owe inexpressible gratitude for providing encouragement at the more difficult moments of my graduate career. The thought of the rest of my family including my father, Stephanie, Ken, Jordan, Avery, Lee, Elihu, and my late grandparents has always kept me going. Thank you Jonathan for paying the bills, arranging exotic vacations and, most of all, hanging in there. Finally, I would like to thank the Summer Institute of Linguistics for their excellent phonetic fonts which are freely distributed on the World Wide Web.

ABSTRACT

PRONUNCIATION MODELING IN SPEECH SYNTHESIS

Corey Andrew Miller

Mark Liberman

This dissertation investigates the area of pronunciation modeling in speech synthesis. By pronunciation modeling, we mean architectures and principles for generating high-quality human-like pronunciations. The term pronunciation modeling has previously been applied in the context of speech recognition (e.g. Byrne et al. 1997). In that context, it describes theories and procedures for handling the pronunciation variation that naturally occurs across speakers. In contrast, our work is in the domain of text-to-speech synthesis, which, as we will show, requires modeling the pronunciation variation of an individual whose speech the synthesizer is attempting to model. We will explain our methodology for learning and reproducing pronunciation variation on an individual basis, and show how most crucial features of such variation can be easily generated using the architecture we describe. Throughout the course of this exposition, we highlight contributions to linguistic theory that such a thorough analysis of individual variation provides. We describe the postlexical module of an English text-to-speech synthesizer. This module is responsible for transforming underlying lexical pronunciations from a lexical database into contextually appropriate surface postlexical pronunciations. This transformation is achieved by machine learning of a corpus of hand-labeled postlexical pronunciations that have been aligned with lexical pronunciations. The machine learning is conducted by a

neural network, whose architecture and data encoding we describe. A thorough analysis of the performance of the postlexical module is offered, with attention to the relative success of the neural network at learning a wide range of postlexical phenomena. We examine the extent to which a symbolic approach to allophony is warranted, and provide an acoustic analysis that attempts to provide an answer to this question. Assessments of the success of currently existing theories of phonetics, phonology and their interface are offered, based on the experience of generating a complete postlexical phonology of English for use in synthetic speech.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Chapter 1. Introduction

1.1. What is speech synthesis pronunciation modeling?

This dissertation investigates the area of pronunciation modeling in speech synthesis. By pronunciation modeling, we mean architectures and principles for generating high-quality human-like pronunciations. Divay and Vitale (1997) have made the following prognostication for speech synthesis, "in the future, in text-to-speech systems, some segments and even syllables will disappear entirely and certain functors will be greatly attenuated." We will describe methods for achieving this kind of speech synthesis.

The term pronunciation modeling has previously been applied in the context of speech recognition (e.g. Byrne et al. 1997). In that context, it describes theories and procedures for handling the pronunciation variation that naturally occurs across speakers. In contrast, our work is in the domain of text-to-speech synthesis, which, as we will show, requires modeling the pronunciation variation of an individual whose speech the synthesizer is attempting to model. This dissertation will show how properly modeling individual pronunciation variation is crucial to the production of synthetic speech that is comprehensible, natural and acceptable to listeners. We will explain our methodology for learning and reproducing pronunciation variation on an individual basis, and show how most crucial features of such variation can be easily generated using the architecture we

describe. Throughout the course of this exposition, we highlight contributions to linguistic theory that such a thorough analysis of individual variation provides.

In this chapter, we first review the interactions of computational phonology and speech technology. We then explore other traditional phonological issues that this dissertation addresses. Subsequently, we discuss the inspiration that this work has drawn from sociolinguistic studies of variation in the phonological domain. We conclude the chapter with an overview of the rest of the dissertation.

1.2. Computational phonology and speech technology

This work may be situated in the developing field of computational phonology (see Bird 1995). One aspect of this field is the use of computers to test phonological theories on large datasets. Due to this orientation, computational phonology places an emphasis on the implementability of theoretical frameworks. This characteristic of computational phonology lends itself to employment in speech technology applications, where speed, clarity and descriptive adequacy are prized. Of course, computational phonology is not simply an application of phonology to real world speech technology problems. It is also engaged in the analysis of current phonological theory with respect to power, tractability and plausibility for human behavior (e.g. Kaplan and Kay 1994, Coleman 1995b).

The computational phonological techniques we will examine fall within the field of machine learning. These techniques can extract patterns from data that can then be

applied to unseen data for application of the original processes. The machine learning technique that we will investigate most thoroughly is the artificial neural network. Neural networks are composed of multiple low-power processing elements that simulate, to some extent, the makeup of the human brain. Neural networks have been shown to provide insight into linguistic problems since the pioneering work of Rumelhart and McClelland (1986) on learning the past tense of English verbs and Sejnowski and Rosenberg (1987) on learning orthography-phonetics conversion.

An important aspect of many computational phonological techniques is their reversibility. For example, in both speech technology applications and models of human linguistic behavior, progressing in either direction between orthography and lexical pronunciation or lexical pronunciation and postlexical pronunciation is desired. If a computational technique has been worked out in one direction, the ease with which it can be applied to the reverse problem, such as between speech synthesis and speech recognition, is an indication of its usefulness and potentially, its explanatory value (e.g. Meng et al. 1996). The reversibility of phonological processes is both a useful goal for engineering as well as a critical feature of human language processing. It would be surprising if our abilities to understand and produce speech used unrelated competencies and structures. The failure of theories of phonology to function reliably in a reversible manner has often been cited as a major criticism of those theories (Bird 1995, Coleman 1995a, Kaye 1989).

In this subsection, we will first explore the linguistic aspects of speech synthesis. We will then review attempts at postlexical modeling in both speech recognition and speech

synthesis. Finally, we will examine the uses that have been made of neural networks in phonological investigation.

1.2.1. Linguistic aspects of speech synthesis

Speech synthesis refers to the creation of human-sounding speech by computer. Liberman (1994b) presents a taxonomy of the various forms of input to this process, such as telephone numbers, concepts or arbitrary text. We will be considering the kind of speech synthesis that takes computer-readable text and transforms it into a speech waveform. In order to perform this task accurately, the synthesis application needs to transform text into a linguistic representation that can then be converted into the acoustic domain. For example, if we want the synthesizer to pronounce the English word *thought*, we convert it into a phonological representation, such as /θɔt/, and then the synthesizer produces a waveform that (hopefully) sounds like /θɔt/.[1]

There are a number of ways in which a synthesizer could be designed to produce an appropriate waveform for /θɔt/ and the other words that it is expected to pronounce. One common approach is concatenative synthesis, which takes acoustic subword units (e.g. phones, diphones, demidiphones, etc.) from a database of recorded speech and strings them together while smoothing the transitions between them.

[1] We will be enclosing phonemic representations in slashes //, and phonetic (allophonic) transcriptions in brackets []. We will sometimes use slashes, as here, when the distinction is not important.

It has been found much more reasonable to store speech for units approximating the phoneme, rather than for words or letters. Storing examples of whole words would be prohibitive in terms of computer storage,[2] given the amount of words a general-purpose application is likely to encounter. In addition, such a system would be at a loss when confronted with novel words, such as names or neologisms. Storing speech examples based on letters would lead to inaccuracy, due to the irregularities of English spelling (consider *thr**ough***, *b**ough***, *c**ough***, *en**ough***).

Consequently, units such as phonemes (or theory-neutral *phones*) are a common intermediary between text and speech. While the conception of text or words as strings of phonemes bundled with other information (such as stress and syllabification) is a common one that has proved useful to both linguistics and speech applications, numerous alternatives have been proposed. From the perspective of Optimality Theory, Golston (1996) proposes that morphemes be represented as constraint violations. Cohen (1995) introduces a non-symbolic phonetic notation derived from a self-organizing map (Kohonen 1988). Shillcock et al. (1993) employ a phonological system based on *elements*, using the apparatus of Government Phonology (Harris and Lindsey 1995).

In contrast to those approaches which attempt to model something smaller than the phoneme (in the sense that a molecule is smaller than a cell), Randolph (1989) advocates use of the syllable as the primary building block in a phonological system for speech recognition. While we will investigate the feasibility of some of these approaches to

[2] However, AcuVoice (www.acuvoice.com/faq.html) claims to have a 152 megabyte "sound bank".

phonological representation, we will maintain the notion of the phoneme as a useful starting point.

All speech synthesizers build upon knowledge of human speech. However, synthesis methods can be situated along a continuum of knowledge-based to data-based systems. For example, Allen et al. (1987) describe synthesis by rule, where analyses of multiple speakers over many years gave the designers the ability to generate acoustic parameters by rule for various speech sounds. The resulting synthesis may resemble the speech of an individual whose voice was studied to learn appropriate acoustic parameters for various sounds, however this is not necessarily the goal of such systems.

The comparison between knowledge- and data-oriented approaches is a recurrent theme in computational phonology and speech technology. For example, Daelemans, Gillis and Durieux (1994) contrast their own data-oriented or *empiricist* approach to the acquisition of stress, with various Principles and Parameters (Chomsky 1981) or *nativist* approaches, such as Dresher and Kaye (1990).

The advantage of data-oriented systems is that they can be applied to new languages and data without extensive revisions. Price (1996) provides an interesting commentary on the differing viewpoints of linguists and engineers who find themselves working together on natural language applications: engineers tend to look at symbolic systems, such as those employed by linguists, with some suspicion, while some linguists might at first be uncomfortable with statistical, data-driven approaches.

This dissertation will use an experimental speech synthesizer being developed at Motorola (Karaali et al. 1996, Gerson et al. 1996, Corrigan et al. 1997, Karaali et al. 1997, Karaali et al. forthcoming) as a test-bed for the ideas and experiments to be described below. The Motorola synthesizer (also known as MotorMouth) distinguishes itself from many other synthesizers through its extensive use of neural networks throughout the processing of text to speech (Karaali et al. 1998). Neural networks are used to perform machine learning of various aspects of the linguistic behavior of an individual whose voice is being modeled. The neural networks are trained on labeled samples of an individual's speech. When a new synthetic voice is desired, the neural networks are retrained on samples of speech from the new voice. We will describe the various neural networks and the training procedure below.

Figure 1-1 is a graphic depiction of the synthesizer, indicating the different modules of which it is composed. In general, the linguistic module of the synthesizer is responsible for transforming orthographic text into a symbolic linguistic representation. The linguistic representation is then passed to an acoustic module that transforms it into acoustic parameters which can be synthesized into audible speech.

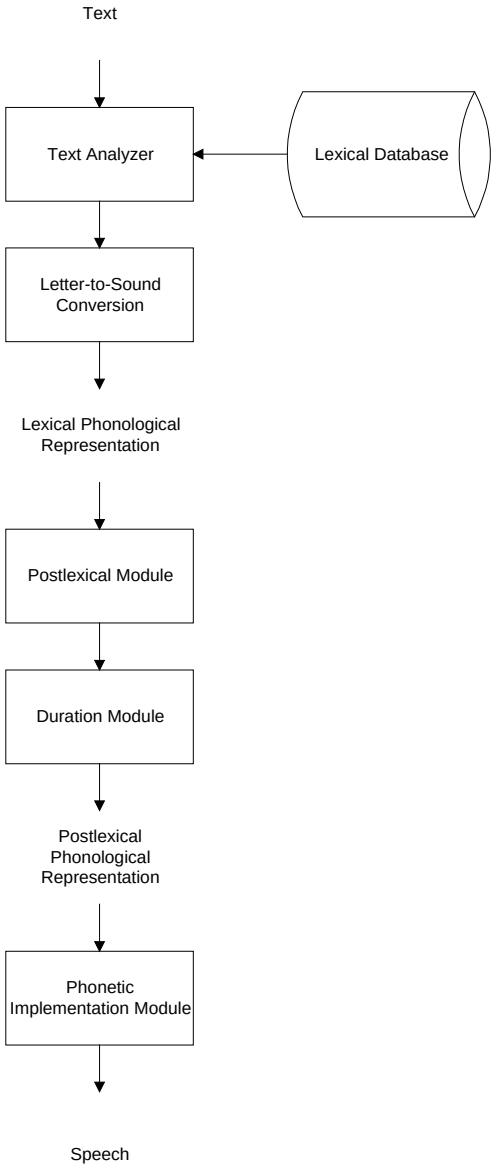

Figure 1-1: Motorola speech synthesizer

Sproat et al. (1998) use the term *text analysis* for many of the processes that take place inside the linguistic module. We divide the linguistic module into a text analyzer, letter-to-sound converter, postlexical module, and a duration module. The text analyzer first performs *text preprocessing*, which involves such operations as translating numbers and symbols into orthographic words. Text preprocessing might also involve the processing of e-mail headers, HTML codes and the ignoring of graphical objects. Preprocessed text is then tokenized into word-sized units that are looked up in the lexical database. A disambiguation module (Karaali et al., forthcoming) assigns words parts of speech (e.g. Church 1988), and in some cases, semantic information (e.g. Yarowsky 1997), to help in selecting the proper pronunciation for homographs.

If a given word's pronunciation is not present in the dictionary at all, a pronunciation is generated for that word by the letter-to-sound conversion module. Once lexical pronunciations have been determined for all the words of a given sentence, a lexical phonological representation is constructed. The lexical phonological representation is organized as a hierarchical prosodic phonological structure, including phones, syllables and phonological words, as well as higher-order constituents.

The lexical phonological representation is transformed and augmented to produce a postlexical phonological representation. In general, the postlexical module is responsible for transforming more or less phonemic, *lexical* pronunciations into *postlexical* pronunciations typical of the connected speech of the individual whose voice is being modeled. The modifications that take place in the postlexical module at present involve

only segmental insertions, deletions and substitutions, such as flapping and *t,d* deletion. These transformations are achieved by means of a neural network that has been trained to learn the correspondences between the pronunciations used in the speech database, to be described below, and the pronunciations used in the lexical database.

In the future, the postlexical module will also be responsible for modeling the prosody of the sentence and larger units of discourse in symbolic terms, such as ToBI (Tones and Break Indices) labels (Beckman and Elam 1997). Ladd (1996, 5) makes explicit the notion of handling intonation in the postlexical phonology. We will not be describing the prosody generation aspects of the postlexical module in this dissertation; however, we will be discussing the use of prosodic information in determining segmental postlexical phenomena.

The phonological representation, having been modified to reflect postlexical phonology, is then submitted to a Duration Module (Corrigan et al. 1997) which assigns durations to each phone, according to the speech of individual being modeled. The duration module uses a neural network that has been trained on postlexical phonological representations to determine appropriate contextual durations.

Finally, the postlexical phonological representation including constituent durations is submitted to a Phonetic Implementation (also known as Acoustic) Module (Karaali et al. 1997), which transforms the phonological representation into spectral parameters that are synthesized into a speech waveform. The acoustic module consists of a neural network

that has been trained to learn the correspondences between postlexical phonological representations and spectral parameters in a hand-labeled[3] speech database of a particular individual.

Essentially, the linguistic module's responsibility in the runtime operation of the synthesizer is to recreate the kinds of linguistic representations that were applied to the speech that the acoustic neural network was trained on. This will assure that the synthesized speech most closely matches the speech of the individual who is recorded in the database, which is the ultimate goal of this kind of speech synthesis.[4] Figure 1-2 summarizes the neural network training that is involved for each voice. Each of the neural networks trains on information acquired from a labeled speech database of the voice being modeled. In fact, the label files are analyzed to form a postlexical phonological representation analogous to the one created from text input described above. A figure of merit, besides ultimate voice quality, that can be used to evaluate development progress on the synthesizer, is the fidelity with which postlexical phonological representations derived from text match those derived from speech database label files.

---

[3] Wightman and Talkin (1997), Ljolje et al. (1997) and Vorstermans and Martens (1994) discuss various *automatic* labeling schemes.

[4] The concept of *voice fonts* that preserve individual speech characteristics has recently caught the public imagination: Andrew Pollack, "Sound Bites and Then Some: Computers May Soon Use Your Voice to Say Anything", *New York Times*, 21 April 1997, Business Day, p. 1.

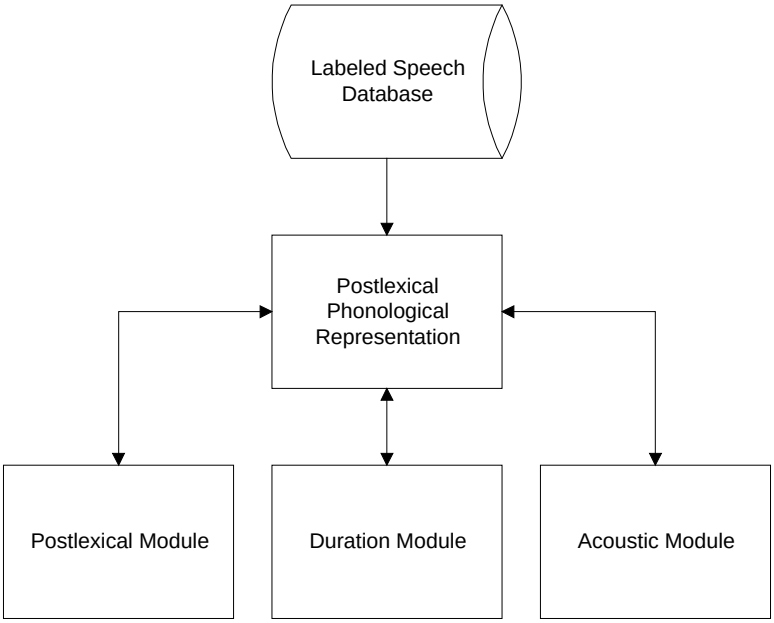

Figure 1-2: Motorola synthesizer training scheme

If a new voice is desired, there are two requirements for customizing the acoustic module. First, a sufficiently sized speech database of the new voice must be labeled in the manner required by the synthesizer. Second, a neural network must be trained to learn the correspondence between the linguistic labels and the spectral parameters of the particular database.

If the language spoken by a new voice has already been handled in the past, the modules that create the lexical phonological representation from text may not require any changes;

otherwise, a new dictionary and preprocessing module will be required.[5] In the case of each new voice, the postlexical module's neural network is retrained to learn the correspondences between the pronunciations used in the current voice's speech database and the pronunciation dictionary. In this way, the postlexical module can account for three sources of potentially idiosyncratic behavior: that found in dictionary transcriptions, that found in the speech database labeling, and that found in the habits of the particular speaker.

Finally, the duration module is retrained so that durations characteristic of the new voice are learned. It is believed that in not only recreating the linguistic habits of the speaker, but also the indexical properties of his or her voice, the synthesizer's intelligibility will be enhanced. Pisoni (1993) reports on the positive contribution of such indexical voice properties to intelligibility.

The integration of neural networks all the way from text to speech is a hallmark of the system (Karaali et al. 1998), and the success that it has may be attributable to the faithfulness with which it attempts to model human behavior. In addition, due to the potential reversibility of the networks of which it is composed, the system has promise for adaptation to some of the tasks of speech recognition.

[5] Of course, certain dialect differences, such as between British and American might well necessitate a new dictionary, or even new sentence tokenization routines: consider British "Mr Jones" vs. American "Mr. Jones".

1.2.2. Review of prior work in postlexical modeling

There are at least three dimensions along which research in postlexical modeling can be described: direction, technology and perspective. Direction involves whether lexical forms are being transformed into postlexical forms, as required in speech synthesis, or whether lexical forms are being determined on the basis of postlexical forms, as in speech recognition. In other words, the problem for speech synthesis is to determine, based on a lexical form, the most appropriate postlexical form, whereas in speech recognition, the most likely lexical form must be selected for a captured postlexical form.

Technology involves the mechanisms employed by the researchers, such as neural networks, finite state machines or decision trees. Perspective involves whether the enterprise has been undertaken by linguists or psycholinguists interested in human language capabilities or speech technologists working in automatic speech recognition or speech synthesis. We will arrange our review of postlexical modeling by discussing synthesis (lexical-postlexical) and then recognition (postlexical-lexical) approaches. Within each approach, we will discuss the technologies and perspectives employed. Finally, we present a section on phonetic studies that have analyzed the postlexical variation in corpora without a particular speech technology application in mind.

Williams (1994) developed a computational implementation of lexical phonology and tested it by trying to generate output forms along a stylistic continuum in both English and Spanish. In the case of English, her system is based on the description of lexical

phonology provided by Halle and Mohanan (1985), and Booij and Rubach (1987). Williams achieved what she described as more casual speech by progressively moving rules to later strata. She acknowledges that some of the forms generated are not actually attested in reports on casual speech.

Gildea and Jurafsky (1996) describe methods for automatically inducing finite state transducers to perform lexical-postlexical conversion from data. They artificially introduced postlexical phenomena, such as flapping, into a lexical database, and then attempted to induce transducers from the converted data. A first attempt failed because important linguistic information was not provided to the algorithm. A subsequent attempt is shown to succeed after the introduction of linguistic knowledge. Gildea and Jurafsky (1996) claim that their model represents an alternative between purely nativist and purely empiricist approaches, in that it attempts to represent prior knowledge as a set of learning biases that guide an empirical learning algorithm. They note that the work of Riley (1991) and Withgott and Chen (1993) is similar in this regard (see below).

Despite the introduction of linguistic knowledge, Gildea and Jurafsky claim that their approach is not necessarily nativist; rather, they suggest that "assuming in the system some very general and fundamental properties of phonological knowledge (whether innate or previously learned) and learning others empirically may provide a basis for future learning models" (1996, 499). This characterizes the postlexical learning methods discussed in this dissertation as well: we provide the postlexical neural network with

various phonological and prosodic information because we suspect that the network will find it useful, but we do not tell the network *how* to make use of it.

Gildea and Jurafsky group the learning biases they introduce as follows: faithfulness, community and context. Faithfulness refers to input-output correspondence, in this case between underlying and surface forms. They achieve faithfulness by aligning input and output phones using dynamic programming. Community refers to the similarities between segments that can be expressed by distinctive features. Context refers to a knowledge of preceding and following phonological environments.

Gildea and Jurafsky aver that the phonological example they elaborate is meant to suggest the kinds of biases that may be added to empiricist induction models, rather than to serve as a practical phonological learning device. They suggest that stochastic algorithms for language induction are much more likely to be a fruitful research direction due to the noise and nondeterminism inherent to linguistic data.

Albano and Moreira (1996) discuss some of the issues involved in selecting appropriate lexical and postlexical pronunciations for a Portuguese speech synthesizer. Since Portuguese phonemics is straightforwardly derivable from the language's orthography, an exception dictionary played a fairly small role in letter-to-sound conversion. Albano and Moreira chose an underspecified phonemic representation as the first stage of output of their letter-to-sound procedure, rather than a postlexical representation more

characteristic of surface phonetic phenomena. One of the reasons they cite for this is the difficulty of referring to segments on the phonetic level.

Albano and Moreira adopted a two-step letter-to-phone conversion, with the first operating on words and the second operating on sentences, producing a postlexical representation. They distinguish between postlexical and phonetic processes, assuming that the former do not introduce any segmental symbols beyond those already in the lexicon, i.e. they are structure preserving. Continuous, or gradient, processes, such as coarticulation phenomena are taken to be phonetic and are not represented in symbolic terms.

Fitt (1997) discusses a method of generating regional accents for an English speech synthesizer. She divides phonological processes into three categories possessing both obligatory and optional components: pre-lexicon transformations, post-lexicon transformations and connected speech rules. Pre-lexicon transformations assist in the generation of a basic pronunciation lexicon for each accent, for example, non-prevocalic /r/ can be removed from the lexicons of non-rhotic accents such as British RP (Received Pronunciation). Post-lexicon transformations introduce allophones, such as flapping in American English. Fitt (1997) offers some examples of connected speech rules, but it is not clear how these are different from her post-lexicon transformations. It is interesting to note that the postlexical neural network that we will be describing here covers some of the functionality of each of Fitt's sets of transformations.

Church (1983) argues that postlexical variation can actually improve recognition performance rather than derailing it. Speech recognizers that are equipped with knowledge about postlexical variation will be more robust for the variety of styles, tempos and dialects (in the speaker-independent case) with which they are confronted. Lippmann (1996) reports results on the relative performance of speech recognizers on various speech recognition tasks, shown below in Table 1-1. These results indicate that the most improvement is required on spontaneous speech recognition tasks, such as the Switchboard corpus (Godfrey et al. 1992). Spontaneous speech is marked by precisely the kinds of postlexical variation that we intend to explore and for whose production we hope to provide accurate and efficient computational techniques.

Table 1-1: Performance on two speech recognition tasks

Task	Human word error rate	Machine word error rate
Wall Street Journal	1%	12%
Switchboard	4%	66%

In the context of speech recognition, Riley and Ljolje (1996), Withgott and Chen (1993), and Cohen (1989) all describe methods for developing *pronunciation networks* based on postlexical pronunciations. This approach encodes much possible variation (perhaps too much) in each lexical entry. The first two methods differ from ours in their use of decision trees rather than neural networks. Gildea and Jurafsky (1996, 527) criticize decision tree approaches:

> One problem with these particular approaches is that since the decision tree for each segment is learned separately, the technique has difficulty forming generalizations about the behavior of similar segments. In addition, no generalizations are made about segments in similar contexts, or about long-distance dependencies.

We hope to show how our neural network approach to postlexical modeling addresses these criticisms.

Riley and Ljolje (1996) observed an improvement in word recognition accuracy of almost 3% when their system's dictionary included multiple postlexical variants as opposed to one phonemic representation per word. Cremelie and Martens (1996) and Fosler et al. (1996) also report improvements when baseform lexica are enhanced with postlexical variation.

Several investigators have examined the mapping between lexical and postlexical forms using the TIMIT[6] database (Withgott and Chen 1993, Randolph 1990, Riley 1989, 1991). The goal in these cases was to enhance recognition dictionaries with postlexical variation. All of these efforts used species of decision trees, such as classification and regression trees (CART, Breiman et al. 1993), rather than the neural network approach that we employ. Sproat and Riley (1996) describe a method for generating weighted finite state transducers for lexical-postlexical conversion (among other things) from decision trees.

In contrast to the above mentioned methods which rely on hand-labeled corpora, such as TIMIT, Tajchman et al. (1995) describe a method for assigning probabilities to optional

phonological rules across speakers that uses automatic speech recognition. Tajchman et al. benchmarked the probabilities generated by their system again hand-transcribed data, and showed a relatively good fit. They note that the optionality of postlexical rules makes the induction problem non-deterministic, whether the rules are considered phonological or phonetic.

Tajchman et al. took a lexicon of underlying forms (which resembles our lexical database in form and content, see section 3.1), and applied various phonological rules to it to produce a new lexicon of (in some cases, multiple) surface forms. Then a speech recognition system was used on a corpus of recorded read speech from multiple speakers to check how many times each surface form occurred in the corpus. Since each surface form was keyed to the rules that had produced it, it was possible to obtain a count for the application of each rule. This count, combined with a count of the times a rule did not apply, yielded a probability for each rule.

Tajchman et al. found that men were more likely than women to employ many of the postlexical rules they examined. Since they also found that male speech was faster than female speech in their corpus, they presumed a relationship between these two facts. Tajchman et al. note that the decision tree work of Withgott and Chen (1993) and Riley (1991) allows a more fine-grained analysis than their rule-based algorithm. They note, however, that a liability of the decision tree method is that it is more difficult to extract generalizations across classes of phonemes to which rules can apply.

---

[6] TIMIT is a blend of TI (Texas Instruments) and MIT (Massachusetts Institute of Technology), two of the

Two studies on variation in the TIMIT database, Keating et al. (1994) and Byrd (1994) bear important resemblances to our own study. TIMIT is a corpus of read speech from 630 speakers of American English, varying along geographical, age and gender lines, which has been transcribed by hand according to the principles described in Seneff and Zue (1988). Each speaker uttered 10 sentences, 2 of which were recorded from all speakers. Because it is a multispeaker database, the conclusions about variation will undoubtedly be different from the ones we draw from our single speaker database. Nevertheless, the fact that both databases are labeled in a similar fashion (see section 3.2), allows us to draw on what Keating et al. and Byrd have learned.

Keating et al.'s study is divided into a transcription study and an acoustic study. The transcription study examines variation in the pronunciation of the word *the*, with particular attention to the quality of the vowel in relation to the following environment. We will compare their results to our own in section 6.5. The acoustic study examined the effect of vowel context on the acoustics of velar stop consonants. It was found that the frequency of the burst peak was higher when the F2 of the following vowel was higher; indeed, following vowels were shown to have a much stronger effect on stop release bursts than those preceding.

Byrd (1994) examined relations of sex and dialect to reduction in the TIMIT database. She found that men had an average rate of 4.69 syllables per second, while women had an average rate of 4.42 syllables per second; a significant difference ($p = .0001$). Dialect

---

institutions that collaborated to create this database (Fisher et al. 1987).

was also found to be significant (p = .0001) for rate differences, ranging from slowest to fastest in the following order: South, South Midland, New York City, North, West, North Midland, North East and "Army Brat" (people who had moved around).

In addition to speaking more slowly, Byrd found that women exhibited more conservative behavior on some postlexical processes. For example, she notes that women released sentence final stops more often than men. She also found that women produced significantly fewer flaps than the men. Byrd also examined the central vowels in TIMIT and found the following distribution from a total of 17,858 central vowels: 55% [ɨ] (ix) , 18% [ʌ] (ah), 27% [ə] (ax). Speakers in New York City and the West appear to have a greater preference for [ɨ] than [ə], while the North Midland behaves in the reverse manner. We will contrast Byrd's results on reduced vowels with our own in section 6.4.

Byrd notes that since in TIMIT, "no phonemic transcription is provided, automating an investigation into many phonological processes is difficult or impossible" (1994, 43). We hope to show in this dissertation that a phonemic transcription *can* be aligned with the phonetic transcriptions in such a database, resulting in much interesting information about such phonological processes.

Strassel (1997) and Fulop and Keating (1996) make use of the Switchboard corpus to examine pronunciation variation. Switchboard is a recorded multispeaker database of spontaneous speech between strangers available from the Linguistic Data Consortium. It has been orthographically transcribed, although some sections have been phonetically

transcribed (e.g. Fulop and Keating 1996, Greenberg et al. 1996). The principle advantage for examining phonological processes in Switchboard over TIMIT is its use of spontaneous speech. Strassel examined flapping in Switchboard and HUB-4, a corpus of broadcast news recordings, and we will discuss her results in comparison to our own with respect to this variable in section 6.4.

Fulop and Keating examined a subset of Switchboard, consisting of up to 40 tokens each of 48 lexical items. The lexical items were approximately evenly distributed between function and content words. The database subset was transcribed independently by two graduate student transcribers. Fulop and Keating examined transcriber reliability for each lexical item. They differentiated between the high reliability words (≥ 90%) in (1) from low reliability words (≥ 70%) in (2).

(1) bear, chips, farmers, goal, like, therefore
(2) but, customer, have, I, to, was, wasn't, what's

They noted that reliability was not correlated with the number of phonemes in a word, but was generally lower for "high-frequency words, words that reduce". While this seems like a reasonable suggestion, we believe that it would be important to examine the function/content distinction between high-frequency words to understand any interactions. From the small sample of words in (1) and (2), it appears that the low-reliability words are more likely to be function words.

Fulop and Keating developed a contextual database of the transcriptions in their study of the Switchboard corpus. This database included information about each word's canonical dictionary pronunciation, in addition to other linguistic information, such as stress, syllabification, word boundaries, function/content distinction, and preceding and following phone and phonemes. The particular information provided was based on Withgott and Chen (1993). We will describe a similar database, also borrowing from a description in Keating et al. (1994), that we set up for analyzing our own results in section 5.1. Fulop and Keating built context trees to analyze the factors contributing to the varying realizations of phones. Their results summarized in Table 1-2.

Table 1-2: Factors contributing to varying realization of phones in Switchboard

Factor	Majority class	Phones
Preceding/following phoneme	vowels	æ, ʌ, ɛ, e, ɪ, o, b, f
Syllable structure position	resonants	l, n, r
position within word and/or syllable	plosives	d, k, t
No major context effects	stridents, labials	m, s, v, w, z

Source: Data from Fulop and Keating (1996).

Fulop and Keating also examined the extent to which surface phones matched phonemes in words' canonical dictionary pronunciations. Figure 1-1 organizes the dictionary phones according to the percentage of tokens that were transcribed using that phone. Contributing to the low faithfulness of stops to the dictionary pronunciation is a postlexical transcription system employing separate symbols for aspiration and release. In

section 5.1, we will discuss a different measure of variability, entropy, which does not suffer to the same extent on account of transcription artifacts between the lexical and postlexical levels.

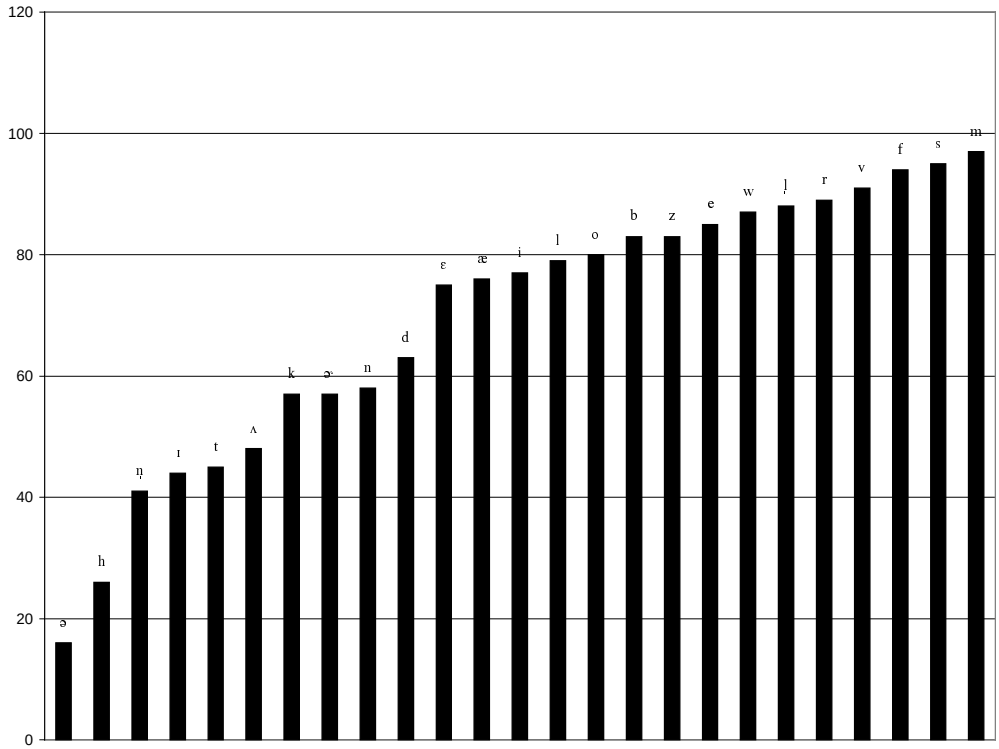

Figure 1-1: Percentage of tokens transcribed using canonical pronunciations

Note: Chart generated using data from Fulop and Keating (1996).

1.2.3. Review of neural networks in phonology

In this section, we will describe various uses to which neural networks have been put in phonology. These include using neural networks to convert orthography to phonetics,

and the use of neural networks as phonological representations themselves. Once again employing the concept of perspective established earlier, we note that the applications of neural networks have been both theoretical, in both psycholinguistics and formal phonology, and practical, for speech technology.

Letter-to-sound conversion refers to the conversion of orthography to pronunciation. It has applications in speech synthesis, for the pronunciation of out-of-dictionary words, and in database search, where it can be used to determine likely variants of a given spelling of a name (Divay and Vitale 1997). Historically, speech synthesizers designed to handle arbitrary text relied on a fairly small "exception" dictionary and letter-to-sound rules. This is the approach used in the MITalk project (Allen et al., 1987). In the case of MITalk, the exception dictionary consisted of "morphs" which were combined with each other and with affixes by means of morphophonological rules. The letter-to-sound rules were explicit contextual rules, roughly of the form: "orthographic *c* is phonetic /k/ unless it is before orthographic *i* or *e*, in which case it is phonetic /s/". Coltheart (1978) proposed a similar model for reading.

Hundreds of ordered letter-to-sound rules are required for acceptable accuracy in English. As computational power and memory have become cheaper, enlarging "exception" dictionaries to the point where they cover most of the vocabulary has become common (Liberman and Church, 1992). Nevertheless, alternative methods need to be available for out-of-dictionary words. Such words include many proper names, such as people, businesses and products, neologisms, domain-specific terminology (or jargon), as well as

abbreviations and acronyms. Hetherington (1995) discusses the out-of-vocabulary problem for speech recognition, and provides data on relative proportions of out-of-dictionary words stemming from such different sources.

Alternatives to letter-to-sound rules have been sought with the intention of reducing the labor of linguists involved in developing and maintaining ordered lists of rules. According to Church (1986), developing a rule-based approach requires "a few years of intense effort by a highly skilled expert. The end result is often very difficult to debug and to maintain."

One important step in moving beyond letter-to-sound rules was NETtalk (Sejnowski and Rosenberg, 1987) which employed a neural network trained on a commercial dictionary to determine pronunciations from text. Sejnowski and Rosenberg attempted to demonstrate how the network, when consulted in early stages of training, simulated a child's speech, and gradually achieved adult-like competence with more training. In a similar vein, Seidenberg (1989) posits a neural network based model for human lexical access and mediation between orthography and phonology.

Sejnowski and Rosenberg (1987) employed a "1 out of *n*" coding scheme for their orthographic input. That is, each letter was represented to the neural network as an atomic unit. In contrast, their phonemic output was encoded in a *distributed* manner, with phonemes represented as bundles of features.

While most prior neural network approaches to letter-sound conversion have focused on a single output (Sejnowski and Rosenberg 1987, McCulloch et al. 1987, Ainsworth and Warren 1992, Adamson and Damper 1996), neural networks can be designed to posit several candidate possibilities for a given input. For example, Deshmukh et al. (1996) describe a neural network implementation designed to provide n-best pronunciations for proper nouns.

Hare (1990) used a Jordan (1986) style neural network to investigate Hungarian vowel harmony, in particular, "transparent" vowels which neither display harmony themselves, nor block the spread of the harmonizing feature to other vowels. Since the same vowels can be harmonic in some contexts and transparent to harmony in others, the problem has typically been dealt with by positing a number of derivational sources for the segments in question. Hare (1990) first performs simulations on bit patterns and then does neural network simulations on bit patterns that represent the features of Hungarian vowels. She arrives at a notion of similarity, in which vowels that are most alike in certain features become more alike in others. She finds that this offers a plausible explanation of the harmony facts. Hare (1990, 148) summarizes the value of her neural network simulations as offering a "dynamic model that allows for testing of hypotheses about phonological processing".

Inspired by Hare (1990), Rodd (1997) used an Elman (1990) net to examine the phonotactics of Turkish vowel harmony. Rodd (1997) constructed nets with varying numbers of hidden layers. She examined how the hidden layers, upon presentation of

Turkish data, became detectors for various phonological features, including consonantal, frontness and sonority, which is involved in determining what segments may appear in consonant cluster sequences. We will examine neural network weights in the hidden layers of our postlexical network in an effort to see what features the network appears to make the most use of in determining postlexical variation.

Goldsmith (1992, 1993) discusses harmonic phonology, which is an attempt to use neural networks as a form of phonological representation. He uses network activations to model stress and syllabification patterns, modeling different systems by altering the network connection weights.

From a psycholinguistic perspective, Gaskell et al. (1995) describe a recurrent neural network designed to perform what they call "phonological inference" from surface forms to underlying lexical forms. Following Lahiri and Marslen-Wilson (1991), Gaskell et al. maintain that "phonological representations in the mental lexicon are structured in accordance with the theory of radical underspecification (Archangeli 1988)". In a manner similar to Gildea and Jurafsky (1996), Gaskell et al. (1995) introduced postlexical variation, in their case, place assimilation, into a database of lexical pronunciations that was in turn generated automatically (with hand corrections) from a text corpus. Shillcock et al. (1993) similarly generated postlexical pronunciations automatically from a text corpus, in an effort to study English phonotactics with a recurrent neural network.

1.3. General phonological issues

Our research into pronunciation modeling in speech synthesis dovetails with several areas in phonological theory. Many of our contributions will be at the interface between phonology and the other fields of linguistics, including phonetics, morphology and syntax. First, we will describe work on the phonetics-phonology interface, which bears upon our treatment of postlexical phonology. Subsequently, we will describe research on the interaction of phonology with syntax and morphology. That discussion will include a treatment of lexical phonology, with particular attention to various taxonomies of postlexical rules and their domains of application. Finally, we will describe the relationship between our work and the studies of phonological variation that have predominantly occupied sociolinguists.

1.3.1. Phonetics-phonology interface

Although they focus mainly on the phonetic interpretation of fundamental frequency, Liberman and Pierrehumbert (1984) advance a view whereby all postlexical processes are to be treated as "facts about the phonetic realization of phonological representations, rather than modifications of phonological representations themselves" (Liberman and Pierrehumbert 1984, 166). In contrast, Kiparsky (1985) argues that there are two types of postlexical rules, one with gradient *phonetic* outputs, the other with categorical *phonological* outputs. We will consider his evidence and examine other subdivisions of postlexical rules in the next section.

Our work on postlexical modeling suggests that there is at least some descriptive adequacy to describing postlexical phenomena in symbolic, categorical[7] terms typical of phonology, rather than the gradient terms of phonetics. We understand that this approach has certain limitations; however we feel that these can be compensated for in the synthesizer's acoustic module in ways to be described below. In a description of casual speech phenomena, Browman and Goldstein (1990) show how apparent insertions, deletions and substitutions of segments can be interpreted as modifications in gestures, some of which are rendered inaudible. We will describe an experiment designed to determine the relationship between the phonological-phonetic and acoustic-phonetic representations in Chapter 4. In this subsection, we will describe some of the issues surrounding the ascription of phenomena to phonetics, phonology or their interface. This debate has also been discussed in terms of the difference between *coarticulation*, a gradient phonetic phenomenon, and *assimilation*, a categorical, phonological phenomenon (e.g. Wood 1996).

Pierrehumbert and Talkin (1992) found that the varying, gradient, realization of /ʔ/ and /h/ was reminiscent of the mapping between underlying tones and surface fundamental frequency contours (Liberman and Pierrehumbert 1984). They found that allophony in /ʔ/ and /h/ was dependent on the prosodic environment in which they were located. For this reason, Pierrehumbert and Talkin argue against a view that segregates tone and

[7] Scobbie (1995) implies he prefers the term *categorial* to *categorical* in his use of *sic* in quoting Zsiga's (1995) use of *categorical*. Perusal of several dictionaries and articles on the topic at hand have not

intonation into a "suprasegmental" component; a view that they claim pervades speech technology (Lea 1980, Allen et al. 1987, Waibel 1988). In order to investigate further Pierrehumbert and Talkin's demonstration of prosodic effects on segmental allophony, we will describe in Chapter 5 the ways in which prosodic information is made available to the postlexical module of the Motorola speech synthesizer, whose function is to predict appropriate postlexical phones.

Pierrehumbert and Talkin conclude that an intrinsically quantitative representation, oriented toward critical aspects of articulation, appears to offer more insight than the traditional fine phonetic transcription. We will investigate the relationship between postlexical transcription and quantitative phonetic implementation in Chapter 4.

In a similar manner to Pierrehumbert and Talkin, Pierrehumbert and Frisch (1997) find that phrasal prosody plays an important role in determining allophonic glottalization, both in the case of vowel-vowel hiatus and voiceless stops before sonorants. In fact, they claim that successful synthesis of contextually appropriate glottalization requires an architecture with a running window over a fully parsed phonological structure. We will describe just such a running window in our postlexical neural network in section 5.4. Pierrehumbert and Frisch (1997) maintain that glottalization is an important allophonic phenomenon in English that is critical for natural-sounding synthesis. We will describe our success at learning when to glottalize in section 6.4.

---

convinced us that one term is preferable to the other, so we will use the more common *categorical* to mean "relating to categories". The connotation of "unconditional, unqualified" is not inappropriate.

Pierrehumbert and Frisch (1997) reiterate the importance of not separating segmental and suprasegmental phenomena. They show how control of fundamental frequency is important for realizing "segmental" glottalization. The Motorola synthesizer's acoustic neural network learns fundamental frequency at the same time as other acoustic parameters, avoiding this separation problem (Karaali et al. 1997).

Zsiga (1995) investigates the phone /ʃ/ in both lexical and postlexical contexts. In an acoustic and electropalatographic study, she finds that lexical /ʃ/, whether underlying as in *mesh*, or derived as in *impression*, appears to have categorical properties. In contrast, postlexical /ʃ/, derived in contexts such as *miss you*, shows a gradient progression between /s/ and /ʃ/ across time. In her analysis of these phenomena, Zsiga (1995) finds that phonological features best handle the categorical situation, while (overlapping) gestures (e.g. Browman and Goldstein 1990) best handle the gradient situation. In particular, she asserts that the pattern of overlap between /s/ and /j/ in *miss you* is such that no postlexical rule of palatalization is required; the effect of palatalization would simply be the acoustic consequence of the normal pattern of overlap.

Zsiga (1997) similarly distinguishes a set of categorical phenomena that are best handled with autosegmental-featural representations from a gradient set of phenomena that are best handled with Browman and Goldstein's gestural model. In particular, she finds that Igbo vowel harmony is categorical, while Igbo vowel assimilation is gradient. Zsiga (1997) concludes that when a straightforward and principled phonetic explanation exists, proposing phonological rules for gradient processes unnecessarily complicates the

phonology. This is consonant with Scobbie's (1995) response to Zsiga (1995), "What do we do when phonology is powerful enough to imitate phonetics?".

Sproat and Fujimura (1993) find that there is no reason to treat the light and dark allophones of /l/ as categorically distinct phonological (or phonetic) entities in English. They find that /l/ is phonetically implemented as a lighter or darker variant depending upon such factors as its position within the syllable and the duration of the prosodic context containing the /l/. As with Zsiga, they offer a gestural explanation for the gradience found. As with Pierrehumbert and Talkin (1992) and Pierrehumbert and Frisch (1997), they find that prosodic information, here, the intrasyllabic position of phonological elements, is necessary to explain the attested allophony.

In a manner analogous to Zsiga (1995) who criticizes "phonological" rules of palatalization, Sproat and Fujimura (1993) contrast their findings with those of Halle and Mohanan (1985) who derive dark /l/'s from light /l/'s via a phonological rewrite rule that adds feature [+back] to /l/'s in postvocalic position. Sproat and Fujimura (1993) ascribe to Halle and Mohanan the assumption that in order to describe how one goes from an abstract phonemic representation to actual pronunciation, one needs an intermediate level of representation where allophones are represented as categorical entities. Sproat and Fujimura (1993) claim that by arguing that much of the variation in quality in English /l/ allophones can be explained by a combination of factors including duration and intrasyllabic position, they have shown that the mapping from an abstract phonological

representation to actual pronunciation does not require the positing of a level of representation where [l] and [ɬ] are distinct phonological or physiological entities.

Sproat and Fujimura (1993, 309) suggest the burden of proof is on those who would claim that allophones are distinct to justify the necessity of that assumption:

> ...it should no longer be taken for granted that two allophones, even ones as apparently different as the [t^h] and [ʔ] of /t/ represent two distinct categories: we expect that variation in such cases will also prove to be graded, once enough contexts are considered and sufficient articulatory data are examined.

Our experiment regarding the allophony of /u/ and /ə/, described in Chapter 4, takes up this challenge.

1.3.2. Lexical phonology and the interface between syntax, morphology and phonology

Our delineation of the problem of determining appropriate contextual postlexical pronunciations from canonical lexical pronunciations is founded in the theory of lexical phonology (Kiparsky 1982). In this theory, lexical rules are seen as occurring on lexical items before they are inserted into contexts with other words. Postlexical rules occur both within and between words once such contexts are established.

Several researchers have looked more closely at postlexical rules in an effort to see whether further distinctions can profitably be made. The main themes for distinguishing

among postlexical rules include whether or not they need to refer to syntactic information and whether their application is gradient or categorical.

We will examine postlexical taxonomies suggested by Kiparsky (1985), Nespor and Vogel (1986), known as prosodic phonology, Kaisse (1985), involving P1 and P2 rules, Mohanan's syntactic and postsyntactic strata and Hayes's (1990) precompiled phrasal rules. We will also discuss relevant issues regarding the phonology-syntax connection in the work of Selkirk (1984).

As stated earlier, Kiparsky takes issue with Liberman and Pierrehumbert's (1984) claim that all postlexical rules are phonetic implementation rules, arguing that "at least some postlexical processes are truly phonological, feature-changing rules" (Kiparsky 1985, 86). To explain this point, Kiparsky (1985) provides examples of lexical and postlexical voicing in English. He illustrates anticipatory lexical voicing assimilation that is triggered by and applies to obstruents, as in a[dz] and a[ps], but not sonorants as in to[kn] or a[pl]. In contrast, he exemplifies perseverative postlexical voicing assimilation that can both apply to sonorants (1a) and be triggered by them (1b):

(1) a. c[r̥]y, p[l̥]ay, sp[l̥]it

b. back[t], bagg[d], bann[d], kidd[ɨd]
back[s], bag[z], bell[z], bush[ɨz]

Kiparsky takes (1a) to be uncontroversially postlexical, due to structure preservation, which "determines point-blank that any rule which introduces marked specifications of

lexically non-distinctive features must be postlexical" (1985, 93). (1a) is an example of a phonetic implementation rule, due to its gradient and variable output. He also takes (1b) to be postlexical, since the same process can occur to cliticized reduced forms of *is* and *has*. However, in contrast to (1a), Kiparsky (1985) takes (1b) to be "postlexical phonological". Anticipating Zsiga (1995) with respect to lexical and postlexical palatalization, he states that "gradient processes...form a 'cline'...(and) appear to be simply the postlexical applications of rules which in the lexicon function in strictly categorical fashion" (1985, 94).

Nespor and Vogel (1986) carried forward Kiparsky's attempts to subdivide postlexical rules with their elaboration of the prosodic hierarchy. It is important to establish the connection between this work and the theory of lexical phonology. Nespor and Vogel do not make it particularly explicit, merely noting in a parenthetical comment that "the rules of postlexical phonology correspond to the prosodic rules of the present proposal" (Nespor and Vogel 1986, 18). Nespor and Vogel make clear that they are not discussing phonological processes that must make reference to syntactic or morphological information (27-33), whether at or below the word level. In general, Nespor and Vogel can be seen as providing a principled way to explain phonological processes that occur across words. They provide a prosodic hierarchy of structures shown in Table 1-1.

Table 1-1: Nespor and Vogel's (1986) prosodic hierarchy

Domain	Symbol
syllable	σ
foot	Σ
phonological word	ω
clitic group	X
phonological phrase	ϕ
intonational phrase	I
phonological utterance	Y

Nespor and Vogel point out that the prosodic domains they describe do not necessarily coincide with syntactic boundaries; in fact, they find the latter insufficient to describe many cross-word phonological phenomena. They present various processes in several languages that are constrained to occur within the various prosodic boundaries they establish.

Kaisse (1985) analyzes the different types of postlexical phenomena from a different perspective from Nespor and Vogel. She recognizes three types of postlexical, or connected speech, phenomena:

1. Variants that can be accounted for by listing suppletive allomorphs in the lexicon; e.g. cliticization.
2. Variants produced by rules of external sandhi. These have syntactic, morphological and or lexical conditions in addition to morphological ones.
3. Fast speech rules, which are entirely dependent on rate and phonological information.

Kaisse calls (2) P1 rules, and (3) P2 rules. P1 rules occur before pause insertion and P2 rules. She claims that P2 rules are the postlexical rules of Mohanan and Kiparsky. Kaisse (1990) suggests that P1 and P2 rules differ in the same way that lexical rules are considered to differ from postlexical rules in the "classical" sense. According to this conception, P1 rules are categorical, while P2 rules are gradient.

Kaisse (1990) compares her postlexical typology with that of Nespor and Vogel (1986). She claims that most of the P1 rules she concentrated on in Kaisse (1985) apply within the phonological phrases of the prosodic hierarchy. However, she asserts that rules operating within clitic groups and phonological phrases belong to a different prosodic hierarchy than those operating within intonation phrases or utterances. According to Kaisse (1990, 129), "the larger prosodic domains are less grammaticalized than clitic groups or phonological phrases. They are not nearly so intimately related to, nor derived from, syntactic categories and syntactic concepts". In other words, the rules applying in the smaller prosodic domains would be P1 rules, while those applying in the larger domains would be P2 rules.

Randolph (1989, 85-86) distinguishes between intrinsic and extrinsic allophones. Extrinsic allophones refer to "phonetic alternations that are attributed to the structural context of an underlying phoneme," for example, its position in a larger phonological unit, such as a syllable. In contrast, intrinsic allophones refer to phonetic alternations resulting from "inherent mechanical constraints imposed on the articulatory mechanism". Randolph (1989) mentions that speaking rate and dialect also play a role in a speaker's

selection of extrinsic allophones. As examples of extrinsic allophones in English, Randolph (1989) mentions the aspirated, unaspirated, glottalized, released, unreleased, flapped and retroflexed variants of /t/, light and dark /l/, and syllable-initial and syllable-final /r/. Although Randolph (1989, 93-94) expresses some discomfort with qualitative allophonic symbols, he makes use of them in a series of experiments on stop realization. While allophonic symbols may be useful for analyzing extrinsic allophony, Randolph (1989, 85) observes that to describe intrinsic allophony, speech must "be treated at the physical level, as multi-dimensional, where the dimensions pertain to individual vocal tract articulators".

Mohanan (1986, 145) proposes that the postlexical module consists of two submodules, namely a syntactic submodule in which syntax is available, and a postsyntactic module of phonetic implementation, in which syntactic information is not available. Mohanan mentions English *a/an* alternation as an example of a rule taking place at the syntactic stratum. Mohanan (1986) claims that this is an improvement over Mohanan (1982), where there was no division in the postlexical module. In the earlier version, *a/an* alternation was assigned to the lexicon, a strategy Mohanan (1986, 180) calls a "contrivance".

Interestingly, Hayes (1990) proposes the solution that Mohanan (1986) rejects. Hayes notes that prosodic phonology (Nespor and Vogel 1986) does not handle rules that require syntactic information, which he calls "direct-syntax rules". He notes that the avowed inability to handle such rules makes the theory harder to falsify, since if the prosodic

hierarchy cannot explain a particular phrasal phonological phenomenon, syntax can be appealed to, potentially indiscriminately. As an alternative, Hayes proposes to eliminate postlexical rules that refer to syntactic information, by precompiling diacritically-marked allomorphic information in the lexicon. According to Hayes (1990, 87), "at the interface of syntax and phrasal phonology, the appropriate diacritically-marked allomorphs are inserted in the relevant syntactic contexts. This is similar to Liberman's (1994a) suggestion that phonological selection occur on the basis of comparison of allomorphs in a single optimality tableau.

Liberman (1994a) discusses the concept of phonological optionality using the example of Latin enclitics. He distinguishes phonological optionality from both sociolinguistic (discussed in section 1.4) and phonetic variation (discussed in section 1.3.1). Our own study approximates that of Liberman in the sense that sociolinguistic variation is minimized due to the use of a corpus of read speech (section 3.2), and we are analyzing the attested variation in phonological terms (with the caveats expressed above). Proper investigation of phonological optionality requires substantial examples in varying contexts:

> ...investigations of this sort are most comfortably carried out in the context of a suitably-annotated corpus. In this context, propensities can be estimated from observed frequencies, and proposed conclusions can be checked or challenged by other scholars. (Liberman 1994a, 88)

Liberman observes that the variation he observes in certain prepositions does not constitute a regular phonological process in Latin. For this reason, he proposes that the variation observed in his study is best considered allomorphy.

Another case in which certain phonological variation occurs only in a subset of the vocabulary is the case of function words. Hayes (1995, 88) discusses McCarthy and Prince's (1986, 1990) observation that word minimality predictions typically hold only for content words, e.g. nouns, verbs and adjectives. Function words can be subminimal, such as English *a* [ə] and *the* [ðə]. Selkirk (1984), in her chapter on "Function words: destressing and cliticization," notes that function words have "strong" and "weak" forms (335). The weak forms are characterized by stresslessness, and the possible deletion of vowels or consonants. Selkirk notes the phrasal aspect of weak form realization, observing that function words "must be sufficiently 'close,' syntactically speaking, to what follows (or, in a few cases, to what precedes)" (336). In these cases, the weak forms may be thought of as clitics. Furthermore, function words exhibit particularly close juncture, increasing their likelihood of participating in sandhi rules.

Selkirk attributes the special behavior of function words to a *Principle of the Categorial Invisibility of Function Words* (PCI). While the reduction of function words bears a resemblance to lexical vowel reduction, it seems to us to be a clear case of postlexical

phonology, given its instantiation in phrasal contexts. In section 6.5, we will examine the various behavior of function words in our corpus, attributing many of their alternations to allomorphy, since the processes they undergo and the contexts in which they undergo them are often specific to function words, or even particular function words (e.g. the discussion of *the* in section 6.5).

As we will describe below, our stipulation of the problem of modeling postlexical variation conflates to some extent what has traditionally been considered allophony (such as flapping), phonologically conditioned allomorphy (such as [ðə]/[ði] alternation), and function word destressing. We provide the syntactic (such as part of speech[8]) and prosodic information that the learning procedure can use, without stipulating how it must be used (cf. Gildea and Jurafsky's 1996 introduction of learning biases to a phonologically learning problem). Building on the work of Pierrehumbert and her colleagues (Pierrehumbert and Talkin 1992, Pierrehumbert and Frisch 1997), we can consider the space of postlexical variation that we set out to learn as *prosodically* conditioned allophony and allomorphy.

1.4. Sociolinguistics/Variation

In this subsection, we will discuss both the relationship of our work to sociolinguistics and theories of variation in general. As discussed above with respect to the observations of Liberman (1994a), sociolinguistic variation may be carefully distinguished from both

phonological optionality and phonetic variation. Sociolinguistic variation includes pronunciation variation arising from stylistic, geographical and social class variation. Perhaps a convenient way of grouping geographical and social class variation would be to call them "speech community variation", since we have learned that even within the same geographical region, there is a complex interaction between sex, race, social class and other factors (e.g. Labov 1966). While it is clear that stylistic variation is an issue both within and across speakers, speech community variation can be subject to the same vacillation, when one considers phenomena such as *accommodation* (e.g. Street and Giles 1982), or *audience design* (Bell 1984).

Another dimension along which variation can be categorized is that of inter- and intraspeaker variation. The approach to modeling individual postlexical variation that we will present should be contrasted with approaches in both sociolinguistics and speech technology that model inter- rather than intraspeaker variation. In sociolinguistics, Guy (1980) showed that in the case of *t,d* deletion, variation in the individual tends to mirror that of the speech community. However, Van de Velde and van Hout (1997) show how different explanations of the variation in Dutch *n* deletion result from an examination of individual, as opposed to aggregate, data. Approaches to pronunciation modeling in speech recognition (e.g. Tajchman et al. 1995), attempt to predict a range of dialect and postlexical phenomena across speakers.

---

[8] Actually, *prominence* (O'Shaughnessy 1976), as discussed in section 3.2.1.

It is important to clarify how the issues of intra- and interspeaker variation affect our own work. Our phonetically labeled database contains the speech of one speaker. Concatenative and neural network synthesizers (*e.g.* Karaali et al. 1996, Gerson et al. 1996, Corrigan et al. 1997, Karaali et al. 1997) have proceeded under the assumption that they are more likely to be considered natural if they strive to learn the habits of one particular speaker, rather than if they attempt to average over several. For recognition, we might want to expand our postlexical model to include variation across speakers, however that is beyond the scope of the present research.

In section 2.3 we discuss Labov's (1989, 1994) research on cross-dialectal comprehension. It is important to distinguish between lexical and postlexical variation in such studies. Labov's studies of vowel shifts (Labov 1991), many of which played a significant part in his cross-dialectal comprehension studies, are of the lexical variety—that is, words in a speaker's lexicon would presumably be stored with the vowels in question. In contrast, much sociolinguistic research has concentrated on postlexical variants, such as studies of *t,d* (e.g. Guy 1980[9]) deletion in English or *s* deletion in Spanish (e.g. Poplack 1980).

Kiparsky (1995) has noted that historical linguistic changes of the lexical diffusion variety are often lexical, while examples of neogrammarian sound change are often postlexical. In parallel with the gradience characteristics of lexical and postlexical

[9] Although Guy (1991) analyzes some instances of *t,d* deletion as lexical and other postlexical.

phonology discussed earlier, Kiparsky finds that regular sound change is gradient, and lexical diffusion quantal (1995, 643).

Variation has been regarded as a nuisance both in formal linguistic theory (Chomsky 1965) and in acoustic phonetics (Klatt 1980). While some variationist work, such as Guy's (1991) explication of the exponential model, have proved the lawfulness of variation with respect to formal theories, separate traditions in both speech technology and speech perception have found acknowledgment of the inherent variability of speech to be beneficial, as we will elaborate below.

Church (1983) contrasted his work on phonological parsing for speech recognition with a tradition that had sought to characterize the automatic speech recognition problem as one in which variable signals needed to be mapped to an invariant, symbolic level, such as phonemes. Church showed how allophonic variants, particularly those involved in juncture phenomena in allegedly confusable sentences, actually provided important keys to their semantic resolution. Elman and McClelland (1986) show how variability on a spectral, or featural, level actually improves recognition performance, by providing important cues as to neighboring segments and coarticulatory phenomena.

Pisoni (1993, 1997) contrasts newer work in speech perception where "episodic" memory is seen as critical in the speech perception process, rather than "analytic" memory. The episodic approach is backed by findings that listeners understand familiar voices better than unfamiliar ones. Rather than reducing speech to invariant cues in order to decode it,

listeners are seen as benefiting from indexical or paralinguistic aspects of the signal in their understanding efforts.

While the views of Elman and McClelland, Church and Pisoni are complementary, it is important to distinguish between linguistic variability, as is characterized by varying treatment of allophony by speakers, and paralinguistic variability, which is evidenced by different voice qualities, for example. The dimensions of style (Abe 1997) and emotion (Murray and Arnott 1995) have also been investigated in the context of synthetic speech. In a speech synthesis application, the goal may be to produce several voices, each one internally coherent. Each voice should command a range of styles appropriate to different topics and interlocutors. In speech recognition, the ideal system must be able to handle a variety of speakers speaking in a variety of styles about a variety of topics.

Speech technology is beginning to show an acute interest in variation. For example, the field of speaker adaptation within speech recognition attempts to identify speaker characteristics in an effort to be able to use more specific pronunciation models (Zhao 1997, Ström 1997). Research into automatic dialect classification is one example of this approach (Huggins and Patel 1996, Miller and Trischitta 1996). Undoubtedly, projects such as the Phonological Atlas of North America (Labov 1996) will inform these investigations. Junqua and Haton (1996) provide a survey of current research on style variation in the context of speech recognition. We hope to contribute to this growing research area.

As high-quality synthetic speech becomes more of a reality, the question of how to introduce what we have learned about variation becomes a question of critical importance. If indeed variation is as important to understanding and explaining language as we have been saying it is, how should we apply our knowledge to speech generation systems? In order to approach this issue, we need to adjust our perspective to describing variation in such a way that it is possible to allow it to serve as a model for a synthetic voice.

1.5. Overview of dissertation

In this chapter, we have circumscribed the problem of pronunciation modeling in speech synthesis. We have shown how we intend to explore it within the context of the postlexical module of a speech synthesizer being developed at Motorola. We have described the developing field of computational phonology and previous efforts at pronunciation modeling on corpora. We have reviewed literature in phonetics, phonology and their interface with syntax and morphology that we believe provides useful means for describing and explaining the kinds of postlexical variation we intend to investigate in this dissertation. In addition, we have examined sociolinguistic and psycholinguistic treatments of variation in an effort to explore the range of pronunciation variation and the factors that control it.

In chapter 2, we will provide rationale for modeling postlexical variation in the way we are proposing. Chapter 3 will cover the data sources involved in our experiments,

including a lexical database and a hand-transcribed speech corpus of one speaker. In chapter 4, we describe an experiment designed to explore the question of gradient and categorical treatments of allophony posed in this chapter. In chapter 5, we characterize the learning problem, and explain the methods we are using in our attempts to solve it; in particular, a neural network designed to map from lexical to postlexical pronunciations. In chapter 6, we provide a phonological analysis of our results, breaking them down by phenomena. In each case, we attempt to relate our findings to other experimental and theoretical reports. In chapter 7, we provide a discussion on the sum of our results, and consider their implications on both linguistics and speech technology.

Chapter 2. Rationale for modeling postlexical variation

In this chapter, we will discuss some of the most important reasons for modeling postlexical variation in the way that we propose for speech synthesis. First, we will discuss the various parameters for evaluating synthetic speech, in an effort to see what past research indicates about the importance of proper modeling of postlexical variation. We will then consider specific aspects of the Motorola speech synthesizer that motivate the approach we have taken. We then look to sociolinguistic studies of cross-dialectal comprehension for motivation of the dialect modeling that our procedure allows.

2.1. Evaluation of synthetic speech

There are four main axes along which speech synthesizers have been evaluated: intelligibility, comprehensibility, acceptability/preference/quality, and naturalness. Intelligibility is usually assessed on the segmental level, and to determine it, listeners are asked to identify, either by transcription or in multiple choice format, particular words with which they are presented. Comprehensibility usually refers to understanding of passages longer than the word; for example, sentences or paragraphs. In comprehensibility tests, listeners may be asked to either repeat or answer questions about what they hear.

Finally, acceptability (also known as quality or preference) refers to the reaction of listeners to the speech— whether they like it or not, or whether they like one synthesizer

more than another. Naturalness is related to acceptability measures; however, attempting to isolate it represents an attempt to separate intelligibility from acceptability. Of course, the notions of intelligibility, comprehensibility, acceptability and naturalness are all intimately related; however, it is useful to identify the different sources of listener behavior with respect to synthetic speech. We will first review some general information about the evaluation of synthetic speech. Then we will examine some case studies in evaluation which bear upon our interest in modeling postlexical variation.

2.1.1. Intelligibility

According to Ralston et al. (1995), "intelligibility provides an index of the lower bounds of perceptual performance for a given transmission device when no higher-level linguistic context is provided." Schmidt-Nielsen (1995) provides a review of intelligibility tests, many of which are employed in various telecommunications applications as well as synthetic speech. Common tests include the Modified Rhyme Test (MRT) (House et al. 1965) and the Diagnostic Rhyme Test (DRT) (Voiers 1983). Both are closed-response tests, with the MRT offering six choices for each item, the DRT offering two. Schmidt-Nielsen notes that as the size of the response set decreases, intelligibility scores will be higher. The DRT provides a diagnostic score on several acoustic features (Jakobson, Fant and Halle 1952).

In a study comparing the Motorola speech synthesizer to several commercial synthesizers, Nusbaum, Francis and Luks (1995) preferred an intelligibility test using a transcription

task in which subjects typed in what they heard. They claim that this allows subjects "a better way of describing the segmental structure of their perception" compared to a forced choice task, as in the MRT. For the intelligibility experiment, subjects were played single words from several synthesizers and one human talker. They were instructed to type into a computer the word they heard in normal spelling, and to type a nonsense word if it did not sound like a word.

While we believe that more natural speech, such as is the result of using postlexical, rather than lexical pronunciations, results in increased intelligibility, this is not necessarily obvious. It might be argued that hyperarticulated pronunciations give listeners a better clue than those that are more casual as to what is being said in the context of poor-quality synthetic speech. It will be important to demonstrate that the acoustic, or signal processing, side of the synthesizer is of sufficient quality that using natural postlexical pronunciations actually improves intelligibility, as we believe it does in natural speech. Proof of these points remains for future work.

Bradlow et al. (1996) claim that speakers with larger vowel spaces are more intelligible than those who have smaller vowel spaces. They also point out how fine acoustic-phonetic differences affect the intelligibility of consonant clusters, particularly with respect to the deletion and release of /t/ and /d/. They also show how listeners' ability to detect syllable affiliation is related to the duration of /s/ between words. Such results lead us to believe that attaining high intelligibility will require more than simple manipulation of the symbolic, or segmental characteristics of postlexical speech, but also

complex manipulation of acoustic phonetic characteristics. In future work, we hope to show how a neural network approach to generating postlexical symbols and a phonetic implementation neural network can achieve the required fineness of distinctions, resulting in maximized intelligibility.

Research into "clear speech", a variety of speech which is hyperarticulated, (for example, the way some people address the hard of hearing) reveals that such speech is more intelligible than normal speech (Picheny et al. 1985, 1986, Krause 1995). Such results might lead one to suspect that speech synthesizers are likely to be more understood the more hyperarticulated, and consequently less casual, they sound. However, we believe that over longer discourses, more conversational speech will prove to be more comprehensible. We hypothesize that hyperarticulated speech would be likely to distract attention from the content of the synthesized message by calling undue attention to how it is uttered.

2.1.2. Comprehensibility

According to Ralston et al. (1995, 240), most models of comprehension are based on data obtained from studies of reading, which typically try to account for only a subset of the factors known to influence comprehension. An important difference between the modalities of listening and reading is that written text is typically static and available for re-inspection, whereas spoken language is very much transitory and ephemeral.

According to Kintsch and van Dijk (1978), comprehension is the process of building up a coherent, connected propositional representation of a text's meaning.

Ralston et al. (1995) discuss the distinction between simultaneous and successive measures of comprehension:

> Successive measures of comprehension are those made after the presentation of linguistic materials. Simultaneous tasks, on the other hand, measure comprehension in real-time as it takes place and usually require subjects to detect some secondary event occurring during the time that the comprehension takes place. (Ralston et al. 1995, 247)

Successive measures have been used much more than simultaneous measures, and they appear to be less sensitive, due to memory decay and reconstruction of information absent in the original materials.

Among successive measure of comprehension, Ralston et al. (1995) identify recall, recognition and sentence verification. Recall involves repeating materials after they have been heard. Recognition tasks assess memory for higher levels of representation, such as propositions derived from text. These include the familiar standardized multiple-choice reading comprehension tests. Finally, sentence verification tasks most approximate simultaneous measures of comprehension. Sentences are presented to listeners who must respond whether they are true or false. Since the judgments are most often trivial, such as "A robin is a bird", the main dependent variable of interest is response latency. Ralston et al. note that successive measures of comprehension have yielded mixed results on

comparisons of synthetic and natural speech. Table 2-1 summarizes some of these results.

Table 2-1: Mixed results on successive comprehension measures

Results	Report
Reliable effects of voice on accuracy, but no training effects	Luce (1981), Moody and Joost (1986)
Differences between voices when synthetic speech first encountered, but differences became smaller with even moderate exposure or training	Jenkins and Franklin (1981), McHugh (1976), Pisoni and Hunnicutt (1980)
No comprehension differences between natural and synthetic speech, nor learning effect after training	Schwab et al. 1985

Source: Ralston et al. (1995, 265).

Simultaneous measures of comprehension include both phoneme and word monitoring tasks, where listeners are asked to identify when a particular phoneme or word occurred in a stream of speech. As in the sentence verification task, the main dependent variable is response latency. Ralston et al. (1995) summarize the results of several experiments including simultaneous measure of comprehension, which they claim are the first reported (Ralston et al. 1991). These experiments showed that natural speech was consistently higher than low-quality synthetic speech (Votrax) on word monitoring accuracy and latency, as well as word and proposition recognition. A significant correlation was found with results from an MRT intelligibility task. An interesting finding regarding successive comprehension measures performed in the same test battery

was that subjects listening to natural speech were more accurate for proposition-recognition sentences than word-recognition sentences, while subjects listening to Votrax speech had the reverse behavior. Ralston et al. (1995) explain this asymmetry between natural and synthetic speech in the following passage:

> The data are consistent with the hypotheses that comprehension is constrained by a limited capacity of processing resources that can be allocated to various comprehension sub-processes as well as other cognitive processes. Subjects listening to synthetic speech apparently allocate a greater proportion of the available resources to acoustic-phonetic processing, leaving relatively fewer resources for analysis of the linguistic meaning of the passage. (Ralston et al. 1995, 271)

We believe that this finding has important ramifications for our work. Obviously, in most practical uses, speech synthesis will be used to communicate propositional content. Therefore, it is important to reduce the amount of attention that listeners must pay to the acoustic-phonetic properties of the speech in order to derive such meaning from synthesized text. We hope that over longer stretches of speech, naturalness, such as that engendered by attention to postlexical detail, will improve comprehensibility and reduce listener fatigue.

Ralston et al. (1995) note some general problems with comprehension studies, especially those that involve responding to published multiple choice tests. Since results on these tests have a good chance of being related to prior real-world knowledge, they suggest pre-screening test questions and discarding those that subjects can correctly answer without listening to the corresponding passages. Better yet, they recommend using fictional passages.

2.1.3. Acceptability

Nusbaum, Francis and Luks (1995) found that the Motorola research synthesizer was judged more acceptable by listeners than three commercial synthesizers. They assumed that acceptability ratings reflect several different characteristics of the speech, "including how well the listeners understand the message and how pleasant (or perhaps appropriate) the speech sounds." However, the Motorola synthesizer lagged behind the others in segmental intelligibility. Schmidt-Nielsen (1995) also notes the possibility for a mismatch between acceptability and intelligibility. She notes that while acceptability (or quality) measures are often highly correlated with intelligibility, there are situations where intelligibility may be high, but speech quality is degraded. For example, noise removal techniques can improve acceptability but tend to lead to lower segmental intelligibility. Given the potential mismatch between intelligibility and naturalness, we hope to investigate ways in which the intelligibility of a system can be improved, without losing naturalness and acceptability.

For the acceptability experiment, subjects were played speech in several domains where synthetic speech might be used, such as finding out movie times, or getting account information from a bank. Subjects were asked to rate the acceptability of each sentence for that domain on a scale of 1 to 7. A rating of 1 was described as "I would not use a service that used this voice", while a rating of 7 was described as "I would prefer to use a service with this voice". A rating of 4 was described as "This voice is acceptable".

Schmidt-Nielsen (1995) discusses another kind of preference test, the paired comparison test. In such a test, the listener hears a sentence for each of two speech conditions and selects the one that is preferred. All pairs should be presented twice so that each system is presented both in first and second place. However, Schmidt-Nielsen notes that the use of ratings or category judgments instead of paired comparisons greatly simplifies the data collection since each system being tested need be rated only once for each speaker. According to Schmidt-Nielsen (1995, 210), experimental evidence and informal experiences indicate that rank orderings assigned by use of rating scales and by paired comparisons are highly correlated.

2.1.4. Naturalness

Speech synthesizers demonstrating knowledge of postlexical variation are likely to be perceived as more natural than those that do not. According to Allen et al. (1987, 87), postlexical rules (in their terms, phonological recoding rules) "are not 'sloppy speech' rules, but rather rules that aid the listener in hypothesizing the locations of words and phrase boundaries." However, as Sorin (1991) demonstrates with respect to the French mute *e*, given the current state of the signal processing end of speech synthesis, simulating natural pronunciations may result in a loss of intelligibility. There is an obvious tension between naturalness and intelligibility, and many current synthesizers have opted for intelligibility at the expense of naturalness. Increased acceptance of synthetic speech will most likely hinge on an improvement on both measures.

Nusbaum, Francis, and Henly (1995) sought to tease apart the notion of naturalness from intelligibility and acceptability, with which it is often confounded. They believed that people can readily hear the source difference between natural speech and synthetic speech, from global levels of prosody to local acoustic-phonetics of spoken words. Nusbaum, Francis and Henly note that from a practical perspective, unnatural voice quality of speech seems more of an aesthetic issue than one that is important to determine the usability of synthetic speech.

However, they note that as the intelligibility of synthesis improves, the naturalness of synthetic speech becomes increasingly more important. As we will discuss in more detail below, this notion also motivated Sorin (1991) to reintroduce the French mute *e* into the CNET (Centre National d'Etudes des Télécommunications) synthesizer once a certain intelligibility threshold had been reached. According to Nusbaum, Francis and Henly, naturalness is particularly important for voice prostheses, and will be an important factor in the acceptability of synthetic speech for use in a wide range of applications.

According to Nusbaum, Francis and Henly, naturalness is a voice quality that is purely subjective; however, some of the factors that may influence the perception of naturalness can be specified analytically. Synthetic speech differs from natural speech in prosodic and segmental structure, as well as source characteristics. Since Nusbaum, Francis and Henly assume that segmental duration and timing, intonation, and amplitude variation are under the control of rules, they claim that "the patterning of these sources of information may show less variability than human speech" (Nusbaum, Francis and Henly 1995, 8).

They note that there are many opportunities for oversimplification and error in the rules of a text-to-speech system, both at the level of acoustic phonetic rules and phonological rules. We hope to show that since our system is based on machine learning of postlexical variation, rather than rules, we can introduce more natural variation into our synthetic speech than would otherwise be possible.

Nusbaum, Francis and Henly attempted to develop tests of naturalness that would eliminate as much as possible the contribution of intelligibility, since at present, the intelligibility differences between natural and synthetic speech are still sufficiently large to affect the perception of naturalness. They also sought to develop tests that would target specific aspects of naturalness, rather than simply provide global ratings. To this end, they developed separate tests to assess the naturalness of source characteristics and the naturalness of lexical prosody.

In their first experiment, on the naturalness of source characteristics, Nusbaum, Francis and Henly sought to reduce contributions of segmental structure and prosody to the perception of naturalness as much as possible. Their view was that even at the level of an individual glottal pulse, there would exist differences between natural and synthetic speech perceptible to listeners. The experiment involved taking a single glottal pulse and concatenating it to produce a sustained vowel, thus eliminating the effects of prosody. By focusing on maximally discriminable single vowels isolated from context, they claim to have eliminated many of the effects of intelligibility on perception.

By iterating a single glottal pulse to form a sustained vowel, Nusbaum, Francis and Henly note that they are eliminating one aspect of source characteristics that could be important to the perception of naturalness, namely the contribution of the perception of variability between glottal pulses. To assess the contribution of variability, they created a second set of stimuli based on sample of five successive glottal pulses that were iterated.

Nusbaum, Francis and Henly created vowel tokens by both concatenating single and five glottal pulses for /i/, /a/ and /u/ spoken by two synthesizers, DECtalk and Votrax, and two male talkers. Each vowel was played to subjects along with a text message identifying the vowel (thus eliminating any need for intelligibility judgments). Subjects were instructed to listen to each vowel and to decide whether it was produced by human or computer.

An analysis of variance revealed that there were no reliable differences in classification performance between the vowels created from the concatenated sets of five glottal pulses compared to those created from a single glottal pulse. Despite this result, Nusbaum, Francis and Henly believe that variability does play a role in naturalness; however, they observe that perhaps five glottal pulses were not enough to demonstrate it.

Nusbaum, Francis and Henly found that only /i/ proved to be diagnostic of naturalness, in the sense that subjects consistently judged the human speech for this vowel more natural than the synthetic speech. In addition, Votrax was judged less natural than DECtalk. Interestingly, DECtalk was actually judged more natural than human speech for /u/.

Nusbaum, Francis and Henly performed a second experiment, designed to assess the contribution of lexical prosody to naturalness judgments. To eliminate the influence of intelligibility, they low-pass filtered spoken words to eliminate as much segmental information as possible, leaving only prosodic structure. As in the first experiment, subjects were instructed to listen to a word and judge whether it was spoken by a human or a computer. The test words included words consisting of mono-, di- and polysyllabic words of varying lengths spoken by DECtalk, Votrax and two male talkers. Results showed that across all words, listeners judged human speech the most natural, followed by DECtalk and Votrax.

Since the measures of naturalness proposed by Nusbaum, Francis, and Henly rely on the elimination of segmental variation from the test materials, it would be difficult to use such methods to assess the contribution to naturalness provided by the postlexical module described in this dissertation, since the phenomena being learned are segmental in nature. In the following section, we will examine various ways in which the contribution of natural postlexical variation to synthetic speech has been assessed.

2.1.5. Case studies involving postlexical variation

We will now review two studies in synthetic speech which have sought to employ models of postlexical variation. Portele (1997) presented five phoneticians and six naive listeners with six sentences produced by a German speech synthesizer. He was focusing on German schwa deletion in the context stop-schwa-sonorant. Each synthetic sentence was

produced in two forms, one with reduced vowel variants and the other without reduced vowel variants. The phoneticians preferred the reduced forms, while the naive listeners preferred the unreduced forms. According to Portele, "Synthetische Sprache wird jedoch anders perzipiert als natürliche Sprache, bei letzterer werden kanonische Realisierungen kaum toleriert, während synthetische Sprache durchaus 'deutlich' klingen darf".[10] From this study, it appears that language professionals may have different, perhaps paternalistic, notions about what people want to hear from a speech synthesizer, and these judgments may not necessarily accord with those of the general public.

It is also possible that synthetic speech is still too unrealistic (i.e. not human sounding enough) to support very human characteristics like vowel reduction. That is, people may not be comfortable with a machine behaving in a manner they deem too human. Even a machine like Data on *Star Trek: The Next Generation*, who is human in so many respects, suggests that he is a machine by his failure to use contractions.

Sorin (1991) reports on the problem of mute *e* in French speech synthesis. She reports that the problem had previously been "hidden" due to the poor quality of synthetic speech. In other words, when a synthesizer's basic segmental intelligibility is still in question, it probably does not make sense for researchers to spend time worrying about certain issues of postlexical variation. Now that quality had improved, she sought to address variation in mute *e*, in an effort to make the CNET synthesizer sound more natural.

[10] "However, synthetic speech is perceived differently from natural speech. In natural speech, canonical

Sorin (1991, 148) uses the term mute *e* for designating "those occurrences of /ø/- or /œ/-type vowels which can be either pronounced or omitted in French without modifying the meaning of the word or word sequence", thus combining the varieties of *e* called by Delattre (1968) 'muet' and 'instable' or 'caduc'. Table 2-1 provides examples of the variable in question.

Table 2-1: Mute *e* in French

Sorin (1991)	Delattre (1968)	Description	Examples
mute	muet	almost always elided	méd**e**cin
	caduc, instable	not always elided	l**e**, appart**e**ment

Source: Sorin (1991, 148). Mute *e*'s in bold.

Most of the time, in colloquial French, the mute *e* is elided if its presence does not facilitate the pronunciation of the word or word sequence. A previous version of the CNET synthesizer only elided those mute *e*'s that appeared at the end of polysyllabic words. While it was thought that intelligibility was maximized by this procedure, listeners perceived the behavior as dysfluent or stumbling.

Sorin ran an experiment with conditions with differing treatments of mute *e* by the synthesizer. The main generalizations derived from this test were:

realizations are hardly tolerated, while synthetic speech must sound thoroughly precise."

(1) When mute *e* is elided in synthetic speech, correct word identification is substantially lower than that obtained with natural speech pronounced the same way.
(2) Systematic pronunciation of intermediate mute *e* in synthetic speech leads to identification scores virtually identical to those obtained with natural speech, in which all examples of mute *e* are elided.

However, when subjects were asked to give a preference rating to the synthetic versions with and without elision, the one with all pronounced mute *e*'s pronounced was judged worst— it did not sound fully natural. While Sorin's results show that introduction of some natural postlexical variation can lead to improved user preference, they also show that there may be an intelligibility cost to the introduction of that variation.

Nusbaum, Francis and Luks (1995) report a similar result to Sorin's in an evaluation of several speech synthesizers, including two from Motorola. Nusbaum, Francis and Luks performed two sets of experiments, one set designed to determine the ranking of synthesizers with respect to intelligibility and another set designed to determine the ranking of the synthesizers with respect to acceptability.

The Motorola synthesizers did reliably better (at $p < .05$) in acceptability than the other synthesizers. In contrast, the Motorola synthesizers performed reliably worse than two of the other synthesizers in intelligibility. In the past, acceptability had been mostly a measure of intelligibility, since in general synthesizers suffered from severe intelligibility problems. Today, intelligibility differences between synthesizers are less glaring, so acceptability judgments may be based more on judgments of naturalness, voice quality

and prosody. The fact that the Motorola synthesizers were lower ranked for intelligibility than acceptability is cited as proof that listeners were judging acceptability and intelligibility along different lines.

Nusbaum, Francis and Luks were surprised that the most acceptable synthesizer was not found to be the most intelligible. They report that this is the first such discrepancy that they are aware of. However, it seems to us that Sorin experienced the same result— the more natural sounding synthesis was in fact not the most intelligible.

2.2. Approximation to training data

Due to the method of training the acoustic module of the Motorola speech synthesizer, described in section 1.2, the approximation of testing data to the data upon which the system was trained is particularly important for synthesizing speech that resembles the original training data. The system is tested, or run, based on text input from users. This input is then looked up in a lexical database, to be described in detail in section 3.1. Rather than synthesizing lexical pronunciations directly, the lexical pronunciations from the dictionary are transformed into postlexical pronunciations typical of the speaker whose voice is being synthesized by means of the neural network based postlexical module.

Performing this transformation not only ensures that output pronunciations will accurately reflect postlexical processes for connected speech, it also ensures that the phonetic data that the acoustic neural network was trained on is matched as much as

possible by the phonetic data produced from text. For example, consider that a speaker whose voice is being modeled has a merger of vowels before /r/, exemplified by pronouncing *Mary*, *marry* and *merry* as [meri]. This means that the acoustic neural network has never seen the sequence [ær] as in some easterners' pronunciation of *marry*, and thus if it were asked to synthesize such a sequence, it would not benefit from the accuracy of training data that it would in sequences that it had seen before.[11] However, if the postlexical module, which is trained on transcriptions of the speaker being synthesized, transforms the tautosyllabic [ær] sequence that is likely to be found in dictionaries to [er], then the phonetic neural network will be tested on a sequence with which it is familiar, improving the likelihood of especially accurate and natural synthesis of such a sequence.

2.3. Cross-dialectal comprehension

Modern text-to-speech systems utilize a pronunciation dictionary for pronunciation of most words (Liberman and Church 1992). Thus, the kinds of pronunciations provided in such a dictionary are critical to the synthesizer's performance. Several studies have aimed to capture dialect variation in synthetic speech by providing a means for storing different pronunciations for different dialects in the synthesizer's lexicon (Williams and Isard 1997, Fitt 1997, Bladon et al. 1987). In contrast to these phonological approaches, Hertz and Huffman (1992) show how detailed information on the phonetic implementation in each dialect is required.

[11] Such sequences are indeed exhaustive as is ensured by the phonetic balance of the training data (Egan 1948).

As Labov and his colleagues in the Cross Dialectal Comprehension project have shown, segmental differences between American dialects can often result in serious misunderstandings (Labov 1989). Labov showed that Chicagoans, for example, had a marked local advantage in interpreting speech from Chicago, as compared to people from Birmingham and Philadelphia. Even when discourse context might have been thought to be a useful disambiguator, listeners often chose nonsense words to describe phonetic forms alien to their dialects (e.g. *budgeroom* over *bedroom* for [bʌdrum]).

Given that synthetic speech has been shown to be less understandable than natural speech (Pisoni 1997), it would seem that avoiding miscomprehension due to dialect differences is an important goal for synthetic speech. Therefore, if a synthesizer is to be used in a particular community, it may be a good idea to employ a synthesis of the local dialect, thus insuring maximum intelligibility.

Nearness to the dialect of the listener may also be thought likely to engender greater acceptability for synthetic speech. For example, Williams and Isard (1997) suggest that a Scottish bank telephone service with synthetic speech would be better off having a Scottish accent than an English one. However, all local dialects are not necessarily prized by their speakers. Labov has shown in New York (1966) and elsewhere that locals react negatively to phonological characteristics of their local dialects.

In section 6.6, we will discuss the success of our neural network based postlexical module at learning dialect features of a particular speaker whose voice is to be synthesized.

Provided that this dialect is not subject to negative connotations on the part of its speakers, it is believed that such a procedure will lead both to higher intelligibility, comprehensibility and acceptability of such a synthesizer.

2.4. Benefits of variability

Pisoni (1993, 1997) suggests that speech variability can actually enhance intelligibility. For example, Pisoni (1993) discusses results of training native Japanese speakers to distinguish English /r/ and /l/. When trained on variable stimuli, both across speakers and from different linguistic contexts within speakers, listeners improved in their perception of the /r/-/l/ distinction. Since, in the synthetic speech domain that we are considering, there is only one speaker, capturing the natural variability of segments in different environments may help improve intelligibility, though the effects of inter-speaker variability may be less applicable. One feature of the Motorola speech synthesizer's neural network system that gives it the chance of reproducing more natural speech variability than a typical concatenation system is the size of the learned unit. The neural network learns the most appropriate 5 or 10 millisecond frame for each environment. Since concatenation systems usually employ speech units of much greater duration, the neural network system has the opportunity to insert more natural variability.

Chapter 3. Data sources

In this chapter, we will discuss the two principle data sources of our study: a lexical database and a hand-labeled speech corpus of one individual. The characteristics of these data sources, and the modifications we have made to them, are important to understanding the major goals of this dissertation; namely to generate postlexical pronunciations typical of the labeled speech corpus on the basis of the lexical pronunciations found in the lexical database.

3.1. Lexical database

We developed a relational database architecture, implemented in Microsoft Access, to contain pronunciation information (cf. Grimes 1988). As shown in Figure 3-1, the database, known as Lexorola, contains four fundamental tables, Variants, Orthography, Pronunciation and Properties. The Orthography table contains the orthography of the word. Each orthography has several potential spelling variants, which are contained in the Variants table. For example, *color* may be spelled *colour*. Each orthography has one or more pronunciations, which are stored in the Pronunciation table, along with the source dictionary of the pronunciation. Each pronunciation has one or more sets of properties, which are stored in the Properties table. These properties include part of speech and sense.

Such properties are important in choosing the correct pronunciation for non-homophonous homographs, such as *record*. As a verb, it is pronounced /rəˈkɔrd/, while as a noun it is pronounced /ˈrɛkərd/.[12] An example where sense influences pronunciation is the word *bow*. When the word is used as a noun meaning 'forward part of a ship' it is pronounced /baʊ/; however, when it is used as a noun meaning 'archer's weapon', it is pronounced /bo/. Homographs which are homophonous, such as *bank* meaning 'riverside' or 'money business' are not annotated in our lexical database.

[12] We will use slashes (/) to enclose lexical pronunciations and square brackets ([ ]) to enclose postlexical pronunciations.

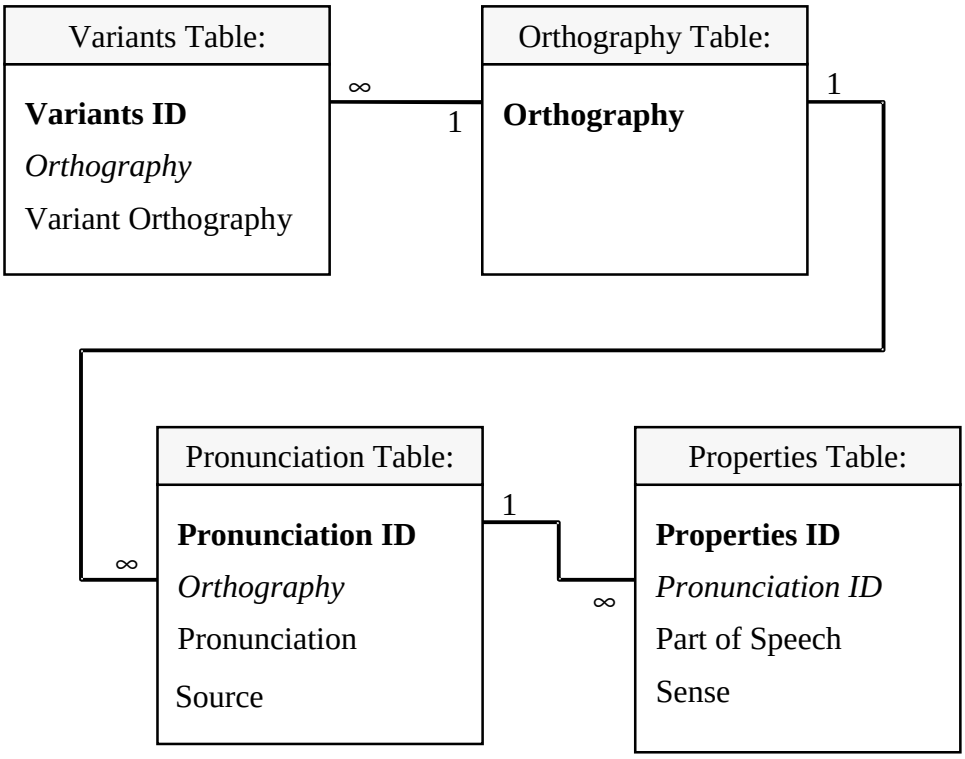

Figure 3-1: Relational lexical database architecture

Note: Primary keys in bold, foreign keys in italics. Relationships are shown by lines connecting associated keys in tables, with an indication of one-to-many directionality. Figure adapted from Karaali et al. (forthcoming).

The relational lexical database was created from three source lexica: The *Carnegie Mellon Pronouncing Dictionary* (Weide 1995), *Moby Pronunciator II* (Ward 1996) and *COMLEX English pronouncing lexicon* (also known as Pronlex, Linguistic Data Consortium 1995). The dictionaries we used tended to list variant pronunciations associated with particular orthographies with no annotation, regardless of whether the variants were sociolinguistic (*e.g.* /kɔt/ and /kɑt/ for caught), based on part-of-speech

distinctions (such as *live* /lɪv/ and *live* /laɪv/) or semantics (such as *lead* /lɛd/ and *lead* /lid/).

This kind of listing of variation is apparently suitable for speech recognition, where the object is to convert speech to text. In such an application, if the pronunciation /lid/ is recovered from the speech stream, it is sufficient to output the form *lead*; the fact that that orthography is also pronounced /lɛd/ does not matter. However, the existence of homophones like *eye* and *I* is important for speech recognition applications to be aware of and correctly disambiguate.

We removed sociolinguistic variants, such as dialect or style variants, in favor of one plausible American English dialect, from which various dialects and styles could be derived (cf. McLemore 1995), which we call "Generalized American". We distinguish this term from "General American" which has been used by researchers to describe a dialect that they believe to be actually spoken by a segment of the community (e.g. Wells 1982, 10).

For example, although it is a minority American pronunciation, we preserved the distinction between voiced /w/ and voiceless /ʍ/, as in 'which witch' /ʍɪtʃ wɪtʃ/. A dialect with a merger of these two phones can be derived from such a lexicon, while a dialect that preserves the distinction cannot be derived from a lexicon with a merger, at least without reference to orthographic information. Of course, a lexicon based on these

principles will end up reflecting no real dialect in particular. As Williams and Isard (1997, 2436) state in response to the idea of developing alternative accents for a synthesizer from a particular real base accent, “There seems to be no single accent containing all possible phonemes and distinctions of English accents”.

There have been several attempts at devising a transcription scheme for English that could accommodate some of the variation found in various dialects. Trager and Bloch (1941) describe a phonemic representation capable of describing a wide range of American dialects: “...it seems significant and convincing that the syllabic phonemes of all English dialects known to us...can be accommodated without forcing in the general system here set up” (Trager and Bloch 1941, 245). Labov (1991, 13) is a contemporary advocate of this system for similar reasons: “It is chosen as the representation that can best generate or relate current American dialects: a jumping-off point at which current sound changes begin”. Chomsky and Halle (1968, hereafter SPE) made use of underlying forms related to orthography, which reflect a much earlier stage of the language. They note that “there has...been little change in lexical representation since Middle English, and, consequently, we would expect (though we have not verified this in any detail) that lexical representation would differ very little from dialect to dialect in Modern English (SPE, 54). In a discussion of /u/ allophony in section 4.2, we will point out how contemporary phonological environments may provide better explanations than Middle English word classes.

Wells (1982) describes vocalic variations in English dialects worldwide using keywords to denote lexical sets containing large numbers of words whose vowels are the same. Williams and Isard (1997) employ Wells' keywords to facilitate developing speech synthesizers for new dialects. They find this approach easier than using phonemic transcriptions, since "it is not possible to choose one accent as a base form and systematically translate to others" (Williams and Isard 1997, 2435). According to Williams and Isard, if all vowels in the lexicon and the letter-to-sound conversion system use "keyvowels", the same lexicon can be used for all accents. While we find this approach interesting, we believe it is a bit too hopeful to expect not to have to make lexical adjustments between major dialect groups, such as British and American. For example, many words such as *corollary* and *ancillary* have initial stress in American and antepenultimate stress in British. In addition, there exist consonantal differences such as between /skɛdʒuəl/ (American) and /ʃɛdʒuəl/ (British) for *schedule,* and /fəlˈe/ (American) and /fˈɪlət/ (British) for *filet*.

We filled in part-of-speech and sense information for those words whose pronunciation is dependent on that information. In addition, frequency-annotated part-of-speech tags for the remaining words were added to the lexical information, to aid in the disambiguation process described in Karaali et al. (1998). At present, the lexical pronunciation database has almost 200,000 pronunciations, including over 1000 of which require disambiguation by part of speech, and over 200 of which require disambiguation by sense. Synthesizers have modules which can tag incoming words for part of speech (Church 1989) or

semantics (Yarowsky 1997) in order to help select the right pronunciations. Of course, this requires that lexica tag non-homophonous homographs for part of speech or semantics, if required. One of the purposes of combining the three source lexica was to identify and tag for part of speech or semantics as many non-homophonous homographs as possible.

Transcription protocols are being developed, based on McLemore (1995) and Schmidt et al. (1993), in order to make dictionary entries stemming from different source lexica consistent. Dictionary transcriptions are being made more phonemic as opposed to phonetic, in order to simplify letter-to-sound learning, and to allow for speaker-specific postlexical allophony and dialect characteristics to be determined in the postlexical module.

We refer to the level of transcription found in the dictionary as phonemic, or *lexical*. We oppose this to the *postlexical* level in accordance with lexical phonology (Kiparsky 1982, 1985). Although the phonemic status of schwa is not entirely clear (e.g. Bolinger 1981, Giegerich 1992, 68-69), we do transcribe vowels that are ordinarily reduced (for example the initial vowel in ***a****bout*) with schwa as a practical matter. Lexical pronunciations are characterized by their appropriateness for use in isolation. They show a lack of possible vowel reduction (for example in function words such as *can* or *had*), and they do not feature segmental reflexes of postlexical rules such as flapping. These are often suitable starting points from which *postlexical* variants due to rate of speech, style and context will be derived. For example, the /t/ in the lexical transcription for *thought*, /θɔt/, might

need to be modified to a flap [ɾ] in the expression "I thought about it" to reflect conversational American speech (see Nespor and Vogel 1986 for more examples).

Now the fact that /t/ has several allophones has some important consequences for our discussion. First, we believe that the traditional phonemic level is a useful intermediary point for both synthesis applications and human speech processing (contra e.g. Cohen 1995). We believe that the phonemic level is a useful one for pronunciation dictionaries and the target pronunciations of an automatic letter-to-sound processing system. That is, we do not think it would be useful to store postlexical pronunciations in a dictionary for speech synthesis, for example both /θɔt/ and /θɔɾ/, although this is an approach that has been considered for speech recognition (see section 1.2). Storing allophonic variants in the dictionary would require more space for the dictionary in addition to requiring a procedure for determining which variant to use in different contexts. We intend to perform the latter contextual determination in a manner to be described below that generates postlexical pronunciations from lexical pronunciations, so the postlexical variants do not need to be stored in the lexicon.

It is important to clarify the status of morphology in our lexical database. Allen et al. (1987), in an influential earlier approach, developed a morph dictionary, containing 10,000 morphs. These morphs were used in combination with a morphological parsing scheme to simulate the coverage a much larger dictionary. As Liberman and Church (1992) point out, the reduced cost of memory has allowed larger dictionaries to become

common for speech synthesis. Since the electronic dictionaries that we are using have inflections and derivations spelled out (perhaps another artifact of their design towards speech recognition), we chose to maintain this characteristic. We have subsequently noted the importance of building morphological structure into the dictionary, as well as developing morphological analysis software.

Morphological sophistication would enable us to ensure paradigm uniformity, for example, avoiding the hypothetical situation shown in Table 3-1, where some members of a paradigm are pronounced with one vowel and the rest with another. This kind of situation is a potential artifact of combining various dictionaries that filled out different paradigms to different extents based on of text corpora or other principles, resulting in inconsistent or incomplete inflections or derivations.

Another advantage to morphological sophistication would be to avoid sending orthographies that are not in the dictionary, but are inflections or derivations of words that are, to the letter-to-sound conversion module. Analyzing a complex word into a stem that might already be in the dictionary, and adding affixes by means of morphophonological rules, would generate a potentially more reliable pronunciation than one that would result from sending the whole word to the letter-to-sound conversion module.

Table 3-1: Hypothetical example of inconsistent pronunciation across morphological paradigms

Orthography	Pronunciation
walk	/wɔk/
walked	/wɑkt/
walks	/wɔks/

Finally, it is important to clarify that although our dictionary represents a state of phonology after the completion of lexical phonological rules, it does not represent postlexical pronunciation. Tajchman et al. (1995) refer to the forms found in dictionaries such as the one we are using as postlexical due to the fact that morphological alternations are spelled out; however, we prefer to reserve that term for the output of our postlexical module, which relies on the context of words in utterances.

3.1.1. Characteristics of source dictionaries

The lexical database retains information about which words came from which dictionary. This information will prove useful in an attempt to see which dictionary's transcriptions are most consistent. For example, the Pronlex database is thought to be of higher quality than the other sources, due to its manner of production and the transcription protocol to which it claims to adhere (McLemore 1995). In contrast, CMU (Weide 1995) notes that many of its pronunciations were in fact generated by a speech synthesizer, while Moby

(Ward 1996) appears to have pronunciations stemming from several different sources, and consequently appears to lack consistency.

3.1.2. Transcription consistency and simplification

One of the transcription characteristics of all of the dictionaries is an inconsistency in the use of the low back vowels, /ɔ/ and /ɑ/. This is, of course, not surprising, given the great variety of patterns in the distribution of these vowels across words and dialects (see Herold 1990, Labov 1996). One solution that we are contemplating is to enforce a distribution similar to that in British Received Pronunciation (RP): where orthographic 'o' would be pronounced /ɑ/, while other "complex" combinations of letters, such as 'augh', 'ough', 'aw', 'au' would be pronounced /ɔ/ (at least when the choice is between /ɔ/ and /ɑ/). Although there is probably no American dialect with just such a distribution, it has the advantage of being plausible and principled, and hopefully offending a minimum of people. Of course, it is also possible simply to eliminate distinctions such as this which are becoming more common among younger speakers (e.g. Hartman 1985). In any case, we will examine the extent to which the postlexical network "irons out" dialect asymmetries in the low back area between the lexical database and our recorded speaker in section 6.6. We examine the acoustic properties of /ɔ/ and /ɑ/ in our recorded database in section 4.2.

We will now discuss some potentially inconsistent dictionary transcriptions that we have rectified. We chose to simplify the transcription system used in the lexical database, that is, to make it more phonemic, in order to improve letter-to-sound conversion performance and to allow the postlexical network to handle allophony in a speaker-specific fashion. The apparent randomness in allophony in some of the source dictionaries used in our lexical database has been criticized, "ɨ/ə are among the most inconsistent transcriptions in any dictionary and there is almost no consensus for ɨ/ə transcription among different dictionaries" (Jiang et al. 1997). In addition, profusion of symbols in a pronunciation dictionary complicates the task of transcribers employed to augment the vocabulary.

Table 3-1 shows the phone sets used in Pronlex, CMU and Lexorola. From Pronlex to Lexorola, there are progressively fewer symbols used. For example, only Pronlex employs syllabic nasals, while CMU and Lexorola express these sounds with schwa + nasal consonant. Pronlex and CMU both employ a single symbol for stressed and unstressed schwa + /r/, while Lexorola uses two symbols. Pronlex uses /ɯ/, while Lexorola uses /ɦ/ + /w/. Pronlex uses /ə/ for the unstressed central vowel and /ʌ/ for the stressed central vowel, CMU uses /ʌ/ for both stressed and unstressed varieties, while Lexorola uses /ə/ for both, a move also proposed by Ladefoged (1982: 29-30).

Table 3-1: Phone sets for three pronunciation dictionaries

Pronlex	CMU	Lexorola	Pronlex	CMU	Lexorola
æ	æ	æ	dʒ	dʒ	dʒ
ʌ	ʌ		ʃ	ʃ	ʃ
ɛ	ɛ	ɛ	θ	θ	θ
ɪ	ɪ	ɪ	ʒ	ʒ	ʒ
m̩			b	b	b
ɔɪ	ɔɪ	ɔɪ	d	d	d
ə˞	ə˞		f	f	f
ʊ	ʊ	ʊ	g	g	g
aʊ	aʊ	aʊ	h	h	h
aɪ	aɪ	aɪ	k	k	k
ɑ	ɑ	ɑ	l	l	l
ɔ	ɔ	ɔ	m	m	m
eɪ	eɪ	eɪ	n	n	n
i	i	i	p	p	p
u	u	u	r	r	r
o	o	o	s	s	s
n̩			t	t	t
ə		ə	v	v	v
tʃ	tʃ	tʃ	w	w	w
ð	ð	ð	j	j	j
ŋ	ŋ	ŋ	z	z	z
ʍ					

Table 3-2 shows the varying treatment of stressed and unstressed vowels in the three dictionaries. In order to avoid problematic treatments of schwa (Jiang 1997, McLemore 1995), Lexorola has collapsed unstressed /ɪ/ with /ə/. McLemore (1995) proposes distinguishing between the these two unstressed vowels by employing /ɪ/ when an

unstressed vowel is on either side of a [+coronal] segment, or when it represents an *i* in the orthography. This rule overgenerates in the case of words like *attain* /ə.ˈteɪn/. Perhaps McLemore's rule could be improved by stating that /ə/ becomes /ɪ/ in the environment of a tautosyllabic coronal; however, we will suggest that leaving schwa-coloring unspecified in the lexicon, potentially in conjunction with a postlexical processor will generate more accurate speaker-specific transcriptions. Acoustic results on schwa realization will be discussed in section 4.2, while symbolic results on schwa realization will be discussed in section 6.4.

Table 3-2: Stress in three dictionaries

	Pronlex			CMU			Lexorola		
vowel	primary	secondary	unstressed	primary	secondary	unstressed	primary	secondary	unstressed
æ	√	√	√	√	√	√	√	√	√
ʌ	√	√		√	√	√			
ɛ	√	√	√	√	√	√	√	√	√
ɪ	√	√	√	√	√	√	√	√	
m̩	√								
ɔɪ	√	√		√	√	√	√	√	√
ɚ	√	√	√	√	√	√			
ʊ	√	√	√	√	√	√	√	√	√
aʊ	√	√		√	√	√	√	√	√
aɪ	√	√		√	√	√	√	√	√
ɑ	√	√	√	√	√	√	√	√	√
ɔ	√	√	√	√	√	√	√	√	√
e	√	√	√	√	√	√	√	√	√
i	√	√	√	√	√	√	√	√	√
o	√	√	√	√	√	√	√	√	√
u	√	√	√	√	√	√	√	√	√
n̩			√						
ə			√				√	√	√

The benefit to the schwa allophones in the dictionary is that the pronunciations can be often be used without postlexical modification. The difficulty is that enforcing such distinctions in the dictionary can be problematic, as McLemore (1995) observes. If a postlexical module can be relied on, as in the present case, perhaps the allophonic variation in the dictionary can be dispensed with.

3.2. Labeled speech corpus

We will now describe the speech database used for training the postlexical neural network, in addition to the duration and acoustic neural networks described in section 1.2. The database consists of both sentence-length materials and single-word utterances. These materials were distributed as shown in Table 3-1, and consist in total of 7088 words. The table shows the total quantity of utterances of each type and shows the split between training and testing sentences for the postlexical network. Postlexical training was performed on a randomly selected 90% of the total materials, while all test results are given for the remaining 10% which were withheld from training.

Table 3-1: Materials in the labeled speech corpus

Sentence type	Total quantity	Train quantity	Test quantity	Mean word length	Source
"Harvard"	453	401	52	7.97	Egan (1944)
semantically anomalous	142	125	17	11.19	Corrigan (1996)
questions	45	43	2	10.18	
news/ manuals	20	19	1	8.09	
words	1066	958	108	1	

A university-educated male speaker who grew up in Chicago was recorded with a close-talking microphone in a soundproof room reading the speech database when he was 36 and 38 years old. The talker was asked to read the materials in a normal voice. The

talker also recorded a similar set of materials with the instructions that he speak “clearly”, but these were eliminated from the current study as well as for acoustic model training due to their unnaturalness.

The speech was labeled at a fairly narrow phonetic level in a manner similar to the TIMIT database (Seneff and Zue 1988) by an electrical engineer with some graduate training in linguistics. Although the database was originally labeled in accordance with TIMIT, some subsequent modifications to the phone set were made in accordance with the CSLU (Center for the Spoken Language Understanding) labeling guidelines (Lander 1997): both aspirated and unaspirated voiceless stops were used, and a distinct symbol was introduced for glottal stops derived from /t/ as opposed to those that emerge intervocalically. Both of these changes were introduced by the labeler by rule before postlexical training had begun. In subsequent databases, aspiration distinctions have been entered by labelers by hand, on an auditory or signal display basis.

In Table 3-2 we have listed the phones used in the labeled corpus. We have translated the ASCII Worldbet symbols (Hieronymus 1993) used in the label files into a modified version of the IPA. IPA modifications include the labeling of stop closures with a ‘c’ suffix, and the labeling of the glottal stop derived from /t/ as /tʔ/. Stop releases are indicated without the ‘c’ suffix, and aspiration is noted by a superscript ‘h’. Stop closures and releases are labeled separately in some databases used for speech technology, due to the important acoustic distinctions between the two events.

Table 3-2: Postlexical phones used in labeled corpus

ɑ	æ	ʌ	ɔ	aʊ	ə	ə̥	ɚ	aɪ	b
bc	tʃ	d	dc	ð	ɾ	ɛ	l̩	m̩	n̩
ŋ	ɝ	eɪ	f	g	gc	h	ɦ	ɪ	ɨ
i	dʒ	dʒc	k	k^h	kc	l	m	n	ŋ
ɾ̃	o	ɔɪ	p	p^h	pc	ʔ	r	s	ʃ
t	t^h	tc	θ	ʊ	u	ʉ	v	w	j
z	ʒ	tʔ							

It is important to emphasize that the recorded database is read speech. Labov (1994, 157-158) reiterates his position that attention paid to speech is the central organizing principle of stylistic variation. Following Labov, spontaneous speech covers a spectrum ranging from casual speech, which is closest to the vernacular, through careful speech, such as in a formal interview. These are distinguished from controlled styles, such as reading, which are at the formal end of the stylistic continuum. According to Labov, "only in spontaneous speech will we find the most advanced tokens of linguistic change in progress" (1994, 158).

Blaauw (1994) discusses some of the prosodic aspects of read speech that allow listeners to discern it reliably from spontaneous speech. Read speech is probably closer to clear speech (Picheny 1985, 1986) than spontaneous speech, and is likely to feature fewer postlexical reductions (Faust 1995, Van Bergem 1993, Koopmans-Van Beinum 1992, Eskénazi 1992). Nevertheless, as will be seen, our read database provides a wide range of postlexical phenomena. Of course, it will be interesting to work on spontaneous

speech in future work, as it will undoubtedly widen the range of phenomena encountered. It is also possible that an acoustic model for a speech synthesizer will be able to be more natural if it is trained on spontaneous speech. According to Laan (1997, 44):

> Although most text-to-speech systems are fairly well intelligible nowadays, they often employ a somewhat "dry" and "mechanical" reading style that requires much concentration of the listener. Most of our knowledge on speech perception and speech production and thus speech synthesis too, is based upon research on read speech, often even with nonsense words. Therefore, studies on different types of speaking styles are requested to complement our basic knowledge about the process of speech production and perception.

We hope to take up this challenge and to test the hypothesis that synthesis based on spontaneous speech requires less listener concentration, when we analyze a new speaker currently being labeled, who was recorded with both read and spontaneous speech.

The manner in which the speech database that we are using was labeled will have an effect on the way in which we view postlexical variation. Lander (1997) points out how both the TIMIT and the more recent CSLU labeling systems for English are largely phonemic, but include several allophones, which are deemed "spectrally distinct" and "frequently occurring". For example, nasalization is not indicated on vowels, but flaps, fronted /u/ and two reduced vowel qualities (ɨ and ə) can be indicated. Seneff and Zue (1988, 4) explain why additional allophonic variants, such as nasalized or lateralized vowels, are not employed: "Such information would surely be useful, but the decision-making process is prone to judgment error, and would require a significant increase in time and effort (for labeling)."

Despite their adequacy for carefully articulated or read speech, Greenberg et al. (1996) discusses the failure of phone-based models to capture "many of the spectro-temporal properties of spontaneous speech typical of informal spoken dialog". However, we feel, following Keating et al. (1994, 135) with respect to the TIMIT database, that a conceptualization involving phonetic segments is useful:

> The TIMIT labels represent a necessarily-arbitrary segmentation and categorization of utterances into phonetic segments. Obviously, any study of these labels will be acceptable only to researchers who accept segmental phonetic transcriptions as a useful record of speech-events. It is possible to admit that segmental transcriptions have no theoretical basis as formal representations of speech, and yet still find them useful: they are a shorthand for, a pointer to, key articulatory and acoustic events.

Laporte (1997, 411-412) also acknowledges the flaws of string-based phonetic transcription, however he notes that they are convenient and useful and "standardized to quite a reasonable degree among linguists". In addition, string-based analysis lends itself to the machinery of regular relations (Kaplan and Kay 1994).

3.2.1. Syntactic and prosodic labeling

Various syntactic and prosodic information was provided in the speech labeling scheme as well. Figure 3-1 illustrates the speech labeling scheme with all of the tiers that were labeled on the present speech corpus. We have discussed the phonetic tier above, and will describe the rest of the tiers below.

Syllable boundaries were provided on perceptual grounds where possible. According to the labeler, in cases where the speech signal did not permit an easy syllable division, *The*

New Merriam-Webster Dictionary (Merriam-Webster 1989) was consulted. This dictionary's syllabification was based on the practices of *Webster's Ninth New Collegiate Dictionary* (Merriam-Webster 1984, David Justice, personal communication). Interestingly, Merriam's syllabification policy changed between the ninth and tenth (Merriam-Webster 1996) editions of the *Collegiate* (Brian Sietsema, personal communication). The ninth edition uses a stress-conditioned resyllabification analysis (see Blevins 1995, 232), while the Tenth Collegiate appears to adhere more strictly to the Maximal Onset Principle (see Blevins 1995, 230).[13]

It should be noted that there is no labeling of resyllabification across words in the corpus. Nespor and Vogel (1986, 64) claim that resyllabification does not occur across words in "normal colloquial" speech; however they do not exclude the possibility that it does occur in "fast or sloppy" speech. Contrary to what might have been expected, Labov (1997) reported that resyllabification across words in a vast vernacular database occurs with low frequency. Given these results, and the fact that our database is at the formal end of the stylistic continuum, the absence of labeling of resyllabification may not be critical.

[13] Examples (using Merriam's transcription symbols) from the ninth *Collegiate* (Merriam-Webster 1984) show that stressed short vowels often have codas: *lemon* ˈlem-ən, *dirty* ˈdərt-ē, *litter* ˈlit-ər, cf. *healer* ˈhē-lər. The same words from tenth *Collegiate* (Merriam-Webster 1996) show strict adherence to the maximal onset principle regardless of the length of the preceding vowel: ˈle-mən, ˈdər-tē, ˈli-tər, ˈhē-lər.

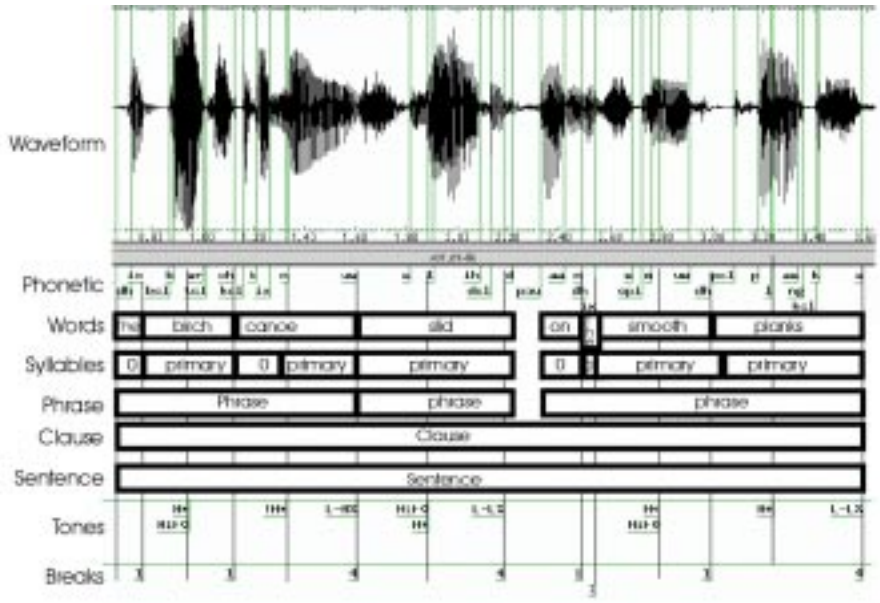

Figure 3-1: Speech labeling scheme

Source: Karaali et al. (forthcoming)

There were two separate passes for marking word-level groupings and above. The first, syntax-oriented, phase involved assigning syntactic phrase, clause and sentence boundaries, function/content tags and prominence rankings (O'Shaughnessy 1976) to words. The second, prosody-oriented, phase involved labeling the database according to the ToBI conventions (Beckman and Elam 1997).

Phrase and clause boundaries were added according to traditional notions. Sentence boundaries were added at the beginning and end of each utterance; which coincided with textual sentences denoted by periods. Words were assigned the labels *content* and *function,* more or less corresponding to open and closed class words (see Haegeman

1991, 105). Each word was also assigned a prominence ranking, according to Table 3-1. These rankings are based on O'Shaughnessy (1976) who found that part of speech of words affected fundamental frequency in approximately the rank ordering shown in the table. Similar rankings were used by Allen et al. (1987).

Table 3-1: Prominence rankings

Ranking	Word types
0	articles
1	conjunctions, relative pronouns
2	prepositions, auxiliaries, nonemphatic modals, vocatives
3	personal pronouns
6	finite verbs, demonstrative pronouns
7	nouns, adjectives, ordinary adverbs, negative contractions
8	reflexive pronouns
9	emphatic modals (e.g. "I *do* like ice cream")
10	quantifiers
11	interrogative words
12	negative adverbs
13	sentential adverbs

Note: Based on O'Shaughnessy (1976) and Allen et al. (1987).

The ToBI labeling includes tiers with tones, breaks and words. The words tier simply contains the orthography for each word. The words tier was useful for obtaining the lexical pronunciation of each word from the lexical database, which was aligned with the postlexical pronunciations (see section 5.1). The breaks tier contains a number between 1 and 4 (including diacritics which were ignored in the postlexical neural network encoding discussed in section 5.4), indicating the relative disjuncture between each pair of words. '4', indicating the greatest possible disjuncture, signals the end of an intonational phrase,

while ‘3’ signals an intermediate phrase (Pierrehumbert and Beckman 1988). ‘1’ signals the normal disjuncture between words, while ‘2’ is either a strong ‘1’ or a weak ‘2’. As will be seen in section 5.4, break values of ‘3’ and ‘4’ were used in the neural network encoding to signal the ends of intermediate and intonational phrases, respectively, while the other breaks were ignored.

The tones tier contains pitch accents, which fall on accented syllables, phrase accents, which fall at the ends of intermediate and intonational phrases, and boundary tones, which fall at the ends of intonational phrases. The tones tier was used in the neural network encoding to mark whether or not a syllable was accented. The type of accent (e.g. H*, L*+H) was not provided to the postlexical neural network, though it was provided to the duration and phonetic implementation networks. While some research has shown the important effects of presence of accent on segmental allophony (e.g. Dilley et al. 1996), it is not clear whether *type* of accent should have a particular effect.

3.2.2. Levels of labeling

Barry and Fourcin (1992) present a useful typology of speech database labeling levels, described in Table 3-1, within which it will be useful to situate our labeling procedure. Our labeling system, as well as the TIMIT and CSLU labeling systems, shares characteristics of Barry and Fourcin’s acoustic-phonetic, narrow-phonetic and broad-phonetic levels. For example, our use of separate symbols for stop closures and releases fits in with the acoustic-phonetic level. Our use of some allophonic labels such as those shown in Table 3-2, coincides with the narrow-phonetic level. An additional

source of allophony in the speech corpus not shown in the table is the laryngealized vowels, marked in the corpus by a preceding glottal stop [ʔ]. As has been stated above, the use of allophonic labels is somewhat arbitrary and incomplete, so our system also corresponds more to the broad-phonetic level for those phones lacking allophonic symbols in Table 3-2. CSLU's (Lander 1997) labeling system actually permits stricter adherence to the narrow-phonetic level by its use of various diacritics to signal nasality, rounding, etc.

Table 3-1: Barry and Fourcin's (1992) Labeling Typology

Level	Characteristics
Physical	multitiered; can include data from speech analysis machinery, such as nasal transmission detectors or electropalatographs; can be linked to descriptions such as Browman and Goldstein (1986)
Acoustic-phonetic	uses established phonetic descriptors such as stop closure, release burst, aspiration; primary criterion for assigning a label to a stretch of speech is acoustic homogeneity; can retain base symbol link so that narrow and broad phonetic levels can be derived (semi-) automatically
Narrow-phonetic	uses IPA or ASCII equivalent; difficult since many processes not categorical; differentiates as much as possible with respect to properties modifying base sound (e.g. rounding, nasality),
Broad-phonetic	only employs symbols that have phonemic status, but uses them to indicate non-phonemic, continuous speech phenomena, e.g. *good boy* [gʊb bɔɪ].
Phonemic	represents functionally distinctive sound units of a language; serves as mediator between sound signal and lexicon; decision on phonemic form of words cannot always be automatic, cf. either /iðə˞/, /aɪðə˞/.

Table 3-2: Allophones in TIMIT labeling system

Phoneme	Allophones	Phoneme	Allophones
/u/	[u], [ʉ]	/d/	[d], [ɾ]
/ə/	[ə], [ə̹], [ɨ]	/l/	[l], [l̩]
/t/	[t], [tˀ], [ɾ], [ʔ]	/m/	[m], [m̩]
/n/	[n], [ɾ̃], [n̩]	/h/	[h], [ɦ]

In our investigation of the linguistic representations useful at various stages in text-to-speech synthesis, we will pay particular attention to the interface between phonemic and allophonic representations.

Chapter 4. Experimental approach to comparing gradient vs. discrete aspects of postlexical variation

A potential benefit to labeling allophonic variation in the speech corpus is that better models for particular acoustic events can be learned by the Motorola synthesizer's acoustic neural network described in section 1.2. For example, if the two /u/ allophones, [u] and [ʉ], are used as labels, it is likely that the vowels corresponding to each label will be more tightly clustered in formant space than if only one phonemic /u/ label were employed. Provided one knows when to request each variant of /u/, the more appropriate allophone can be used. In addition, it may be suspected that training two separate allophones will allow /u/'s to cover more of the vowel space than they might if only one /u/ model were trained. If there were only one /u/ model, it is likely that it might reflect an averaging of fronted and unfronted /u/'s, therefore representing a compression of the vowel space for /u/. Given Bradlow et al.'s (1996) finding that speakers with larger vowel spaces are more intelligible than those with smaller vowel spaces, training an acoustic module to have two /u/ models might well aid intelligibility.

In this section, we describe spectrographic analyses performed to examine the clustering of vowel allophones in an effort to assess the contribution of the allophonic labels. We describe an experiment in which we trained the acoustic neural network differentially using phonemic and allophonic labels, in an effort to see how well the network can

capture allophony in acoustic space without the aid of allophonic labels. Based on the results of these experiments, we will explain our position on the use of allophonic labels.

It is a truism in phonetics that no two utterances, even of the same word, are identical. Nevertheless, linguists have sought to categorize sounds that have particular commonality. For example, phonemes can be posited on perceptual grounds via minimal pair tests. This research aims to shed light on the question of whether allophones of phonemes are in fact infinite or profitably limited to a small number, perhaps depending on the phoneme involved. As an example of the dual contribution of these possible extremes to the grammar, Labov (1994, chapter 7) discusses how both a gradient approach to phonetic implementation and a discrete approach to the phonological description of chain shifts results in the best description of these phenomena.

In addition to providing empirical evidence for characterizing the nature of phonetic processes, this research can be of practical benefit to speech technology applications, where such results can influence the modeling of postlexical processes as well as the representation used in pronunciation dictionaries and speech databases. These experiments will also provide acoustic phonetic evidence that bears on the debate between Liberman and Pierrehumbert (1984) and Kiparsky (1985), among others, regarding the categorical, gradient or mixed nature of postlexical processes

We employ a speech synthesis system featuring a neural network for generating acoustic parameters for 5 millisecond frames of speech on the basis of current and surrounding linguistic information (Karaali et al. 1997) as a laboratory in which to assess whether particular postlexical processes are best treated as gradient phonetic implementation rules or discrete, symbolic phonological processes. We envisage the four possible conditions with respect to the symbolic representation of allophony in our speech synthesis system shown in Table 4-1. We will elaborate on each of these conditions below. We believe that Conditions 1 and 2 are the most viable, and those will be the ones distinguished in the experiments below.

Table 4-1: Phonemicity and allophony in speech synthesis data sources

Data source	Condition 1	Condition 2	Condition 3	Condition 4
dictionary	phonemic	phonemic	allophonic	allophonic
labeled corpus	phonemic	allophonic	allophonic	phonemic

Condition 1 is characterized by the use of phonemic labels in both the dictionary and the labeled corpus. In this case, all allophony will need to be handled by the acoustic neural network. As Barry and Fourcin (1992, 10) indicate, a phonemic labeling system allows transparent lexical access between the dictionary and the acoustic models generated from the corpus. Our experiments below will shed light on the extent to which relying on the acoustic module to handle all allophony is sufficient.

Condition 2 combines a phonemic dictionary with an allophonically labeled corpus. In this condition, a postlexical module to convert between phonemic lexical symbols and the allophonic symbols of the corpus and acoustic module is required. This condition is the one currently being pursued most intensely. We believe that keeping the dictionary phonemic allows it to be reused for a variety of voices. In addition, the postlexical and acoustic modules can be easily retrained for each speaker, in order to learn their individual postlexical phonology.

Condition 3 combines an allophonic dictionary with an allophonically labeled corpus. While lexical access might appear to be transparent, as in Condition 1, this approach is problematic. For example, it would be difficult to design an allophonic dictionary that would handle cross-word phenomena with ease. Due to the speaker-specific nature of postlexical phonology, the lexicon might need to be revised for each speaker, a potentially labor-intensive process. It might be suggested that a dictionary with a wide range of allophonic variants would be useful; however, a principled means by which to select the appropriate variant that takes into account a constellation of linguistic and paralinguistic information would need to be developed. Cohen (1989, chapter 5) discusses the notion of allophonic cooccurrence, whereby the association between such constellations of information and particular linguistic forms can be modeled in a speech recognition system.

Finally, Condition 4 uses an allophonic dictionary with a phonemically labeled corpus. The problems with an allophonic dictionary discussed for Condition 3 still apply. In

addition, there would be no meaningful acoustic models for the various allophones in the dictionary to be associated with in the acoustic module. Given this situation, we feel that this condition would be unsatisfactory.

It is assumed that the acoustic neural network is affected by these labeling choices. We will analyze the clustering of the allophones in formant space in both Conditions 1 and 2 in order to assess the relative contribution of the various labeling schemes. This experiment will be informed by an analysis of the original speech.

There are several different perspectives from which to consider postlexical variation. One way is to look at the structural symbolic, or transcription, differences between lexical and postlexical representations. For example, certain kinds of deletion can be viewed as the deletion of a symbol, while flapping could be viewed as the substitution of one symbol for another. Of course, such a symbolic approach might be faulted for ignoring the finer-grained articulatory and acoustic changes that are taking place. Some postlexical variation might best be considered from an acoustic or articulatory perspective. For example, the study of vowel allophones' formant measurements might shed light on their patterning.

In the case of the subphonemic vowel symbols, we intend to compare the acoustic neural network's performance with and without such symbols, in an effort to see how much the network's learning of postlexical variation is assisted by explicit labeling. We will attempt to determine the relative value of achieving postlexical variation by symbolic

means versus acoustic, phonetic means. For example, we will examine the formant values of fronted /u/ when given a special label [ʉ], versus letting the acoustic neural network determine phonetic allophony based on a single *phonemic* label /u/.

We propose to examine the first and second formant values of certain pairs of vocalic allophones produced by the speech synthesizer under three conditions. The first "original" or "natural" condition is the original natural speech. The second "normal neural network" or "allophonic label" condition involves training the acoustic neural network on the original TIMIT allophonic labels in our speech database. The third "collapsed neural network" or "phonemic label" condition will involve training the acoustic neural network on a version of the speech database where allophonic labels have been replaced by a "collapsing" label. This collapsing label might be the underlying phoneme, as in Table 3-2, or an arbitrary choice between the allophones to be collapsed. Note that in the collapsed condition, the features of the collapsing label are applied to both allophones.

We will examine the clustering in formant space of allophones of the same phoneme under the three conditions. If the allophonic label condition is more similar to the natural distribution than the phonemic label condition, the use of a 'phonological' label will be seen to contribute positively to the description of events. However, if the phonemic label condition results in a distribution at least as similar to the natural condition as the

allophonic label condition, we can assume that this allophony is perhaps of a gradient type; that is, not necessary to describe in symbolic terms.

4.1. Acoustic neural network[14]

The role of the acoustic neural network (also known as the phonetic implementation network) is to convert the linguistic representation of speech into an acoustic representation of speech. Specifically, the output of the phonetic network is a vector of vocoder parameters for a five-millisecond frame of speech which can be converted into a speech waveform by using the synthesis portion of a vocoder (Karaali et al. 1997). The input to the acoustic neural network contains phone information, stress information, and syntactic information (it is also possible to include prosodic information, but the network described here was intentionally reduced in size and complexity in order to make training faster) for each frame of speech.

In order to convert the linguistic input into the acoustic output, a multi-layer neural network is used, as shown in Figure 4-1. In order to keep the neural network architectural descriptions manageable, we have adopted a block notation that allows nodes to be grouped into a more abstract level of representation (Karaali 1989). The notation incorporates two fundamental symbols. Rectangles represent a single layer of neural network nodes and encapsulate weight, bias and activation characteristics. All other components in our architectures, including blocks that facilitate input and output, are

[14] This section was prepared with the help of Noel Massey.

denoted with hexagons. Labeled arcs are used to describe the dimensionality of input and output. Blocks that receive more than one arc indicate input vectors that are concatenated from multiple sources. For reference, each block in each diagram is also indexed with a number surrounded by square brackets. This notation allows neural network architectures to be represented in a compact block format.

In order to enhance the neural network's concept of phones, the phone information is supplemented with vectors of features, similar to those shown in Table 5-1. This somewhat redundant information improves the network's ability to learn similarities between phones. The phone to feature blocks (blocks 3 and 4) convert the phones into sets of features. Note that in the collapsed conditions, the features take on values corresponding to the collapsing label.

Blocks 1 and 2 in Figure 4-1 represent the other linguistic inputs to the neural network. Block 1 utilizes a time-delay neural network (TDNN) window. The TDNN window provides a running window over 30 non-uniformly sampled frames over 415 milliseconds. More samples are taken towards the middle of the window than at the ends. This method provides a reasonable context over which to determine the appropriate acoustic parameters. Block 1 has the following inputs: prominence ranking (discussed in section 3.2), syllable stress (primary, secondary or unstressed), function/content specification, and the phone label. It is this phone label that is either modified or not in the different experimental conditions to be described.

Input block 2 provides information about the duration of the preceding and following four segments, as well as their distance to the current frame. This block has the possibility of providing information more distant than that provided in the TDNN window of Block 1. In addition, information is provided about the distance to preceding and following syntactic boundaries.

Output block 12 contains the coder parameters for each frame of speech. In the training situation, these are presented as inputs at the same time as the inputs to blocks 1 and 2, and the neural networks weights in blocks 5-11 are adjusted to minimize error. In the testing, or runtime, situation, when input is fed into blocks 1 and 2, output is emitted from block 12 which is then passed to a coder in order to be synthesized into speech.

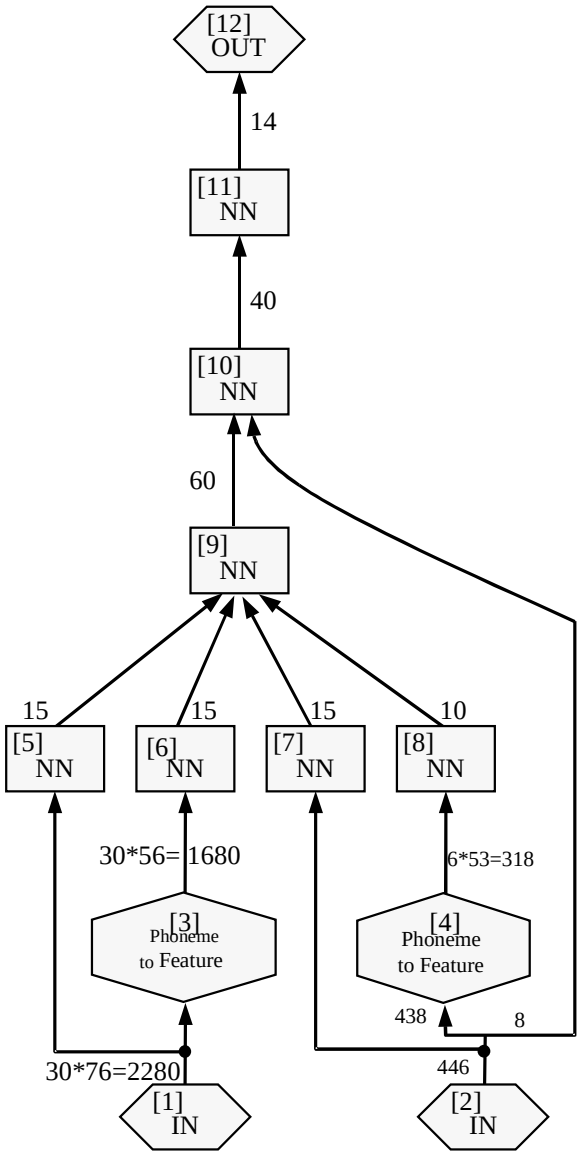

Figure 4-1: Acoustic Neural Network

Source: Adapted from Karaali et al. (forthcoming).

4.2. Experimental procedure

We examined first and second formant values of four pairs of vowels in three separate conditions. Since this experiment was designed to investigate the contribution of allophonic labels to phonetic implementation, we will refer to the pairs of vowels in each experiment as "allophones", even if they are not strictly allophones in a theoretical sense. The first condition, "original" (speech), was a testing subset (97 out of 470[15] Harvard sentences) of the original recorded speech in our database, downsampled to 8000 samples per second. The second condition, "normal" (acoustic neural network), involved training our acoustic neural network on the training data using original allophone labels and testing it on the testing subset. The third condition, "collapsed" (acoustic neural network), was the same as the second condition, except that the phone labels and features of allophones of the phone under study were collapsed to those of a single phonemic label, thereby preventing the network from benefiting from whatever contribution separate allophonic symbols might provide. Table 4-1 shows the quantities of each of the allophones examined in this set of experiments. These are approximately 20% of the entire data, which were withheld from acoustic neural network training.

[15] This number is larger than the number of Harvard sentences used in postlexical training, discussed in section 3.2, due to the fact that several files need to be eliminated from postlexical training due to various problems aligning lexical and postlexical forms (see section 5.1) in those files.

Table 4-1: Allophone quantities in testing subset

Collapsing Phone	"Allophone"	Quantity
/ə/	[ɨ]	123
	[ə]	84
/u/	[u]	11
	[ʉ]	26
/ɑ/	/ɑ/	50
	/ɔ/	43
/o/	/o/	38
	/i/	72

In one set of experiments, we examined the formant values of [ə] and [ɨ] in the first two conditions and collapsed them both with the /ə/ label in the third condition. In a second set of experiments, we examined the formant values of [u] and [ʉ] in the first two conditions and collapsed them both with the /u/ label in the third condition. Besides altering the phone label input, the third condition also provided the features of the replacement phone during network training.

After the initial sets of experiments were conducted, we chose to run two additional sets in order to gain more insight into the nature and value of this method for assessing the extent of symbolic mediation in phonetic implementation. We chose to consider /ɑ/ and /ɔ/ as allophones for the purpose of this experiment, and to collapse them as /ɑ/ in the collapsed condition. We thought that since they really are phonemes, with overlapping distributions, that chaos should result in the collapsed condition. Contrary to our

expectations, as we discuss below, the collapsed condition maintained a significant distinction between /ɑ/ and /ɔ/, indicating that in our speaker's dialect the environments for /ɑ/ and /ɔ/ may be somewhat complementary.

Finally, we chose to identify a pair of vowels whose distribution was not only patently not complementary, but whose realizations in formant space were substantially distinct. For this experiment, we used /o/ and /i/, and collapsed them as /o/. As we will discuss below, chaos does indeed result in the collapsed condition. If nothing else, this lends some credence to the results in the more meaningful experiments where the collapsed condition does not result in chaos.

Below we describe the results for both sets of experiments. To assess the effects of the different training situations, we use formant values predicted by the Entropic ESPS version 5.2 *formant* program. F1 and F2 measurements were taken at the vowel midpoints for each occurrence of the vowels under study. A window duration of .1 seconds and a frame step of .005 seconds (in order to be synchronized with the frame size of the voice coder) were used. In the figures below, formant values are all placed on grids with same dimensions, representing the entire vowel space of our speaker, in order to highlight their relative location in his vowel space.

In order to provide further graphical indications of each of the vowels' extent in space, we wrote a program to draw ellipses around the vowels according to the following: the ellipse midpoint is the intersection of the F1 and F2 means for the vowel in each

condition. The major and minor axes of the ellipse are two standard deviations along F1 or F2. The degree of rotation of the ellipse is derived from the standard rotation of axes where the angle θ is determined from Equation 4-1. The degree of rotation therefore captures the covariance of F1 and F2.

Equation 4-1: Rotation of axes

$$\tan 2\theta = \text{covariance}(F1, F2) / (\text{variance}(F1) - \text{variance}(F2))$$

We employed two different statistical tests to assess whether the means of two samples were actually sampled from the same population—the null hypothesis. In our case, the null hypothesis applied to formant measurements from two allophones of the same phoneme would mean that the allophones are not significantly distinct in formant space. Using the tests we selected, we take probabilities of less than .05 to indicate that the null hypothesis is rejected, and that the allophones are distinct in formant space.

In order to examine both F1 and F2 simultaneously, we required a multivariate statistic. We employed a two-dimensional Kolmogorov-Smirnov (K-S) test (Press et al. 1992) to determine whether two samples described by two variables are significantly different. The K-S test returns a value D, whose probability determines the acceptance of the null hypothesis. We also ran a univariate statistic, the t-test (using Matlab), on each formant in each condition to see which formant(s) contributed to differences between allophones.

In the terms of this experiment, rejection of the null hypothesis on either formant for a pair of vowels indicates that the vowels are not sampled from the same population. Of course, this does not mean that they could not be of the same *phoneme*—as is well known, some allophones of the same phoneme are acoustically very different, such as [ɾ], [ʔ] and [t] as allophones of /t/. It merely indicates whether they are a significantly distinct subgroup of that phoneme. If an allophone is distinct from another allophone of the same phoneme in the original speech, the acoustic neural network's goal can be seen to be to preserve such a distinction in its output. The purpose of this experiment is to see to what extent the phone category label of the specific allophone has an effect on this result.

As another method of assessing the differences between the normal and collapsed conditions with respect to the original speech, we chose to examine Euclidean distance between F1 and F2 values in the different conditions. Euclidean distance, as shown in Equation 4-2, is a method for assessing the difference between two vectors, *u* and *x*, containing *j* elements.

Equation 4-2: Euclidean distance

$$d = (\sum_{j} | u_j - x_j |^2)^{1/2}$$

In our case, the values of F1 and F2 midpoints for a given vowel in a given condition constitute such a vector. For each set of F1 and F2 values for a given vowel, we will compare the Euclidean distance between the original condition and both the normal and collapsed conditions. We assume that larger Euclidean distances in formant space will be associated with increased distortion from the original speech. Note that Euclidean distances can only be properly compared when the number of elements in a vector is the same, and the quantities are comparable. This is the case when comparing the difference between the collapsed and normal conditions with respect to the original speech for each formant separately.

4.2.1. Experiment on [ə]/[ɨ]

First we will report on the experiment considering [ə]/[ɨ] allophony. All plots and numerical results are based on training the acoustic neural network for 35 epochs. Table 4-1 shows the results of the hypothesis tests for this experiment in each condition. As can be seen from the significant values for the K-S statistic in each condition, we can reject the null hypothesis that the formant values for [ə] and [ɨ] come from the same distribution.

Therefore, we can say that [ə] and [ɨ] are distinct in the original speech, and that distinction is preserved by the "normal" acoustic neural network with allophonic labels, as well as the "collapsed" acoustic neural network with the single phonemic label for both

allophones. It is interesting to note that the significance of the distinction goes down from the original to the normal to the collapsed condition, indicating that the distinction is less as one proceeds through the conditions.

As can be seen from the t-test results in Table 4-1, F2 is significantly distinct for each allophone in all three conditions. In contrast, F1 is significantly distinct between the two allophones in all but the collapsed condition. This result indicates that the F2 (front-back) difference is a more robust indicator of the difference between [ə] and [ɨ] than F1. Not unexpectedly, the acoustic neural network, whether in the normal or collapsed condition, tended to blur the less robust distinction in the F1 (high-low) domain.

Table 4-1: Hypothesis tests for [ə]/[ɨ] experiment

condition	K-S	F1 t	F2 t
original	9.49E-20	1.84E-12	0
normal	1.61E-15	4.58E-06	0
collapse	6.10E-07	0.09	2.21E-12

Note: Insignificant results are shaded gray.

Table 4-2 shows the Euclidean distances between each of the formants in the normal and collapsed condition from the original condition. As might be expected, the collapsed condition was more distant from the original condition than the normal condition for the F2 of [ɨ] and both F1 and F2 of [ə].

Table 4-2: Euclidean distances between original speech and normal and collapsed conditions

vowel	formant	normal	collapse
ɨ	F1	727	699
	F2	1524	1633
ə	F1	947	1097
	F2	1287	1417

Figure 4-1 shows the F1/F2 distributions of [ə] and [ɨ]. Note that the legend on this and subsequent charts uses the TIMIT symbols ([ə] = 'ax' and [ɨ] = 'ix'). As can be seen in the figure, the distributions of the two vowels overlap, but [ɨ] is fronter and somewhat higher than [ə].

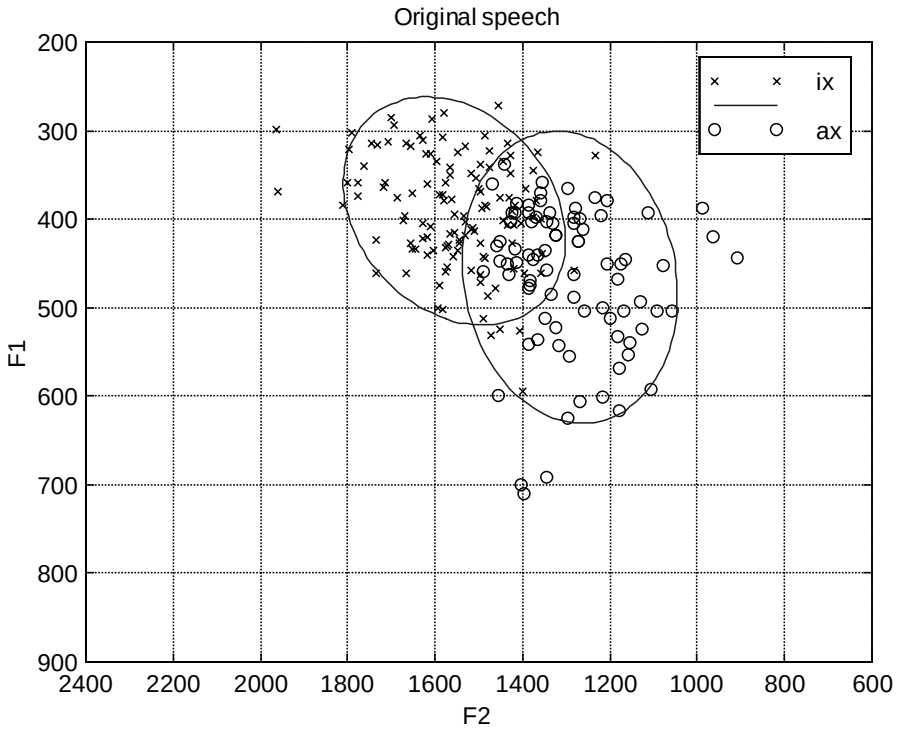

Figure 4-1: Formant distribution of [ə] and [ɨ] in original speech

Note: Legend uses TIMIT labels.

Figure 4-2 shows the distribution of [ə] and [ɨ] in the normal neural network with allophonic labels for these phones preserved. In this case there appears to be more overlap between [ə] and [ɨ] than was seen in the original speech.

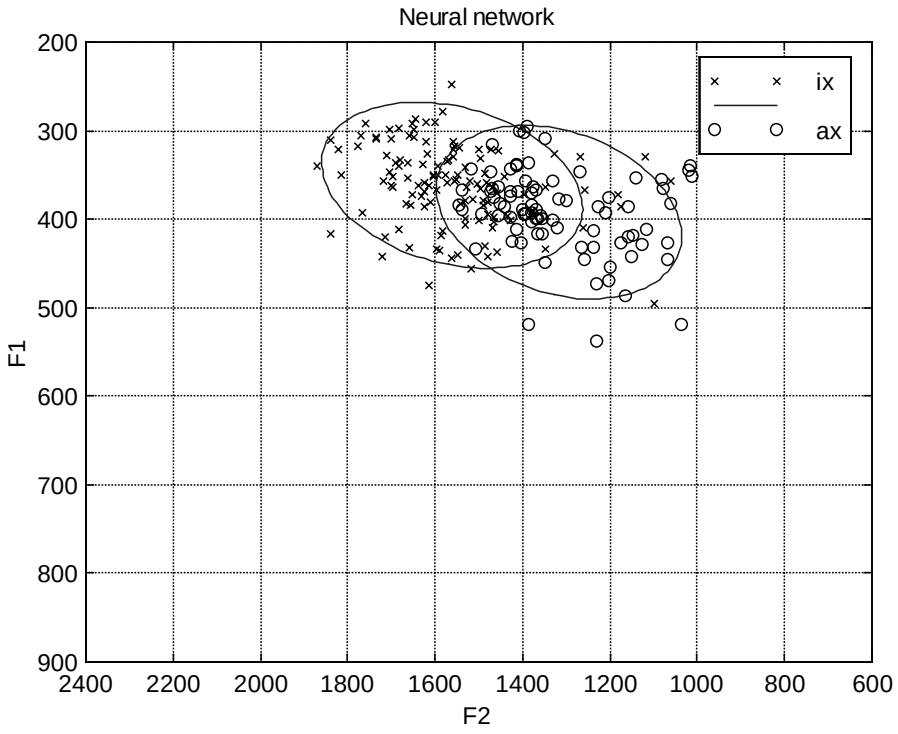

Figure 4-2: Formant distribution of [ə] and [ɨ] in normal neural network with allophonic labels

Note: Legend uses TIMIT labels.

Figure 4-3 shows the distribution in formant space of [ə] and [ɨ] after training with an acoustic neural network where tokens of both allophones were represented by a single phonemic label, /ə/, and features for /ə/. The clouds of formant values overlap to a large degree; however, there still appear to be two distinct clumps, as borne out by the K-S statistic.

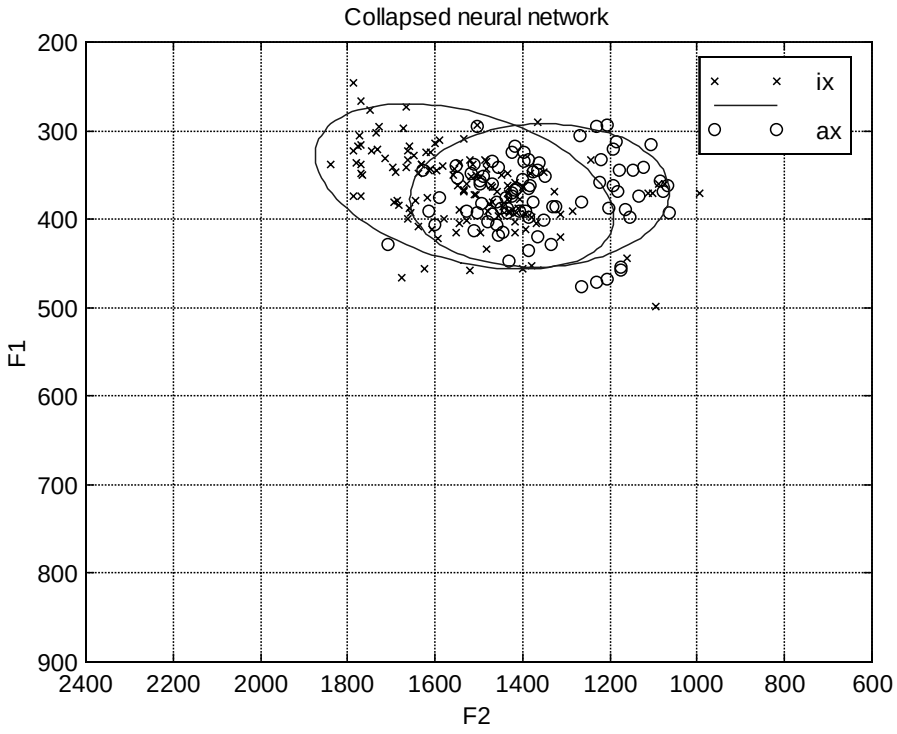

Figure 4-3: Formant distribution of [ə] and [ɨ] in collapsed neural network with single phonemic label used in training for both

Note: Legend uses TIMIT labels.

4.2.2. Experiment on [u]/[ʉ]

Labov (1991, 1994) has noted fronting of /u/ in southern, northern and western American dialects as well as British dialects. Labov (1991, 18-19) notes that there are two norms for /u/, one back and one front, in the speech of two young Chicago women, Debbie Wolf and Betty S. For a third young Chicago woman, Jackie H. with two /u/ norms, Labov (1994, 193) provides both **iw** and **uw** word class labels on a chart of her vowel space.

From his figure 6.16 (Labov 1994, 193), it is apparent that the word class distinction between **iw** and **uw** is not diagnostic of whether /u/ tokens are fronted or not. Although he does not provide information on the conditioning of the bimodal distribution of /u/ for these Chicago speakers, Labov (1991, 26) observes that in Philadelphia and London, /u/ nuclei before liquids retain their back position, as in *fool*. Several speakers in London, Norwich, North Carolina and Atlanta described by Labov, Yaeger and Steiner (1972) appear to have bimodal /u/ distributions. However, Labov (1991, 34) notes that fronting of /u/ in Chicago is a relatively new phenomenon:

> Recordings made in Chicago in the 1970s show no fronting at all of /uw/, but in 1988, 1989, and 1990, interviews carried out by the Project on Cross Dialectal Comprehension in Chicago showed uniform[16] fronting of /uw/ to upper central position from all younger female speakers.

It is clear that our speaker, a Chicago male a few years older than the oldest Chicago female found to feature /u/ fronting, is fully participating in this process.

Table 4-1 summarizes the results for hypothesis testing in the [ʉ]/[ʉ̟] experiment. As with [ə]/[ɨ], the K-S statistic is significant in every condition, indicating that the null hypothesis that [ʉ] and [ʉ̟] come from the same distribution can be rejected. As with [ə]/[ɨ], the level of significance decreases from the original condition through the normal and collapsed conditions, indicating that the neural network blurs the distinction between

[16] It is not clear what is meant by "uniform", since Labov (1991) notes that /u/ is bimodally distributed among speakers who feature fronting of /u/.

the two allophones to some extent, but the elimination of the distinguishing allophonic labels appears to cause even more blurring.

When the formants are individuated, we see that F2 (front-back) is significantly distinct, as with [ə]/[ɨ]. F1 is not significant for the original or collapsed condition, but is significant in the normal condition. It is not clear what this indicates, but clearly F1 is not a reliable discriminator between [u] and [ʉ].

Table 4-1: Hypothesis tests for [u]/[ʉ] experiment

condition	KS	F1 t	F2 t
original	8.40E-06	0.54	1.81E-10
normal	3.53E-05	1.31E-04	5.20E-07
collapse	0.04	0.68	0.0045

Note: Insignificant results are shaded gray.

Table 4-2 gives the Euclidean distances between each of the formants in the normal and collapsed condition from the original condition. Interestingly, only the F2 of [u] is more distant from the original in the collapsed condition than in the normal condition. We might have expected the collapsed condition to always be more distant from the original speech than the normal condition.

Table 4-2: Euclidean distances between original speech and normal and collapsed conditions

vowel	formant	normal	collapse
ʉ	F1	207	197
	F2	989	739
u	F1	161	48
	F2	847	1334

Figure 4-1 shows the F1/F2 distributions of [u] and [ʉ]. Note that the legends for this and the subsequent two figures use the TIMIT labels ‘uw’ for [u], and ‘ux’ for [ʉ]. In this figure, the clouds of the two allophones appear not to overlap at all, although they are contiguous.

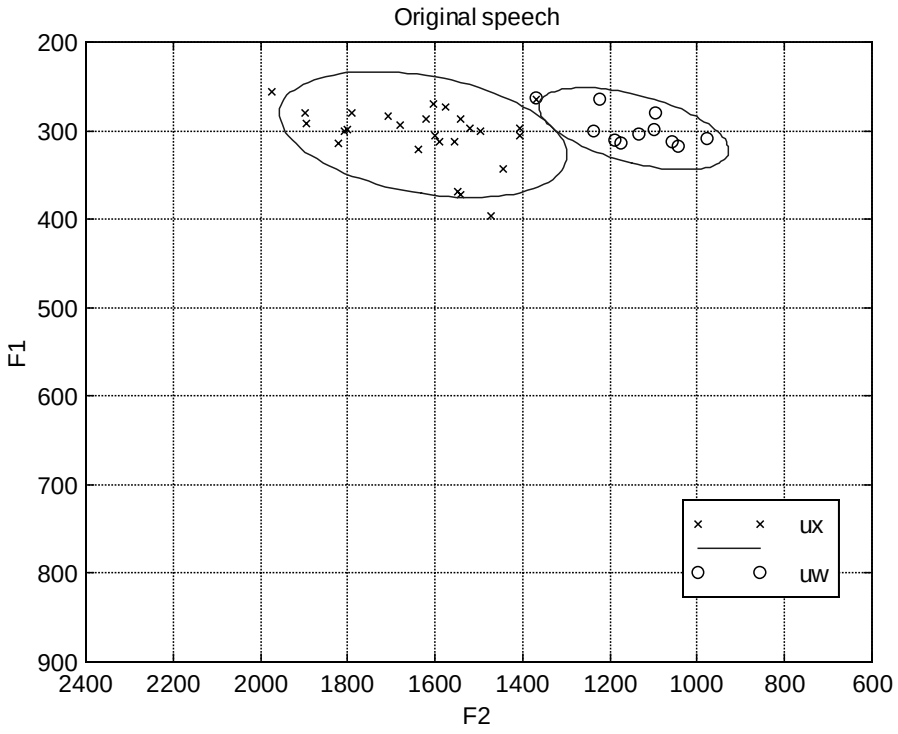

Figure 4-1: Formant distribution of [u] and [ʉ] in original speech

Note: Legend uses TIMIT labels.

Figure 4-2 shows the F1/F2 distribution for [u] and [ʉ] after they were trained with separate allophonic labels. The clouds for each allophone now appear to overlap slightly.

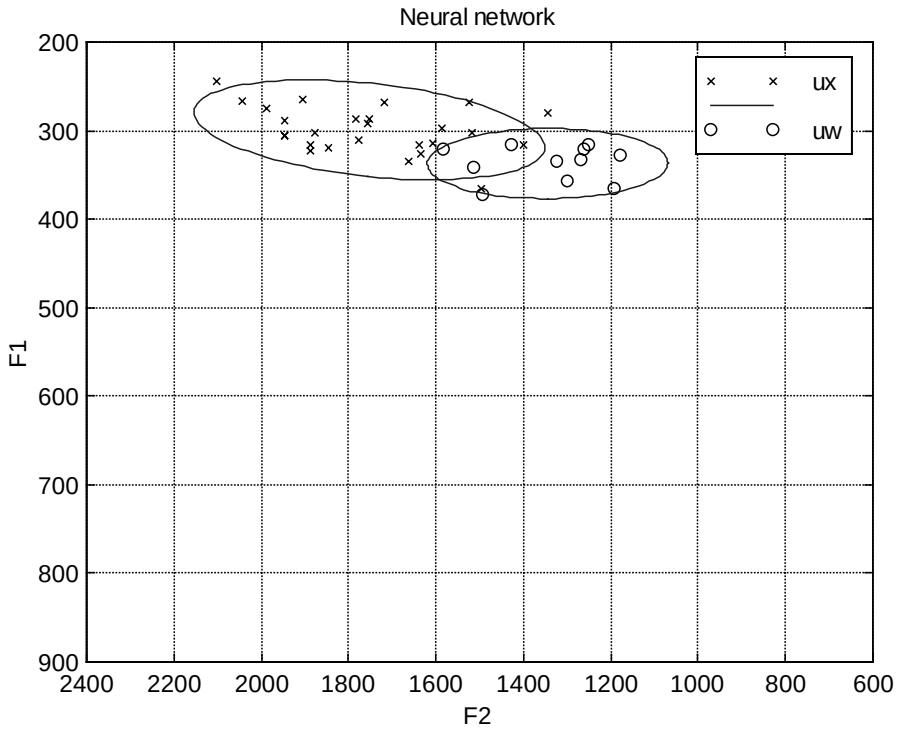

Figure 4-2: Formant distribution of [u] and [ʉ] in normal neural network with allophonic labels

Note: Legend uses TIMIT labels.

Figure 4-3 shows the F1/F2 distribution for [u] and [ʉ] in the collapsed condition, that is, when both were trained with a single phonemic /u/ label and features. The [ʉ] cloud appears to be almost wholly within the [u] cloud. Nevertheless, the two clouds have internal cohesion and there does not appear to be a random mix.

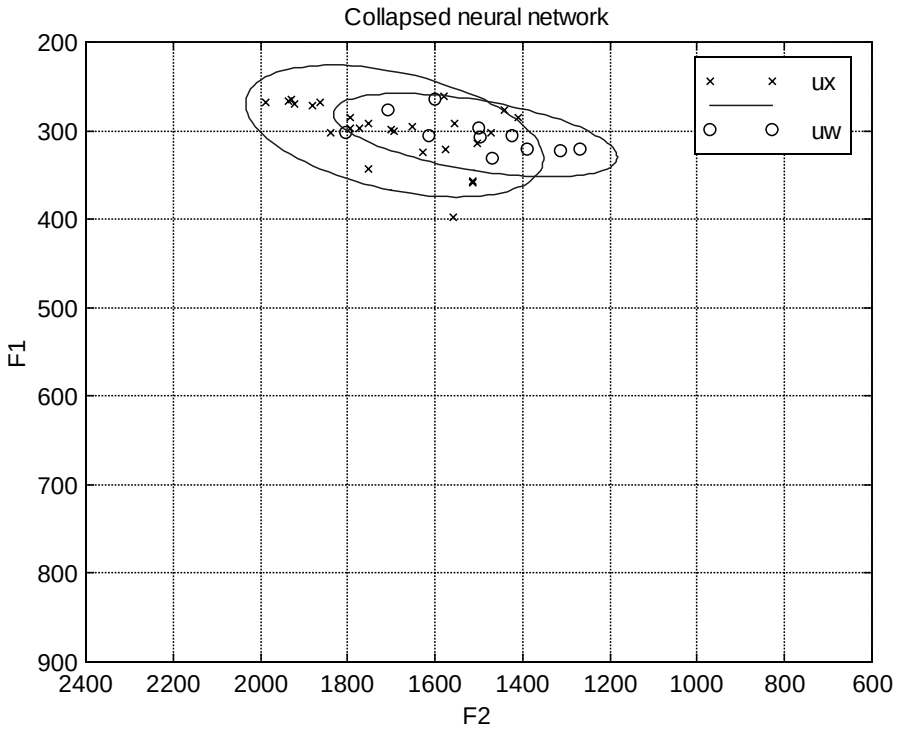

Figure 4-3: Formant distribution of [u] and [ʉ] in collapsed neural network with single phonemic label used in training for both

Note: Legend uses TIMIT labels.

4.2.3. Experiment on /ɑ/ and /ɔ/

With the initial intention of providing a control on the previous two experiments, we performed an analogous experiment using the phonemes /ɑ/ and /ɔ/. In the collapsed case, we collapsed both to /ɑ/. Due to the phonemic status of the two collapsed entities, and the fact that they would have presumably overlapping distributions, we hypothesized

that the collapsed neural network would be unable to preserve a distinction between them based on context alone as in the cases of schwa and /u/ allophones.

As shown in Table 4-1, the statistical results from the formant values of /ɑ/ and /ɔ/ actually resemble those of the allophonic cases examined previously. As might be expected, F1 and F2 taken together with the K-S statistic or separately with the t-test are significant for the original and normal conditions. Contrary to expectations, F1 and F2 taken together and F2 taken alone were still significantly distinct in the collapsed condition. We had thought that since /ɑ/ and /ɔ/ are phonemes, they would occur in similar environments and it would be impossible for the collapsed neural network to establish a distinction based on environment alone, as it did in the allophonic cases examined above. We will discuss the extent to which the environments in which /ɑ/ and /ɔ/ occur actually overlap below. Before that, we will examine the formant displays for the /ɑ/ and /ɔ/ experiments.

Table 4-1: Hypothesis tests for /ɑ/ and /ɔ/ experiment

condition	KS	F1 t	F2 t
original	3.99E-16	2.23E-07	0
normal	4.97E-12	6.42E-04	0
collapse	0.002	0.1077	2.44E-05

Note: Insignificant results are shaded gray.

Table 4-2 gives the Euclidean distances between each of the formants in the normal and collapsed condition from the original condition. As expected, the distance between the original and collapsed condition is greater than that between the normal and original condition for F1 and F2 of both vowels.

Table 4-2: Euclidean distances between original speech and normal and collapsed conditions

vowel	formant	normal	collapse
ɑ	f1	711	966
	f2	774	874
ɔ	f1	472	476
	f2	492	886

Figure 4-1 shows the distribution in formant space of /ɑ/ and /ɔ/. /ɔ/ tends to be higher and backer than /ɑ/, though there is some overlap. It is interesting to note a strikingly similar overlap in the ellipses for /ɑ/ and /ɔ/ drawn by Labov (1991, 19, Figure 1.5) for Betty S., a Chicago woman about ten years younger than our speaker. The closeness of particular phonemes in formant space recalls the concept of near-mergers (e.g. Labov, Karan and Miller 1991), in which there is shown to exist a possible asymmetry between production and perception with respect to phonemic contrasts that are rather small acoustically. Conducting a commutation test (Labov 1994, 356) for /ɑ/ and /ɔ/ for the database speaker in order to assess whether /ɑ/ and /ɔ/ are in a near-merger situation for him remains for future work.

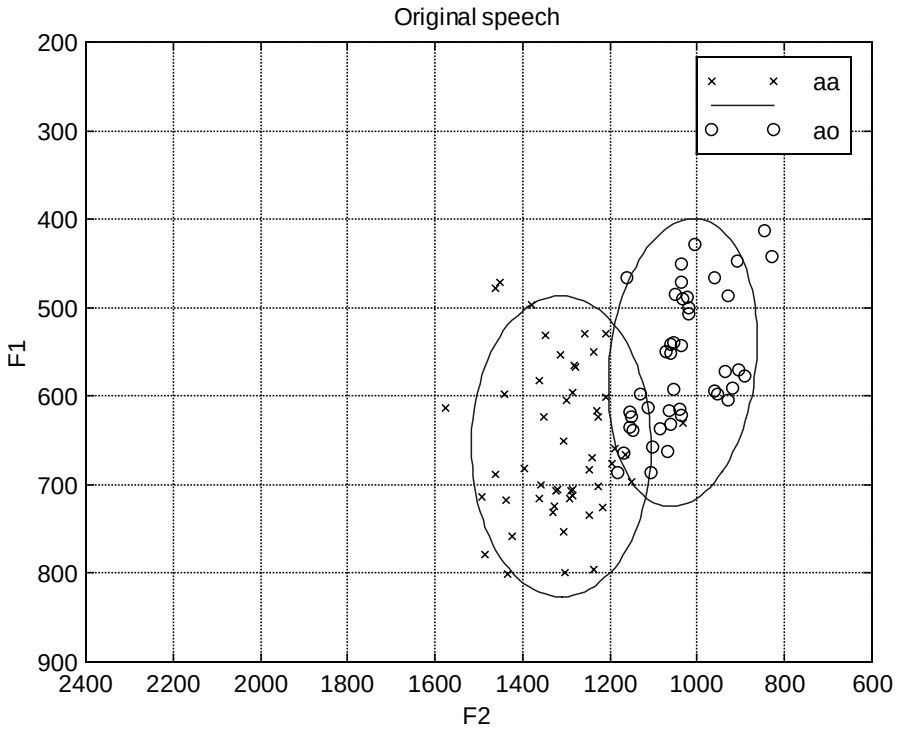

Figure 4-1: Formant distribution of /ɑ/ and /ɔ/ in original speech

Note: Legend uses TIMIT symbols.

As seen in Figure 4-2, the normal neural network slightly enlarges the overlap area between /ɑ/ and /ɔ/, however the two phonemes are largely distinct.

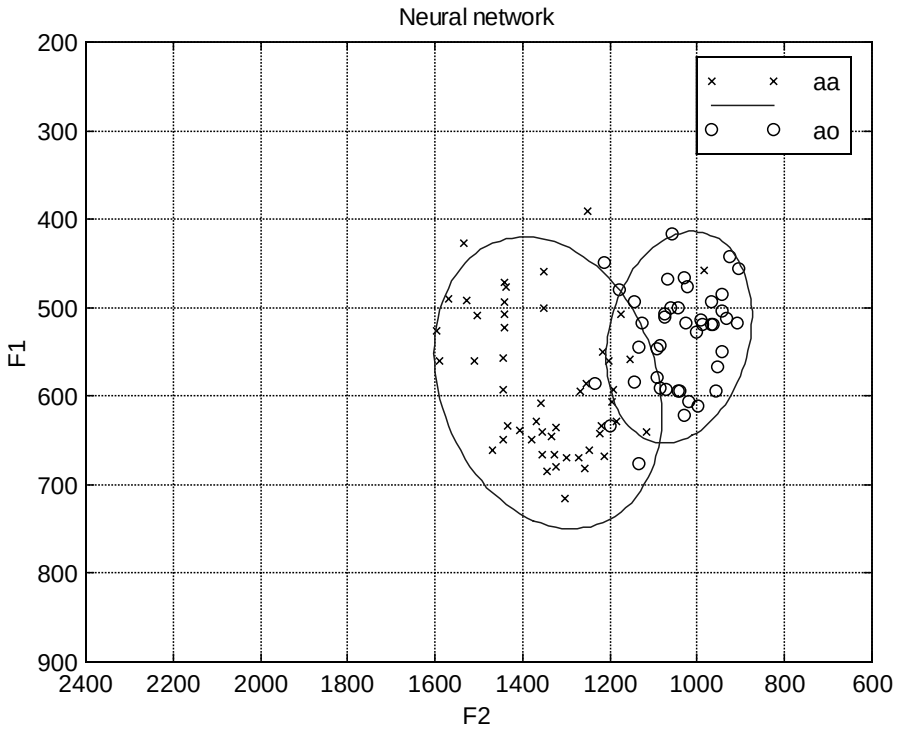

Figure 4-2: Formant distribution of /ɑ/ and /ɔ/ in normal neural network

Note: Legend uses TIMIT symbols.

In Figure 4-3, we see that the collapsing of the phone labels of /ɑ/ and /ɔ/ to /ɑ/ has resulted in substantial overlap between the two phonemes in formant space. In fact, /ɔ/ appears to be contained almost completely within the space for /ɑ/. This recalls Figure 4-3, in which [u] was seen to be contained in [ʉ]. Despite the visual overlap, the formant midpoints of /ɑ/ and /ɔ/ were seen to be statistically distinct, both when F1 and F2 were taken together and F2 was taken alone, as shown in Table 4-1. In addition the

shape and orientation of the /ɑ/ and /ɔ/ ellipses appears to be maintained between the normal condition in Figure 4-2 and the collapsed condition in Figure 4-3.

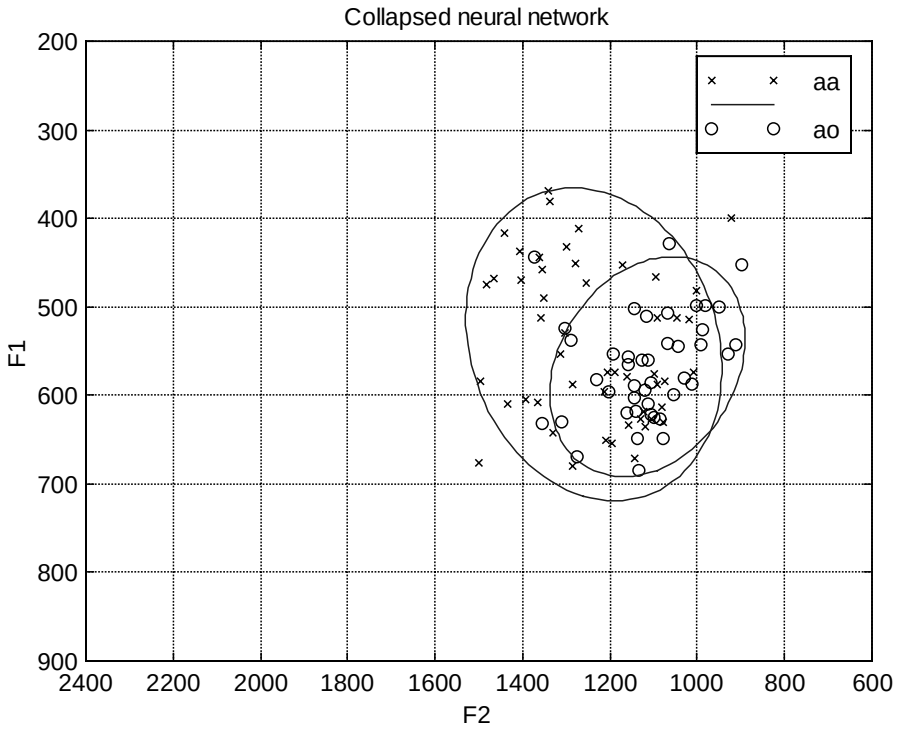

Figure 4-3: Formant distribution of /ɑ/ and /ɔ/ in collapsed neural network

Note: Legend uses TIMIT symbols.

How could the collapsed neural network have preserved a distinction between /ɑ/ and /ɔ/ unless it could appeal to distributional facts? We examined the distribution of /ɑ/ and /ɔ/ in the entire database in an effort to see to what extent environments for each phoneme were overlapping or complementary. Due to the fact that following environment has been shown to explain various mergers and splits such as the British

broad *a,* involving [ɑ] and [æ], and the Middle Atlantic short *a,* involving [æ] and [æ̝][17] (Labov 1994, Chapter 11) we chose to examine it with respect to the /ɑ/ and /ɔ/ distribution in our speaker.

Figure 4-4 shows the number of occurrences of /ɑ/ and /ɔ/ before various phones as they were labeled in our corpus, not crossing word boundaries.[18] /r/ was by far the most common following environment, but because it was approximately equally represented following /ɑ/ and /ɔ/, it was eliminated from the figure. As can be seen, certain environments are exclusively preceded by /ɑ/, such as /m/ and /p/. One environment, following /d/, was exclusively preceded by /ɔ/, but these may well be a result of a paucity of relevant data.

Nevertheless, several patterns are clear: preceding /l/, /ŋ/, /g/, and the voiceless fricatives /s/, /f/ and /θ/, /ɔ/ is much more common than /ɑ/; while preceding /n/, /k/ and /b/, /ɑ/ is much more common than /ɔ/. In a discussion of short *o* in American English, Thomas (1958, 119-120) notes that /ɔ/ is the usual pronunciation before voiceless fricatives and nasals, while /ɑ/ is usual before the stops /p/, /b/, /t/ and /d/. Bronstein (1960, 164) notes that orthographic *o* followed by "velar /k/, /g/ or /ŋ/, /ɑ/ predominates in Eastern

[17] The tensed raised version of [æ] can be realized in several different ways— this is just one example.

New England and New York City...while /ɔ/ predominates elsewhere". Given that the environments for /ɑ/ and /ɔ/ are somewhat skewed, though apparently not in complementary distribution, perhaps it is less surprising that the collapsed neural network was able to do as well as it did in distinguishing these two phonemes.

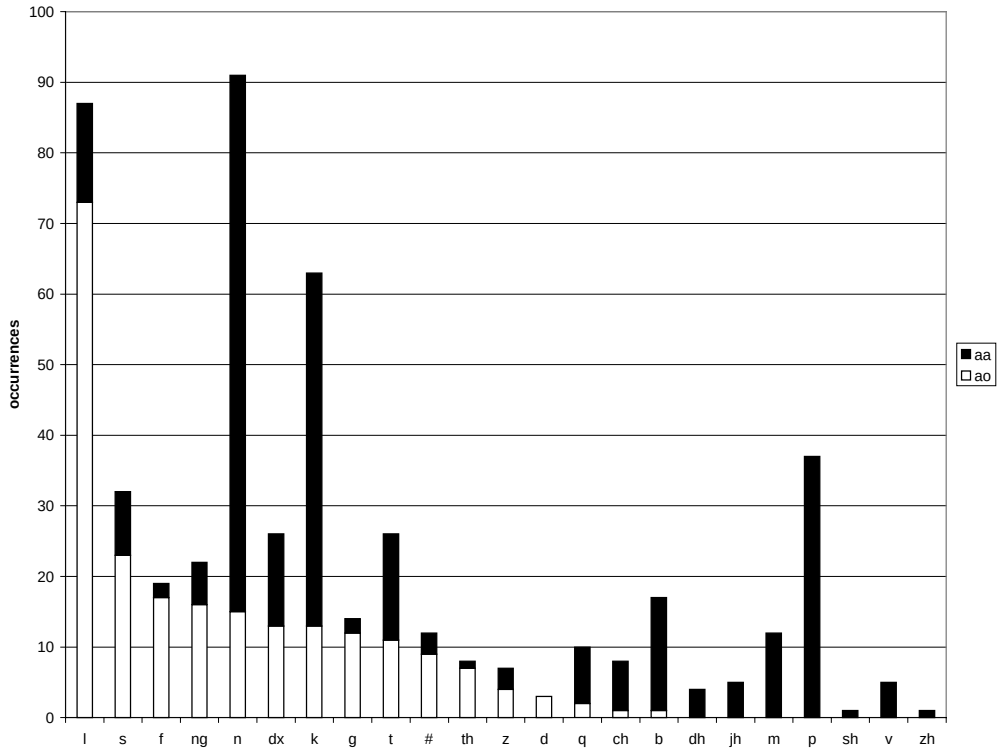

Figure 4-4: Phones following /ɑ/ and /ɔ/

Note: Figure uses TIMIT labels.

[18] We did simplify various varieties of stops for the purposes of this figure, for example, conflating released and unreleased versions.

4.2.4. Experiment on /o/ and /i/

As mentioned earlier, the original intention of the /ɑ/ and /ɔ/ experiment was to demonstrate how collapsing two phonemes would result in chaos in the collapsed condition, due to the fact that phonemes are expected to have overlapping distributions. As we saw, the fact that the distributions of /ɑ/ and /ɔ/ in our speaker's dialect are somewhat skewed, and that /ɑ/ and /ɔ/ are adjacent in acoustic space, resulted in collapsed condition results similar to those encountered with "true" allophones such as [u]/[ʉ] and [ə]/[ɨ].

In order to find a better control experiment, we chose to examine /o/ and /i/, in the original, normal and collapsed condition, where both vowels were labeled /o/. As can be seen in Figure 4-1, in the original speech, /i/ is high and front, and /o/ is low and back and there is absolutely no overlap between them. If these the labels for these two phones were collapsed as /o/, for example, the neural network would be presented with acoustically very different vowels with the same label, even more so than in the previous examples. In addition, in the collapsed case, contextual clues that a vowel is underlyingly an /o/ or an /i/ would be minimal, since both vowels are found in identical contexts throughout the vocabulary.

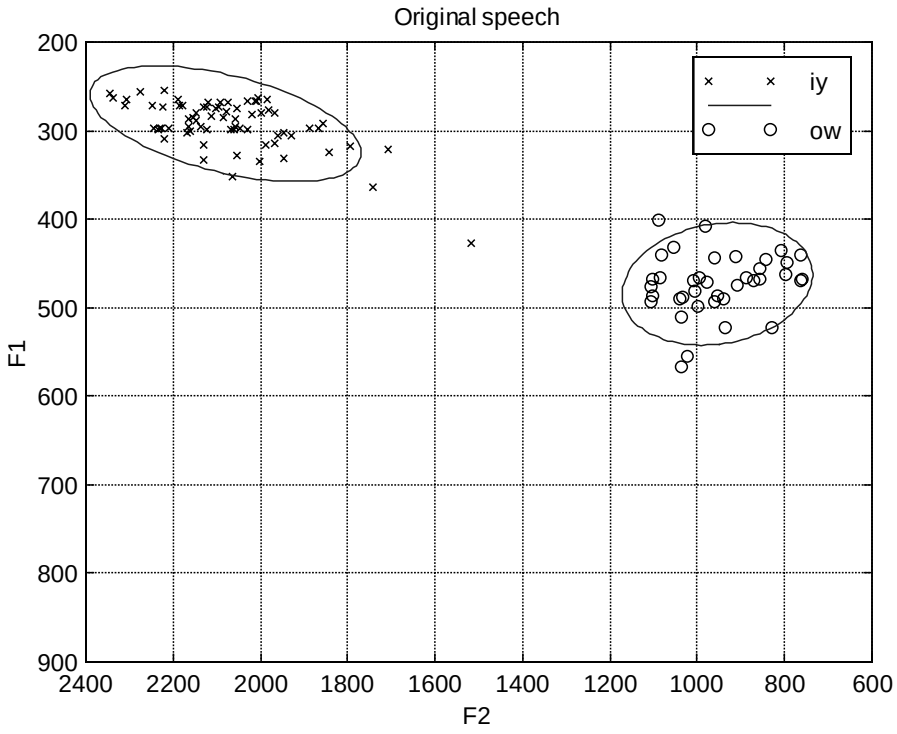

Figure 4-1 Formant distribution of /o/ and /i/ in original speech

Note: Legend uses TIMIT symbols.

The hypothesis tests in Table 4-1 indicate that whether taken together or separately, the F1 and F2 values of /o/ and /i/ are significantly distinct in both the original and normal conditions. Perhaps somewhat unexpectedly, the collapsed results are also significantly distinct. This results may be due to the fact there were almost twice as many /i/ tokens as /o/ tokens, resulting in some skewing of the collapsed data.

Table 4-1: Hypothesis tests for experiment on /o/ and /i/

condition	KS	F1 t	F2 t
original	4.74E-18	0	0
normal	4.50E-18	0	0
collapse	0.0003	1.20E-05	3.20E-05

As can be seen in Table 4-2, the distances between the collapsed and original conditions are much greater than those between the normal and original conditions. It is not surprising that such a great distortion would be introduced by confounding labels of such discrete phonemes.

Table 4-2: Euclidean distances between original speech and normal and collapsed conditions

vowel	formant	normal	collapse
o	F1	271	692
	F2	693	3501
i	F1	278	619
	F2	1510	4703

Figure 4-2 shows that the normal neural network has not much changed the relationship between /i/ and /o/ that held in the original speech.

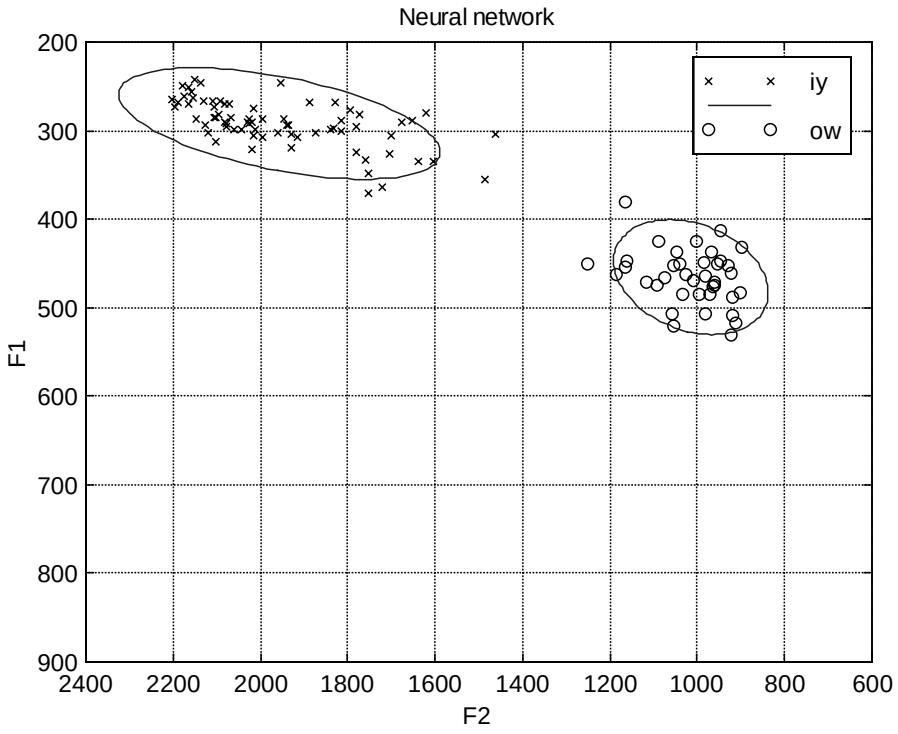

Figure 4-2: Formant distribution of /o/ and /i/ in normal neural network

Note: Legend uses TIMIT symbols.

As can be seen in Figure 4-3, collapsing /i/ to /o/ results in a display that still contains two distinct clouds, but each cloud contains what appears to be a random mix of underlying /o/ and /i/ tokens. The collapsed neural network saw one label, /i/, associated with two acoustically distinct groups of phones. Interestingly, the network has preserved two acoustically distinct groups in the output, though these groups appear not to have to do with whether the original label was /o/ or /i/. In addition, the /i/-like cloud is somewhat lower and backer than in the normal condition, and the /o/-like cloud is somewhat higher and fronter.

These results indicate that collapsing /i/ and /o/ is disastrous for speech synthesis, in the sense that one could not be sure to get an /i/ or /o/ like token when it was required. Of course, this experiment was not performed with the hope of instantiating such a collapse in a working system, rather it was meant to demonstrate, perhaps indirectly, the validity of the other experiments, in which the collapsed results indicated that symbolic mediation was not absolutely required to maintain the distinction between allophones or quasi-allophones, as in the case of /ɑ/ and /ɔ/.

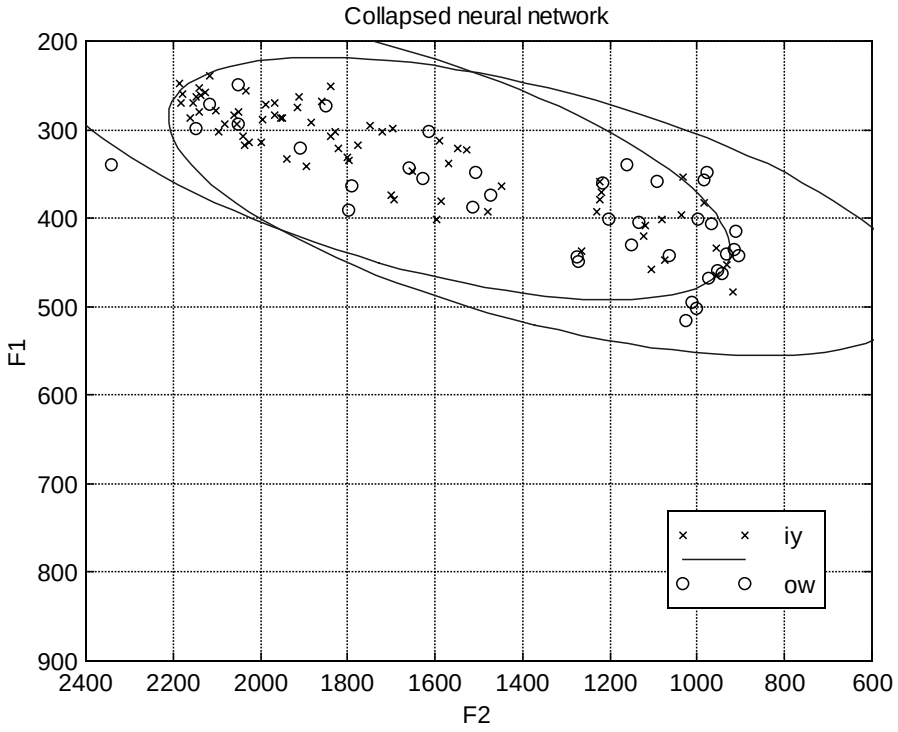

Figure 4-3: Formant distribution of /o/ and /i/ in neural network with /i/ collapsed to /o/

Note: Legend uses TIMIT symbols.

4.3. Conclusions from acoustic analyses of allophony

Perhaps the most interesting outcome of the acoustic investigations of allophony described here is that the allophone distributions in F1/F2 space remained distinct for [u]/[ʉ], [ə]/[ɨ] and /ɑ/ / /ɔ/ despite the removal of the individual allophone labels and features in the collapsed condition. This indicates that the allophony in those cases is predictable from context alone. This is of course what phonemic theory predicts. Indeed, in our implementation, there are other factors, such as duration, available to the acoustic neural network. Perhaps the coherent durational properties of the allophones assisted in their being distinguished by the collapsed neural network. The experiment on /o/ and /i/ showed that if truly contrastive phonemes, with overlapping distributions, are collapsed, the acoustic neural network will not be able to preserve the distinction between the two phones.

The Euclidean distance measures generally showed that the collapsed condition was more distant from the original speech than the normal neural network condition. This result indicates that collapsing phone labels often reduces the ability of the neural network to reproduce the original behavior.

The graphical results also show that the acoustic neural network appears to keep the allophones more apart when it is trained with separate allophonic labels in the normal condition. If indeed vowel space dispersion is important to intelligibility, as Bradlow et al. (1996) claim, perhaps it is worth preserving the allophonic labels in the labeled

database. It is difficult to envisage subjective evaluations involving discriminating the true allophones from each other, since phonemic theory predicts that they will be imperceptible to the lay listener. However, discrimination tests involving /ɑ/ and /ɔ/ might prove revealing.

There are reports that marginal distinctions do exist between [u]/[ʉ] and [ə]/[ɨ] among some speakers; Giegerich (1992, 246), for example, struggles with the phonemic status of schwa, given a distinction among some speakers between the second vowels of *purest* [ə] and *purist* [ɨ]. Wells (1982, 167-168) discusses other pairs such as *Lennon/Lenin*, and claims that General American occupies an intermediate position between merging the two phones, what he calls the Weak Vowel Merger, and treating them as distinct.

As for [u] and [ʉ], Kenyon and Knott (1953, xliii) describe a possible distinction between the vowels in *brewed* and *brood,* and *lute* and *loot,* with the former member of each pair having [ʉ], and the latter [u]. Labov (1994, 162) considers the fronted members of each pair to belong to the **iw**[19] word class, which are either reflexes of French *u* or the Middle English diphthongs exemplified in *few* and *dew*. These are contrasted with the **uw** word class, of which words like *boot* and *do* are members. According to Labov (1994, 162), "The phonemic distinctiveness of **iw** was the result of the loss of the conditioning /j/ glide after apicals, creating a contrast between *dew* and *do, crude* and *croon, lute* and *loot*. The

[19] As a notational convention, we will place Labov's phoneme classes (1991, 13, derived from Trager and Bloch 1941) in bold.

contrast disappeared in many dialects and remained marginal in others." We will explore the correlation of the **iw** word class with fronting of /u/ further in section 6.4.

In future work, we would like to examine whether other non-vocalic allophones, such as the flapped or glottalized variants of /t/, can be learned by the acoustic neural network without an allophonic label. In such cases, perhaps a speech recognizer could be used to assess the degree to which the allophonic target was reached in the different experimental conditions. Zue and Laferriere (1979) discuss some of the gradient acoustic properties of flapping, while Pierrehumbert and Frisch (1997) discuss acoustic correlates of glottalization.

Removing real allophonic labels (such as [ʉ]/[ʉ̟] and [ə]/[ɨ], and perhaps /ɑ/ and /ɔ/) from the labeled corpus, at least those that are not structure preserving (Kiparsky 1985), might simplify the labelers' task, and move us closer to Barry and Fourcin's broad-phonetic labeling level. As concrete evidence of the added difficulty of allophonic transcription, we can examine the data in Table 4-1. This table contains results reported by Fulop and Keating (1996) on inter-transcriber reliability on a Switchboard transcription task at UCLA compared with results for English obtained at CSLU reported by Lander et al. (1995). At both locations, inter-transcriber reliability suffered when agreement on diacritics was counted. Fulop and Keating used diacritics to mark particular nondistinctive phenomena including nasalization and aspiration in English.

Table 4-1: Inter-transcriber reliability at two locations

Condition	UCLA	OGI
counting diacritics	77.5%	55%
excluding diacritics	80.1%	67%

Lander et al. (1995, 169) state the importance of a small label inventory to achieving labeling agreement, even above knowledge of the language to be labeled:

> Label inventory seems to influence agreement more than knowledge of the language, because although transcribers were familiar with Spanish and English, they agreed more often in Spanish, with its smaller label inventory.

In Chapter 6, we will examine the postlexical neural network's ability to learn some of the symbolic distinctions analyzed here in acoustic space. We hope to show that the level of accuracy achieved in symbol prediction, coupled with the higher fidelity of the uncollapsed neural networks for phonetic implementation, will provide a strong basis for a text-to-speech system capable of capturing the subtleties of speaker-specific variation.

Chapter 5. Methods for learning segmental postlexical variation

In this chapter, we will describe the mechanisms by which we learned postlexical variation. In Chapter 3 we described the training data for this enterprise; namely, a lexical database of lexical pronunciations and a labeled speech corpus of postlexical pronunciations. First, we will describe how these two data sources were combined to form a postlexical pronunciation database, with each postlexical phone aligned with its corresponding postlexical phone. We will then attempt to characterize the learning problem by examining details about lexical-postlexical correspondences. Finally, we will describe the neural network architecture we have designed to learn a mapping between the lexical and postlexical domains.

5.1. Creation of postlexical training materials

In order to learn a mapping between lexical and postlexical phones, we needed to combine the postlexical information from the labeled speech corpus with the lexical information in the lexical database. This information needed to be combined in a way that would indicate the relationship between particular lexical phones and particular postlexical phones. That is, we needed to encode the faithfulness, or input-output alignment, between the two levels of representation. We will first describe how the lexical-postlexical alignment was achieved. We will then describe how the aligned lexical and postlexical information was combined into a postlexical database.

5.1.1. Alignment of lexical and postlexical phones

As was discussed in section 3.2, one of the labeling tiers of the labeled corpus contains the orthography for each word.[20] Using the Motorola speech synthesizer's text analyzer (Karaali et al. forthcoming), the disambiguated lexical pronunciation associated with each orthography was obtained from the lexical database. At this stage, we had lexical and postlexical pronunciation strings for each word. Table 5-1 shows examples of such lexical and postlexical pronunciation strings for the word *friendly*. We call the juxtaposition of these pronunciations "unaligned" because we have made no effort to ensure that corresponding lexical and postlexical phones are associated with one another.

Table 5-1: Unaligned lexical and postlexical phones

level	1	2	3	4	5	6	7
lexical	f	r	ɛ	n	d	l	i
postlexical	f	r	ɛ	n	l	i	

Since the learning method we will describe below is based on predicting the appropriate postlexical phone for each lexical phone, it is important that our training data associate postlexical phones with the lexical phone from which they are derived. Although the first four phones in Table 5-1 exhibit the correct lexical-postlexical associations, a mismatch

[20] In the read speech of the current database, this is usually fairly easy to assign. However, in corpora of spontaneous speech, some deliberation is often required due to the extremely close juncture of certain words and cliticization.

has developed at phone 5, due to the deletion[21] of lexical /d/ at the postlexical level. The impact of such a misalignment is great in our approach since the various postlexical phones that occur opposite particular lexical phones are considered allophones of those lexical phones. So, in the case of position 5 in the unaligned strings in Table 5-1, [l] would be considered an allophone of /d/.

In order to rectify this situation, we developed a method for aligning lexical and postlexical phones so that we might achieve an alignment as shown in Table 5-2. The alignment has inserted a placeholder in the postlexical level at position 5. Thus lexical /d/ can now be considered to have "deletion" as a possible allophone; an improvement over the unaligned situation, and one that is consistent with treatments of *t,d* deletion in phonology, sociolinguistics and speech technology (e.g. Guy 1980, 1991, Reynolds 1994, Randolph 1989).

Table 5-2: Aligned lexical and postlexical phones

level	1	2	3	4	5	6	7
lexical	f	r	ɛ	n	d	l	i
postlexical	f	r	ɛ	n		l	i

[21] In accordance with the symbolic approach to allophony that we are taking in this part of the dissertation, we are simplifying and calling this a deletion. Of course, there is likely to be some remnant of a /d/ gesture (cf. Browman and Goldstein 1990). Sociolinguistic studies of *t,d* deletion (e.g. Guy 1980, 1991) make this same simplification.

Since manual alignment is a long and tedious process, several proposals have been advanced to perform this task automatically. Many of these attempt to align orthographies with their corresponding phonetic or phonological transcriptions. In our view, this is a more challenging problem than our current one of aligning lexical and postlexical pronunciations. For example, the orthography-phonetics alignment problem features many-to-one mappings in both directions: orthographic 'ough' can map to phonetic [ɔ] as in *bought*, and phonetic [ks] can map to orthographic 'x' as in *tax*. To avoid this complexity, some researchers have chosen to simplify this problem and impose a one-to-one mapping between orthography and phonetics (Sejnowski and Rosenberg 1987, Laporte 1997), resulting in a potential loss of explanatory value of their proposals. In contrast, we feel that a one-to-one alignment between lexical and postlexical phones is less problematic, given the more straightforward mapping and taking into account certain modifications to the aligned phone set to be described below.

There have been many different methods proposed for the alignment of different levels of representation as well as tokens from different languages. Dedina and Nusbaum (1991) describe a simple approach that parses spellings and pronunciations into separate groups of consonants and vowels, and then maps consonant spelling groups to consonant phoneme groups, and vowel spelling groups to vowel phoneme groups. In the domain of historical linguistics, Covington (1996) similarly found that a simple consonant/vowel distinction was sufficient for aligning words from different languages presumed to be cognates. Knight and Graehl (1997, 131-132) use the estimation-maximization (EM)

algorithm (Baum 1972) to align English sounds with their transliterated Japanese equivalents. Lawrence and Kaye (1986) describe a table-based approach to aligning orthographies and their pronunciations.[22]

The approach we have chosen to follow employs dynamic programming (Kruskal 1983). Dynamic programming is a general method used for comparing sequences. It attempts to find the least costly path between two sequences, employing the operations of insertion, deletion and substitution. Each of these operations entails a particular cost. As will be seen below, the attribution of these costs is often application-specific. Table 5-3 presents one of Kruskal's examples, a comparison of the orthographies *industry* and *interest*, including placeholders inserted in the alignment process. Table 5-4 shows the operations required to transform *industry* to *interest*.

[22] Van Coile (1991) uses Hidden Markov Models, Lucas and Damper (1992) allow a neural network to come up with its own alignment, and Luk and Damper (1991, 1992, 1993) employ image-processing techniques and dynamic time warping to infer letter-phoneme correspondences.

Table 5-3: Sequence comparison of two orthographies

i	n	d	u	s	t		r		y	
i	n				t	e	r	e	s	t

Table 5-4: Operations required to transform *industry* to *interest*

delete *d*
delete *u*
delete *s*
insert *e*
insert *e*
substitute *y* by *s*
insert *t*

When the sequences to be aligned use the same alphabet, as in the orthographies in Table 5-3, only two costs are usually provided for substitution: a standard cost in the case where the sequence symbols are distinct, and no cost where they are identical. However, this policy prevents certain symbols from being considered closer to one another than other symbols. In the domain of aligning phonetic transcriptions, this would mean that perceptually small differences, such as [ə] and [ɪ], would be considered just as erroneous as large differences such as [b] and [k]. Gildea and Jurafsky (1996), Riley and Ljolje (1996), and Nerbonne and Heeringa (1997) all discuss cost functions that take the features of the phones to be aligned into account. The fewer differences in features between two symbols, the lower the cost of their substitution.

In the case of the alignment of lexical and postlexical phones, two distinct alphabets are being used. The cost function must take this into account. Before describing the nature

of the cost function used here, we will discuss some modifications to the postlexical alphabet that were made to ensure a useful alignment. To facilitate the discussion, we will refer to input alphabet, in our case lexical, as the source alphabet. The output alphabet, in our case postlexical, is the target alphabet.

Recall that the postlexical alphabet encodes both closure and release information for stops with separate symbols. Table 5-5 shows hypothetical lexical and postlexical forms for *top*. In order to achieve an appropriate alignment in this case, a placeholder would be required on the lexical level at position 1 or 2, thus shifting the other lexical phones to the right. At first glance, this may appear to be similar to the solution proposed in Table 5-2, however it is unacceptable. The problem is related to the task at hand. We are seeking to predict postlexical phones given lexical phones, that is, we are predicting the target given the source. If some lexical phones, such as /d/, sometimes surface as null, indicating they were deleted, that is fine. However, there would be no principled way to predict when to have placeholders in lexical strings, or any source strings for that matter.

Table 5-5: Unaligned lexical and postlexical phones without pseudophones

Level	1	2	3	4
Lexical	t	ɑ	p	
Postlexical	tc	t^h	ɑ	pc

To alleviate this problem, certain pairs of postlexical phones were collapsed as one "pseudophone". This strategy was adapted from Sejnowski and Rosenberg (1987) who

employed it in the domain of letter-to-sound conversion. They collapsed certain phones, which in their case were the target alphabet, like /k/ and /s/ for alignment with the letter *x*, which was part of their source alphabet. Riley and Ljolje (1996, 296) also handled this problem in a similar away for phonemic-phonetic mapping. Table 5-6 shows the postlexical phones that were collapsed as single pseudophones in our experiments. In general, they are stop closures and releases, and glottal stops and vowels. Use of the pseudophones permits straightforward alignments such as is shown in Table 5-7.

Table 5-6: Postlexical pseudophones

ʔ ɑ	ʔ æ	ʔ ʌ	ʔ ɔ	ʔ aʊ	ʔ ə	ʔ ɚ	ʔ ɨ	ʔ i	ʔ ɪ
ʔ ə̯	ʔ aɪ	ʔ ɛ	ʔ l̩	ʔ m̩	ʔ ɝ	ʔ eɪ	ʔ o	ʔ ɔɪ	ʔ ʊ
ʔ u	ʔ ʉ	ʔ l	ʔ m	ʔ j	ʔ r	ʔ w	ʔ tʃ	bc b	dc d
gc g	pc p^h	pc p	tc t^h	tc t	kc k^h	kc k	dʒc dʒ	tʃc tʃ	

Table 5-7: Aligned lexical and postlexical phones with pseudophones

Level	1	2	3
Lexical	t	ɑ	p
Postlexical	tc t^h	ɑ	pc

Outside of cases where the target postlexical string requires more symbols than the source lexical string due to the transcription conventions, source placeholders, or insertions, are not generally problematic. Since most postlexical phenomena are reductions, the source lexical string often has more phones than the postlexical target string. However, due to the reading style of the postlexical database, as well as possibly due to idiosyncrasies of

individual speakers, postlexical pronunciations (at least as defined here) can occasionally have more phones than the pronunciations in the lexical database.

Table 5-8 provides an example of this phenomenon. The speaker pronounced *exhume* with a glide following the /s/. To solve this problem, the lexical database could have been modified to contain the pronunciation with a glide, since the more common form (in American English) without the glide could be derived from it. This solution was rejected due to the rarity of such a pronunciation in American English (Wells 1982). Alternatively, a pseudophone encompassing postlexical /sj/ could have been introduced. This was rejected, since it would be harmful to a lexical-postlexical alignment of words like *misuse*, where lexical /s/ should align with postlexical [s] and lexical /j/ should align with postlexical /j/. Due to these problems, words like *exhume* pronounced as in Table 5-8 were eliminated from training.

Table 5-8: Example of source insertion

lexical	ɛ	k	s		u	m
postlexical	ɛ	k	s	j	u	m

Table 5-9 shows the cost table used for lexical-postlexical substitutions in our experiments. In accordance with Young et al. (1995), we used a cost of 7 for insertions and deletions, and 10 for substitutions not otherwise covered in Table 5-9. It was

developed by iteratively listing what were thought to be probable lexical-postlexical correspondences, and then examining resulting alignments. While the automaticity of a feature-based alignment (e.g. Riley and Ljolje 1996, Gildea and Jurafsky 1996) might be desirable, the flexibility permitted by the "freer" approach described here was preferred. In addition, it is by no means clear that postlexical reflexes of a given lexical phone are necessarily close in feature space— consider the allophones of /t/, including flap and glottal stop.

Table 5-9: Cost table for lexical-postlexical alignment

lexical	postlexical	cost	lexical	postlexical	cost	lexical	postlexical	cost
ɑ	ɑ	0	aʊ	aʊ	0	ə	ɪ	0
ɑ	ʔ ɑ	0	aʊ	ʔ aʊ	0	ə	ʔ ɪ	0
æ	æ	0	ə	ə	0	ə	ə̰	0
æ	ʔ æ	0	ə	ʔ ə	0	ə	ʔ ə̰	0
ɔ	ɔ	0	ə	ɨ	0	aɪ	aɪ	0
ɔ	ʔ ɔ	0	ə	ʔ ɨ	0	aɪ	ʔ aɪ	0
b	b	0	tʃ	tʃ	0	ð	ð	0
b	bc b	0	tʃ	tʃc	0	d	d	0
b	bc	0	tʃ	tʃc tʃ	0	d	dc d	0
ɛ	ɛ	0	eɪ	eɪ	0	d	dc	0
ɛ	ʔ ɛ	0	eɪ	ʔ eɪ	0	d	ɾ	0
f	f	0	g	g	0	ɪ	ɪ	0
h	h	0	g	gc g	0	ɪ	ʔ ɪ	0
h	ɦ	0	g	gc	0	ɪ	ɨ	0
i	i	0	dʒ	dʒ	0	ɪ	ʔ ɪ	0
i	ʔ i	0	dʒ	dʒc dʒ	0	ɪ	ə	0
k	k^h	0	dʒ	dʒc	0	ɪ	ʔ ə	0
k	kc k^h	0	l	l	0	m	m	0
k	kc k	0	l	l̩	0	m	m̩	0
k	k	0	ŋ	ŋ	0	n	n	0
k	kc	0	ɔɪ	ɔɪ	0	n	ɾ̃	0
o	o	0	ɔɪ	ʔ ɔɪ	0	n	n̩	0
o	ʔ o	0	r	r	0	ʃ	ʃ	0
p	p	0	r	ɚ	0	s	s	0
p	p^h	0	r	ɝ	0	θ	θ	0
p	pc p^h	0	r	ʔ ɚ	0	t	t	0
p	pc p	0	ʊ	ʊ	0	t	t^h	0
p	p	0	ʊ	ʔ ʊ	0	t	tc t^h	0
u	u	0	v	v	0	t	tc t	0
u	ʔ u	0	w	w	0	t	tc	0
u	ʉ	0	j	j	0	t	ʔ	0
u	ʔ ʉ	0	ʒ	ʒ	0	t	ɾ	0
z	z	0	ə	ɚ	0	n	ŋ	1

There are certain cases of artifactual postlexical deletion that are a product of some of the lexical database transcription simplifications discussed in section 3.1. For example, syllabic liquids and nasals are expressed with two symbols in the lexical database in order to reduce the size of the lexical phone inventory and consequently to increase the consistency of lexical pronunciations. Given our alignment procedure, the initial lexical schwas in such transcriptions are often "deleted" at the postlexical level, while the lexical liquids and nasals are associated with syllabic liquids and nasals at the postlexical level.

Sproat and Riley's (1996) description of a decision tree for determining the postlexical realization of lexical /ɑ/ in the multispeaker TIMIT database points out an additional issue encountered in aligning lexical and postlexical pronunciations. One of the postlexical reflexes is found to be /ɔ/. Now, /ɑ/-/ɔ/ alternation does not appear to be a lexical-postlexical issue. It seems more likely that the TIMIT speakers differ on whether they store the initial vowel of a word like *sausage* in the lexicon as /ɑ/or /ɔ/.[23] Nevertheless, this represents the kind of dialect smoothing that our postlexical network performs, as described above in section 2.2, in addition to general postlexical processes.

5.1.2. Creation of postlexical training database

We created a relational database in Microsoft Access to contain the aligned lexical and postlexical information from the labeled corpus and lexical database. The database

structure was inspired by Keating et al. (1994), who used a similar database to analyze the non-speech information contained in the TIMIT database. Our database differed from that of Keating et al. in that we had no need to supply a table containing names of speakers, since our procedure only analyzes data from one speaker at a time. An innovation with respect to Keating et al. was the inclusion of lexical phone information in addition to postlexical labeled phones, allowing for queries directly probing the success of learning postlexical variation.

Table 5-1 displays the fields in the main table of the database. These include fields for both original postlexical phones and neural network hypotheses of postlexical phones. The generation of such hypotheses will be discussed in the following sections. Field names followed by a question mark (?) indicate fields containing binary "yes/no" values. The previous phone ID and next phone ID fields allow for querying all of the field information of the next and previous phones, including the next and previous original postlexical phones, neural network postlexical phone hypotheses and lexical phones.

[23] *Sausage* (like *chocolate*) is an example of a word whose initial vowel tends to be /ɑ/ in Philadelphia and /ɔ/ in New York, neither of which cities is currently undergoing a merger in the low back area.

Table 5-1: Fields in lexical-postlexical database

orthographic word
original postlexical phone
neural network postlexical phone hypothesis
lexical phone
previous phone ID
next phone ID
syllable stress
word prominence
word type
syllable accented?
left syllable boundary?
right syllable boundary?
left word boundary?
right word boundary?
left phrase boundary?
right phrase boundary?
left clause boundary?
right clause boundary?
left sentence boundary?
right sentence boundary
left intermediate phrase boundary?
right intermediate phrase boundary
left intonation phrase boundary?
right intonation phrase boundary?
phone ID
file

5.2. Characterization of learning problem

One way of posing the problem of mapping from lexical pronunciations to postlexical pronunciations is as follows: given a lexical phone, what postlexical phone is most appropriate? In some cases, this choice will be fairly easy due to a lack of allophony occurring with respect to particular lexical phones. For example, Fulop and Keating (1996) found that stridents did not undergo much alteration between the lexical and

postlexical levels. In their study of transcription in the Switchboard corpus, Fulop and Keating examined each lexical phone and counted the number of times the lexical symbol appeared in the postlexical transcription for each phone. While this metric provides some idea of where the variability lies, we feel that it is not the best measure to use. One problem is that certain lexical symbols, such as /t/, have a different meaning when used on the lexical, as opposed to the postlexical level. For example, on the lexical level it combines both the closure and release aspects of /t/, whereas on the postlexical level it indicates an unaspirated *t* release.

The metric that we propose to examine the variability among the different phones with respect to lexical/postlexical mapping is entropy. Entropy can be seen as a measure of uncertainty. In this case, given a lexical phone, how certain can we be of its postlexical output? Entropy is an information-theoretic measure that describes the lower bound on the number of bits needed to encode a message. In our case, we will look at each lexical phone as a message whose content could be any one of its postlexical reflexes. Borrowing from the discussion of entropy in Charniak (1993, 29), a message can be thought of as a random variable W that can take on one of several values V(W) and has a probability distribution P. In our case, W is a lexical phone, and the values V(W) are the postlexical reflexes. The probability distribution will be the probability of each reflex occurring on the postlexical level for the lexical phone in our testing database. Equation 5-1 contains the formula for calculating entropy, H, for each lexical phone W.

Equation 5-1: Entropy

$$H(W) = -\sum_{w \in V(W)} P(w) \log_2 P(w)$$

Table 5-1 lists each lexical phone with the number of each of the postlexical reflexes with which it was aligned in a testing subset of our lexical/postlexical database. These values were supplied to a program that calculated the entropy of each lexical phone, the results of which are shown in Figure 5-1. This can be compared to Figure 5-2, which shows the number of postlexical reflexes of each lexical phone. As can be seen in these figures, entropy does not correlate entirely with number of postlexical variants. It provides a measure that takes into consideration the distribution of the variants as well.

Table 5-1: Postlexical reflexes of each lexical phone

ɑ		æ		ə		i		eɪ	
ɑ	405	æ	532	del.	622	i	642	eɪ	521
ʔ ɑ	17	ʔ æ	33	ɨ	453	ə	387	ɛ	59
del.	7	ɨ	67	ʌ	354	ɨ	308	ɨ	46
ɨ	4	ɛ	9	ə	337	ɪ	51	ə	38
ʌ	4	ə̰	8	ɪ	79	ʔ i	9	ʔ eɪ	5
ə	3	ə	6	ʔ ʌ	22	ʌ	4	i	4
ɔ	3	del.	3	i	20	ɛ	2	ʔ ɛ	3
aʊ	1	ʔ ɨ	4	ʔ ə	13	del.	1	ʔ ə	2
æ	1	ʔ ɛ	1	ə̰	6	ʔ ɪ	1	ʌ	2
		ɑ	1	ɛ	5	aɪ	1	del.	1
				ə˞	5	eɪ	1		
				ʔ ɪ	3				
				ʔ ɨ	3				
				r	3				
				ɑ	2				
				ʊ	1				
				aɪ	1				
				eɪ	1				
				o	1				
				dc d	1				
				ʉ	1				

ɪ		t		o		d		k	
ɪ	487	tc t^h	566	o	314	dc d	504	kc k^h	557
ɨ	187	tc	273	ɔ	157	dc	235	k^h	187
ə	43	tc t	214	del.	29	d	168	kc k	180
ʔ ɪ	30	ʔ	170	ʔ o	12	del.	137	kc	131
del.	24	t^h	134	ʔ ɔ	6	ɾ	73	k	2
ʔ ɨ	5	ɾ	103	ə	1	tc	2	del.	1
ə̰	2	del.	51	ɑ	1	tc t^h	1	gc g	1
aɪ	2	t	7	ɨ	1	ʌ	1	gc	1
ɛ	1	dc	1	aʊ	1				
ʔ ɛ	1	dc d	1						

u		ɛ		r		g		p	
ʉ	214	ɛ	411	r	1243	gc g	165	pc p^h	386
u	119	ʔ ɛ	23	ɜ˞	415	g	112	pc p	86
ɨ	55	ɨ	17	ə˞	164	gc	16	p^h	79
ə	9	ə	4	del.	15	del.	7	pc	78
ə̰	2	i	4	ʔ ə˞	4	r	1	p	2
ʊ	2	ʔ ɨ	1	ʔ r	1				
w	1								

s		ʊ		ɔ		tʃ		dʒ	
s	1228	ʊ	92	ɔ	210	tʃc tʃ	147	dʒc d ʒ	110
del.	8	del.	23	ɑ	23	tʃ	34	dʒ	35
ʃ	1	ɨ	10	ʔ ɔ	9	ʃ	1	ʒ	7
z	1	ə̰	2	ʔ ɑ	1	tc t^h	1	θ	1
ʌ	1	ʉ	1	aʊ	1				

l		n		ʃ		ð		aɪ	
l	1013	n	1251	ʃ	183	ð	916	aɪ	417
l̩	149	ɾ̃	86	tʃ	1	θ	4	ʔ aɪ	8
del.	4	n̩	37	θ	1	ə	1	ɑ	2
ʔ l	2	ŋ	2	tʃc tʃ	1				

b		h		m		θ		v	
bc b	321	h	216	m	566	θ	178	v	361
b	259	ɦ	105	m̩	2	del.	1	del.	9
bc	16	del.	43	ʔ m	2	s	1	f	2

z		aʊ		f		ɔɪ		w	
z	679	aʊ	151	f	549	ɔɪ	48	w	516
s	10	ʔ aʊ	5	ʌ	1	ʔ ɔɪ	2	ʔ w	4
del.	5								

j		ŋ		ʒ					
j	130	ŋ	221	ʒ	15				
ʔ j	1								

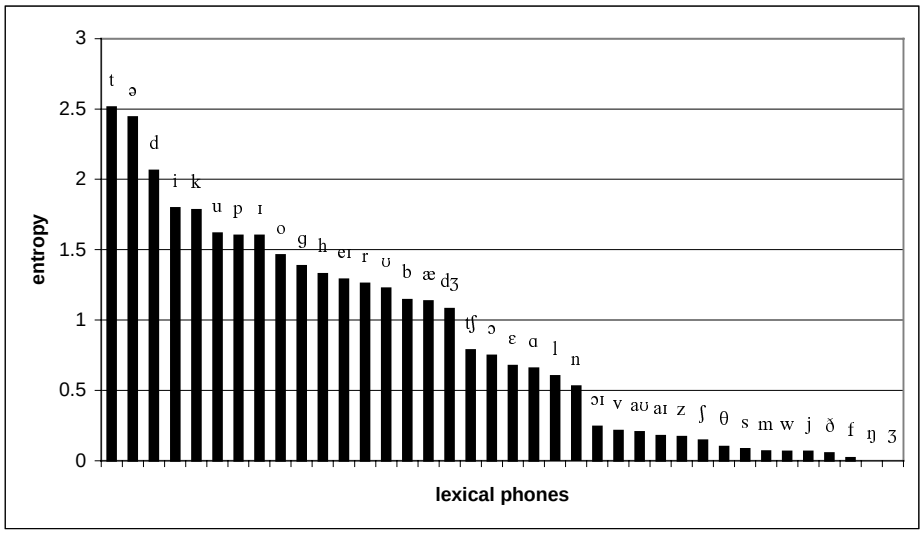

Figure 5-1: Entropy of lexical phones

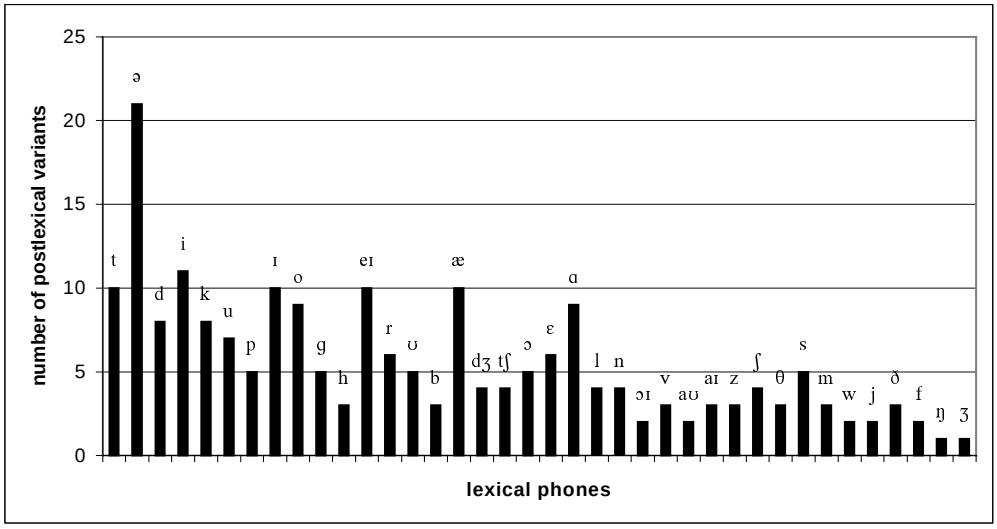

Figure 5-2: Number of postlexical reflexes of each lexical phone

5.3. Neural network architecture

For the experiments described below, we will be using Monnet, a highly advanced neural network simulator being developed at Motorola. The simulator is designed to allow for rapid training, as well as ease of experimentation with different topologies and architectures. The network can be trained with decreasing learning rate and learning momentum in a novel mixture of sequential and random training modes (Karaali 1994).

Figure 5-1 illustrates the postlexical neural network, in the block notation used to describe the acoustic neural network in Chapter 4. In the following subsections, we will describe the lexical input information provided to blocks 2-5 which results in an output lexical phone at block 1. Blocks 6-11 are hidden layers which help to process the input information.

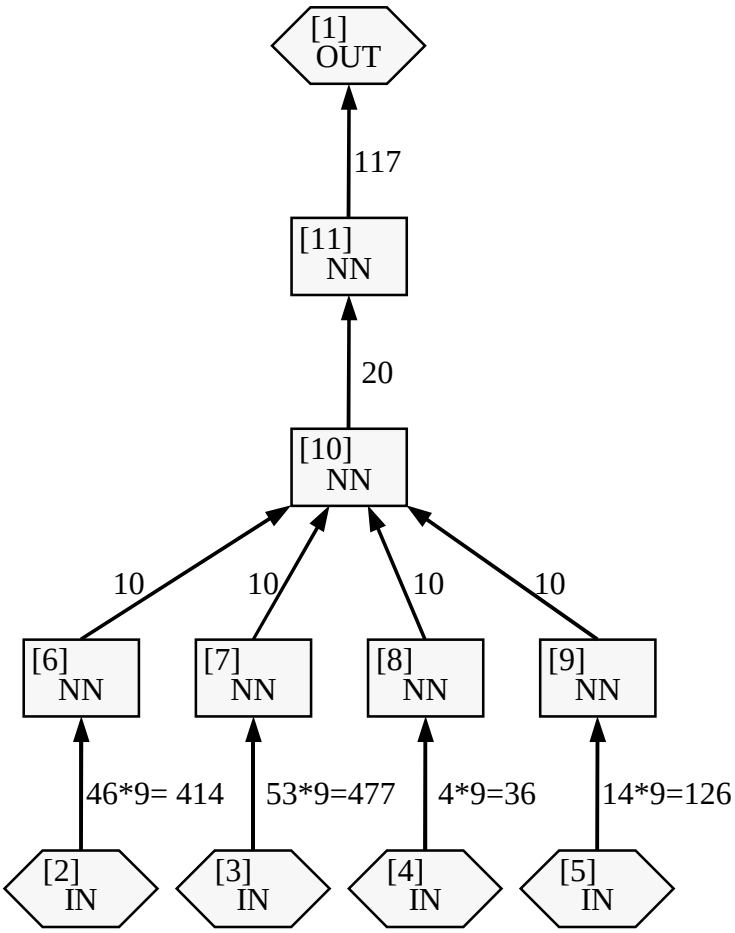

Figure 5-1: Postlexical neural network

5.4. Data encoding

One of the most important aspects of working with neural networks is using an appropriate coding scheme for the input and output representations. This is an area where domain knowledge, in this case phonological and linguistic, is critical. In fact, according to Cohen (1997), “it is better to focus on issues such as representations and prior

knowledge rather than the learning dynamics, since there are no general purpose learning algorithms which are good for an arbitrary prior."

With regard for example to the encoding of syntactic and prosodic information, we draw on various proposals, such as Nespor and Vogel (1986), Pierrehumbert and Frisch (1997), and Dilley et al. (1996) that indicate that prosodic information has an important effect on segmental postlexical rules. However, we merely provide the network with this information, letting it draw its own conclusions about which of this information is useful, to what extent, and for what processes, based on the patterns it finds in the training data. In this way, we are following an empiricist approach that benefits from the introduction of learning biases (in the sense of Gildea and Jurafsky 1996). Our approach differs from a nativist, or rule-based approach, in that we are providing the tools to determine the effects of boundaries and the like, without enforcing how the boundaries and other information must be used. In fact, in many ways, we are conflating the various kinds of postlexical rules, both allophonic and allomorphic, which various proposals in lexical phonology (e.g. Kaisse 1985, Mohanan 1986) have sought to tease apart.

5.4.1. Features for lexical phones

Block 2 is supplied lexical phone label information in a 1 of *n* encoding. Since this encoding is particularly sparse, we sought to permit the network to generalize beyond the specific phone label information by providing featural information for each lexical phone to Block 3. Initial experiments relied on a fairly traditional feature set with values for

activated, unactivated and unspecified features, shown in Table 5-1. The feature table may appear to be rather overspecified. This is because the features shown are actually a subset of a larger feature set that is being used for both phonemes and allophones across a variety of languages for acoustic neural network training. In order to be able to encode diphthongs, each of the vowel features has a variant 1 and a variant 2, e.g. high 1, high 2. Variant 1 contains the feature value for the diphthong's nucleus, while variant 2 contains the feature value for the diphthong's glide. Monophthongs contain identical values for variants 1 and 2 for each feature. In future work, we hope to explore featural representation that employ the innovations of feature geometry, in order to better establish the relationships between features and clusters of features.

Table 5-1: Features for lexical phones

Phone	ə	r	ɔ	æ	ɑ	ð	ɛ	ɪ	ŋ	ʃ	θ	ʊ	ʒ	aɪ	a ʊ	b	d	dʒ	e	f	g	h	i	l	m	n	o	ɔɪ	p	s	t	tʃ	u	v	w	z
Vocalic	+	-	+	+	+	-	+	+	-	-	-	+	-	+	+	-	-	-	+	-	-	-	+	-	-	-	+	+	-	-	-	-	+	-	+	-
Vowel	+	-	+	+	+	-	+	+	-	-	-	+	-	+	+	-	-	-	+	-	-	-	+	-	-	-	+	+	-	-	-	-	+	-	-	-
Sonorant	+	+	+	+	+	-	+	+	+	-	-	+	-	+	+	-	-	-	+	-	-	-	+	+	+	+	+	+	-	-	-	-	+	-	+	-
Obstruent	-	-	-	-	-	+	-	-	+	+	+	-	+	-	-	+	+	+	-	+	+	+	-	-	+	+	-	-	+	+	+	+	-	+	-	+
Continuant	+	+	+	+	+	+	+	+	-	+	+	+	+	+	+	-	-	+	+	+	-	+	+	+	-	-	+	+	-	+	-	+	+	+	+	+
Affricate	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-
Nasal	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	+	-	-	-	-	-	-	-	-	-	-
Approximant	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	+	-
Front1	-	0	-	+	-	0	+	-	0	0	0	-	0	+	+	0	0	0	+	0	0	0	+	0	0	0	-	-	0	0	0	0	-	0	0	0
Front2	-	0	-	+	-	0	+	-	0	0	0	-	0	-	-	0	0	0	+	0	0	0	+	0	0	0	-	+	0	0	0	0	-	0	0	0
Midfront1	-	0	-	-	-	0	-	+	0	0	0	-	0	-	-	0	0	0	-	0	0	0	-	0	0	0	-	-	0	0	0	0	-	0	0	0
Midfront2	-	0	-	-	-	0	-	+	0	0	0	-	0	+	-	0	0	0	-	0	0	0	-	0	0	0	-	-	0	0	0	0	-	0	0	0
Mid1	+	0	-	-	-	0	-	-	0	0	0	-	0	-	-	0	0	0	-	0	0	0	-	0	0	0	-	-	0	0	0	0	-	0	0	0
Mid2	+	0	-	-	-	0	-	-	0	0	0	-	0	-	-	0	0	0	-	0	0	0	-	0	0	0	-	-	0	0	0	0	-	0	0	0
Back1	-	0	+	-	+	0	-	-	0	0	0	+	0	-	-	0	0	0	-	0	0	0	-	0	0	0	+	+	0	0	0	0	+	0	0	0
Back2	-	0	+	-	+	0	-	-	0	0	0	+	0	-	+	0	0	0	-	0	0	0	-	0	0	0	+	-	0	0	0	0	+	0	0	0
High1	-	0	-	-	-	0	-	-	0	0	0	-	0	-	-	0	0	0	-	0	0	0	+	0	0	0	-	-	0	0	0	0	+	0	0	0
High2	-	0	-	-	-	0	-	-	0	0	0	-	0	-	-	0	0	0	+	0	0	0	+	0	0	0	-	+	0	0	0	0	+	0	0	0
Midhigh1	-	0	-	-	-	0	-	+	0	0	0	+	0	-	-	0	0	0	+	0	0	0	-	0	0	0	+	+	0	0	0	0	-	0	0	0
Midhigh2	-	0	-	-	-	0	-	+	0	0	0	+	0	+	+	0	0	0	-	0	0	0	-	0	0	0	+	-	0	0	0	0	-	0	0	0
Midlow1	+	0	+	-	-	0	+	-	0	0	0	-	0	-	-	0	0	0	-	0	0	0	-	0	0	0	-	-	0	0	0	0	-	0	0	0
Midlow2	+	0	+	-	-	0	+	-	0	0	0	-	0	-	-	0	0	0	-	0	0	0	-	0	0	0	-	-	0	0	0	0	-	0	0	0
Low1	-	0	-	+	+	0	-	-	0	0	0	-	0	+	+	0	0	0	-	0	0	0	-	0	0	0	-	-	0	0	0	0	-	0	0	0
Low2	-	0	-	+	+	0	-	-	0	0	0	-	0	-	-	0	0	0	-	0	0	0	-	0	0	0	-	-	0	0	0	0	-	0	0	0

Phone	ə	r	ɔ	æ	ɑ	ð	ɛ	ɪ	ŋ	ʃ	θ	ʊ	ʒ	aɪ	a ʊ	b	d	dʒ	e	f	g	h	i	l	m	n	o	ɔɪ	p	s	t	tʃ	u	v	w	z
Bilabial	0	-	0	0	0	-	0	0	-	-	-	0	-	0	0	+	-	-	0	-	-	-	0	-	+	-	0	0	+	-	-	-	0	-	+	-
Labiodental	0	-	0	0	0	-	0	0	-	-	-	0	-	0	0	-	-	-	0	+	-	-	0	-	-	-	0	0	-	-	-	-	0	+	-	-
Dental	0	-	0	0	0	+	0	0	-	-	+	0	-	0	0	-	-	-	0	-	-	-	0	-	-	-	0	0	-	-	-	-	0	-	-	-
Alveolar	0	+	0	0	0	-	0	0	-	-	-	0	-	0	0	-	+	-	0	-	-	-	0	+	-	+	0	0	-	+	+	-	0	-	-	+
Postalveolar	0	+	0	0	0	-	0	0	-	+	-	0	+	0	0	-	-	+	0	-	-	-	0	-	-	-	0	0	-	-	-	+	0	-	-	-
Retroflex	-	+	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Palatal	0	-	0	0	0	-	0	0	-	-	-	0	-	0	0	-	-	-	0	-	-	-	0	-	-	-	0	0	-	-	-	-	0	-	-	-
Velar	0	-	0	0	0	-	0	0	+	-	-	0	-	0	0	-	-	-	0	-	+	-	0	-	-	-	0	0	-	-	-	-	0	-	+	-
Uvular	0	-	0	0	0	-	0	0	-	-	-	0	-	0	0	-	-	-	0	-	-	-	0	-	-	-	0	0	-	-	-	-	0	-	-	-
Glottal	0	-	0	0	0	-	0	0	-	-	-	0	-	0	0	-	-	-	0	-	-	+	0	-	-	-	0	0	-	-	-	-	0	-	-	-
Aspirated	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	+	-	+	-	-	-	-	-
Closure	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Lateral	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	+	-	-	-	-	-	-	-	-	-	-	-	-
Voiced	+	+	+	+	+	+	+	+	+	-	-	+	+	+	+	+	+	+	+	-	+	-	+	+	+	+	+	+	-	-	-	-	+	+	+	+
Round1	-	0	+	-	-	0	-	-	0	0	0	+	0	-	-	0	0	0	-	0	0	0	-	0	0	0	+	+	0	0	0	0	+	0	+	0
Round2	-	0	+	-	-	0	-	-	0	0	0	+	0	-	+	0	0	0	-	0	0	0	-	0	0	0	+	-	0	0	0	0	+	0	+	0

5.4.2. Stress

We sought to provide the neural network with various stress, accent and prominence information in Block 4. This block contains quantitative answers to the questions shown in Table 5-1.

Table 5-1: Neural network block 4 input

What is the stress of syllable containing this phone: 1 if primary stressed, .5 if secondary stressed, 0 if unstressed.
What is the prominence (as described in O'Shaughnessy 1976) of the word containing the phone: 0-14.
Is the word containing the phone a function (closed class) word (0) or content (open class) word (1).
Does the syllable containing the phone carry a pitch accent (e.g. H*, L* in the sense of Pierrehumbert 1980)

5.4.3. Syntactic and prosodic information

Block 5 contains binary (yes/no) answers to the questions dealing with adjacency to various prosodic and syntactic boundaries as shown in Table 5-1.

Table 5-1: Neural network block 5 input

Does the phone have a (prosodic) syllable boundary on its left/right?
Does the phone have a (prosodic) word boundary on its left/right?
Does the phone have a (syntactic) phrase boundary on its left/right?
Does the phone have a (syntactic) clause boundary on its left/right?
Does the phone have a (syntactic) sentence or (prosodic) utterance boundary on its left/right?[24]
Does the phone have a prosodic intermediate phrase boundary on its left/right?
Does the phone have a prosodic intonation phrase boundary on its left/right?

5.4.4. Windowing

Machine learning procedures might tend to give weight to the position of a phone in a word while learning letter-phoneme correspondences. While there are certain boundary, or edge, effects in English and other languages, absolute position (*e.g.* the third phone in a word) tends not to be important. In order to ensure that neural networks learn the importance of context without undue attribution of effects to absolute position, many researchers have proposed moving data through *sliding windows* composed of several phonemes or letters (in letter-to-sound conversion). Windows of length 7 (Sejnowski and Rosenberg 1987, McCulloch et al. 1987), 9 (Lucassen and Mercer 1984) and 10 (Cohen 1997) have been used. One of the motivations for such long windows, at least in

letter-to-sound conversion, has been the observation of long-distance dependencies for determination of stress placement (*e.g.* Church 1986). However, Pierrehumbert and Frisch (1997, 22), in a study on synthesizing allophonic glottalization, describe the importance of using a running window for postlexical prediction:

> In a faithful TTS implementation of the glottalization rule, the phonological structure will have to be parsed up to the intonation phrase, and the rule will have to have access to both the immediate segmental context and the prosodic strength of the surrounding syllables. This information is available, for example, if the system is implemented with a running window on fully parsed phonological structures, as in Pierrehumbert and Beckman (1988).

Since much of our data is based on isolated sentences, it is important to introduce the notion of padding. We padded each isolated sentence with enough "blank" phones so that assimilatory effects could not be learned across sentences which were contiguous simply by virtue of the training procedure.

We experimented with windows of length 3, 5 and 9. The overall results for three networks are shown in Table 5-1. These results indicate that the largest window size attempted, length 9, was the most successful at learning postlexical processes.

---

[24] Since all the materials used in the current speaker are from read speech, primarily lists of sentences as described in section 3.2, syntactic sentence boundaries largely coincide with prosodic utterance boundaries (Nespor and Vogel 1986).

Table 5-1: Relative performance of postlexical neural network with three different window sizes

Window size	% matching vectors
3	61.2
5	92.9
9	94.9

Note: % matching vectors is an internal measure of success that is correlated with the number of phones correct. It is greater than that number due to the fact that it includes matches on padding vectors.[25]

At first, it is not entirely clear why a window of 9, that is 4 phones in each direction, achieved the best results. It would seem that most postlexical processes are sensitive to neighboring phones alone, although findings such as that of Pierrehumbert and Frisch (1997) above show that the situation is more complex. In addition, Magen (1997) finds vowel-to-vowel coarticulation effects of V1 on V3 in /bV1bəV3b/ contexts— a distance of three phones, across a foot boundary.

Since the best results were achieved with a window of 9 phones, analyses will be based on such a network. All of the sources of information in blocks 2-5 are windowed in this way, and then passed through three hidden neural network layers, blocks 6- 11, and then block 1 emits one of 117 possible postlexical phones.

[25] Padding vectors would have been factored out of the success measure if there had not been errors on them as well in the networks with smaller window sizes.

Chapter 6. Results

In this chapter, we will discuss the results of using the postlexical neural network described in Chapter 5. We will first examine the weights of the neural network after it has been trained in an effort to understand the extent to which it uses the various inputs described above. We will then examine the range of processes learned by the network, including allophony, allomorphy and dialect smoothing. Finally, we will assess the postlexical neural network's performance at learning these processes, comparing our results with analogous results from other research.

6.1. Analysis of neural network

By examining the neural network's distribution of weights after training, we can get an idea of the relative importance of different input information. Figure 6-1 shows the source processing element (PE) weights summed over all phonemes for each of the 9 TDNN window locations in block 6, which receives input from Stream 2. The current phone, which is in the middle of the window (position 5), has the most influence on the network's output, represented by the largest absolute value of the sum of the weights. As might be expected, the influence of the other phones in the window decreases with distance from the current phone. Interestingly, the following phone appears to have more influence than the preceding phone. Giegerich (1992, 214) notes that anticipatory assimilation is much more productive than perseverative assimilation in English. Hock

(1986, 63) makes the same observation across languages. We will also note the importance of following environment in our discussion of vowel fronting below.

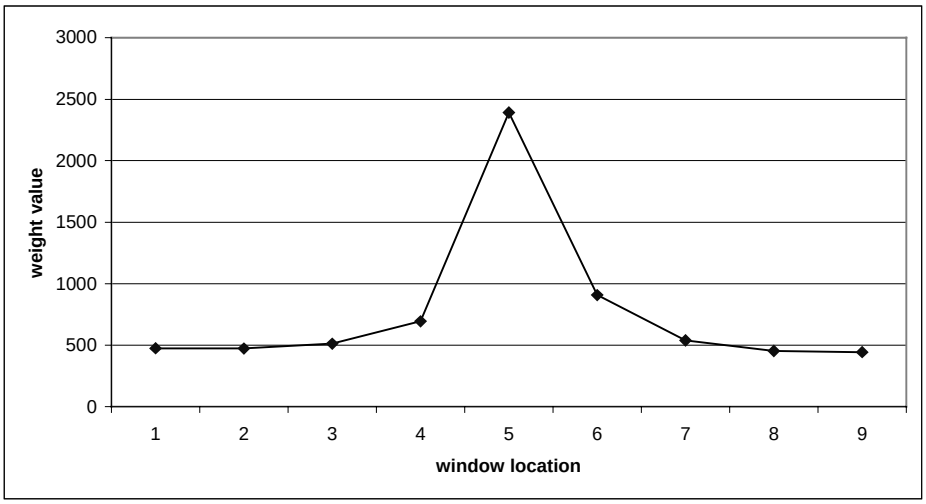

Figure 6-1: TDNN window weights for phone label stream

Figure 6-2 shows the weight allocation in block 10, which receives input from blocks 6, 7, 8 and 9, which in turn receive input from the various data streams. Each of block 6, 7, 8 and 9 supplies 10 PE's to block 10 which represent the values related to the input lexical phones, features, stress and boundary information. The 10 inputs from each of the 4 categories are summed and these 4 sums are displayed in the figure. As shown in the figure, the contribution of lexical phones and lexical features appears to be greater than the stress and boundary related inputs in streams 4 and 5.

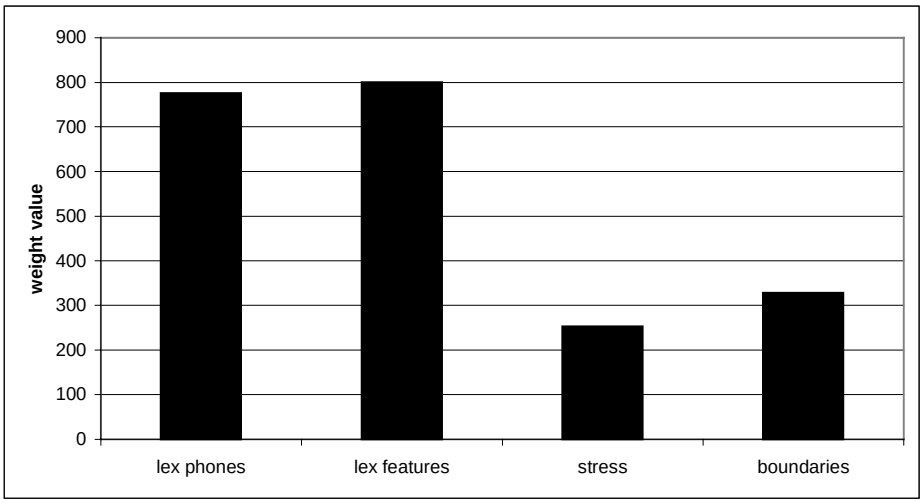

Figure 6-2: Weight distribution by input type

Note that these methods of analyzing the impact of various predictor variables have analogues in other phonological learning and analysis techniques. For example, Randolph (1989, 113) presents statistics on a *percent information extracted* measure for various phonological predictor variables based on learning English stop realization in several corpora using CART (Classification and Regression Trees, Breiman et al. 1993.). Guy and Boberg (1997, 159) present the factor weights for different features of preceding consonants in predicting *t,d* deletion using VARBRUL (e.g. Sankoff 1988).

6.2. General phonological analysis

We will first summarize the breadth of processes that have been learned by the postlexical neural network. Subsequently, we will examine particular processes in detail

with attention to the neural network's performance at learning each process. Table 6-1 illustrates several of the allophonic and allomorphic phenomena learned by the postlexical neural network. These phenomena are all well-known postlexical phenomena of English. Many of them occur variably in the speech of individual speakers along a stylistic continuum.

Table 6-1: Postlexical phenomena learned by postlexical neural network

Phenomenon	Lexical	Postlexical	Orthography
unreleased stops	fɛd	fɛd˺	fed
glottalized vowels	ænd	ʔænd	and
glottalized consonants	stret	streʔ	straight
d deletion	ænd fɑlo	æn fɑlo	and follow
t deletion	əbrʌpt stɑrt	əbrʌp˺ stɑrt˺	abrupt start
destressing and assimilation	ði tæŋk	ðɨ tæŋk	the tank
destressing and assimilation	ði waʲndɪŋ	ðə waʲndɪŋ	the winding
t flapping	dəˠti	dəˠɾi	dirty
nasal flapping	kornəˠ	kɔrr̃əˠ	corner
h voicing	ɪn həˠ	ɨn ɦəˠ	in her
schwa epenthesis	kəˠlz	kəˠl̩z	curls
/u/ fronting	dun	dʉn	dune
syllabic consonants	pudəlz	pʉdl̩z	poodles

Table 6-2 illustrates another class of phenomena learned by the postlexical neural network. These include dialect smoothing, and smoothing of what might well be considered idiosyncrasies (or conventions) of labeling either the lexical database or the labeled speech corpus. These phenomena learned by the postlexical network would

probably not be considered postlexical in a description of natural language. These phenomena are the results of the neural network learning some of the idiosyncrasies of the particular dictionary-labeler-speaker triad at hand. For example, neutralization of vowels before /r/ may not be a postlexical process in the speaker's actual grammar, since his mental lexicon probably stores neutralized versions of these vowels. However, when coupled with a generalized American English pronunciation lexicon, as described above, it is necessary to perform this neutralization actively, which is what the postlexical network does.

Table 6-2: Dialect/labeling/lexical idiosyncrasies learned by network

Phenomenon	Lexical	Postlexical	Orthography
marry/merry/Mary merger	bærəl	bɛr̩l	barrel
schwa deletion/metathesis?	čɪldrənz	čɪldərnz	children's
laxing before /r/	kornə˞	kɔr̃ə˞	corner
wh voicing	ʍaʲlst	waʲlst	whilst

6.3. General error analysis

Analyzing performance on multiple postlexical phenomena at one time can be a daunting task. Tuning the data encoding to improve performance on one process can result in damaging performance on another. Nevertheless, we need some way of assessing the performance of a given architecture of the postlexical neural network. In this way, the architecture can be adjusted and improvements or degradations noted. If particular architectural or data changes result in improved performance, they can be preserved,

otherwise, they can be discarded. The kinds of changes have been found to affect network performance include providing new data, such as prosodic boundaries, or changing the way existing data is used, for example, by changing the TDNN window size.

Perfect performance would be for the postlexical network to predict, based on lexical phones, the postlexical phones that were labeled in the labeled corpus. To be fair, we test for this on the 10% of testing data that was withheld from training. Perfect performance so defined is difficult, if not impossible, to attain.

Table 6-1 summarizes neural network performance on the testing subset given the network architecture described in Chapter 5. As can be seen, 87.8% of the predicted phones matched the phones in the original hand-labeled files. Since this result means that more than 1 in 10 predicted phones will differ on average from the original files, it is important to understand the nature of the incorrect phones, in order to assess the potential perceptual impact of this error rate. We will show how 84% of the errors, or 10% of the total phones, are actually acceptable variants of the original phones, increasing the overall success rate to 98% acceptable phones.

For example, 8.6% of the predicted phones were another allophone of the original phone. Table 6-2 provides details on these kinds of errors. As can be seen in the table, most of these errors are fairly subtle, such as the choice of a different reduced vowel allophone from the one actually selected by the speaker. We will discuss the conditioning of such choices in the data below.

Table 6-1: Summary of postlexical network results

Result	instances	percentage
original phone = nn phone	2303	87.8%
nn phone = another allophone of original phone	225	8.6%
nn phone = another allomorph of original phone	37	1.4%
nn phone = another dialect's version of original phone	6	0.002%
total acceptable phones	2571	98%
total unacceptable phones	52	2%

Table 6-2: Allophonic errors

neural network	original	quantity	neural network	original	quantity
ə	ɨ	27	ɾ̃	n	2
ɨ	ə	12	kc	kc k^h	2
ɾ	dc d	9	ɾ	tc t^h	2
ɨ	ɪ	9	tc t^h	ɾ	2
ɪ	ɨ	8	pc p^h	p^h	2
ɜ˞	ə˞	8	pc p	pc	2
ʔ	tc t^h	7	ʔɛ	ɛ	2
dc	dc d	7	dc d	ɾ	2
h	ɦ	6	ɔ	ɑ	2
tc t^h	t^h	6	ə	ʌ	2
u	ʉ	5	ə	ʔʌ	2
ʉ	u	5	tc	ʔ	1
tc	tc t	5	l̩	l	1
tc t^h	ʔ	5	k^h	kc	1
gc g	gc	5	i	ʔi	1
dc d	d	5	ɾ	ʔ	1
ə˞	ɜ˞	5	ʌ	ʔʌ	1
n	ɾ̃	4	tc t	tc t^h	1
bc b	b	4	pc p^h	pc	1
ʔ	tc	3	kc k^h	kc k	1
pc	pc p	3	dʒc dʒ	dʒ	1
kc	kc k	3	ʔɨ	ɨ	1
ɦ	h	3	ʔɪ	ɪ	1
b	bc b	3	ʔɪ	ʔɨ	1
tc t	tc	3	ɛ	ʔɛ	1
kc k	kc	3	dc d	dc	1
ɑ	ʔɑ	3	tʃc tʃ	tʃ	1
ə	ʔə	3	bc b	bc	1
dc	deleted	3	ʔæ	æ	1
t^h	tc t^h	2	æ	ʔæ	1
tc	tc t^h	2	ə	ɪ	1
pc	pc p^h	2	tc	deleted	1
ɾ̃	n	2	ʔ	deleted	1
kc	kc k^h	2	s	deleted	1
ɾ	tc t^h	2	tc t	deleted	1
tc t^h	ɾ	2	ɪ	ʔɪ	1
pc p^h	p^h	2			

Another source of acceptable errors involves mismatches in optional vowel reduction. 1.3% of the incorrect phones represented choice of a either a reduced vowel by the neural network where the original speech had an unreduced vowel or vice versa. These errors also include a different choice of allomorph for *the*. For example, *the* can surface as [ðə], [ði] or [ðɨ]. Function words are subject to certain destressing patterns (e.g. Selkirk 1984) which are not found in content words, at least in the formal style of speech analyzed here. While the kinds of vowel reductions generally found in function words may not be subject to the clear segmental conditioning that governs *the* allomorphy, we might say that they are prosodically conditioned. Finally, particular morphemes such as *re-* and *be-* are subject to destressing patterns similar to those undergone by function words.

Table 6-3: Destressing errors

neural network	original	quantity	examples
e	ə	4	a
ɨ	æ	4	had, than, and
æ	ɨ	4	at, than, and
i	ɨ	3	resigned, the ink, the arrow
ɨ	e	3	a
u	ɨ	2	to
e	ɨ	2	a
ɨ	ʉ	2	to
r	deleted	2	from
v	deleted	1	of
ʉ	ɨ	1	to
aɪ	ɑ	1	I'll
ɨ	i	1	resumed
ɨ	ɛ	1	then
ɨ	ɑ	1	on
ɛ	ɨ	1	when
ɑ	ɨ	1	on
ə	e	1	a
ɚ	r	1	are
ə	deleted	1	from

A final class of acceptable errors is minor dialect differences between the neural network's choice and the original speech. Since the neural network bases its hypotheses on lexical forms from a dictionary of Generalized American, as described in section 3.1, such errors are not unexpected. We will discuss the success of dialect smoothing below; however, Table 6-4 shows some cases that were not successfully smoothed.

Table 6-4: Dialect errors

neural network	original	quantity	examples
ʃ	tʃ	1	lectured
ʃ	tʃc tʃ	1	attention
dʒc dʒ	ʒ	1	changed
ɪ	ɛ	1	England
gc g	deleted	1	England
ʔæ	ʔɛ	1	arrow
deleted	ə˞	1	expiration

While we hope to have shown that the majority of mismatched phones between the neural network and original speech is acceptable, this notion is actually fairly complex. Traditional phonemic theory differentiates between phonemes and allophones by saying that substituting one phoneme for another would result in a semantic difference, whereas substituting the "wrong" allophone might merely sound odd. Sociolinguistic studies of variation have shown that allophony is governed by a complex constellation of linguistic and social factors. Given this, an allophonic choice might be stylistically, socially or phonologically odd. Determining the oddness or acceptability of particular allophonic choices would require psycholinguistic analysis that we hope to pursue in future work.

In addition, in the future we hope to investigate methods for postprocessing or constraining the output of the postlexical neural network. While at present we are using the top-ranked phone for each slot, the neural network can provide multiple phone hypotheses and likelihoods for each slot. We can use n-gram statistics on English

postlexical phones in order to traverse such networks, in effect realizing phonotactic constraints on the neural network output. Such a procedure might help eliminate a portion of the unacceptable errors that remain.

6.4. Allophony

In this section, we will provide a detailed analysis of the neural network's performance on several allophonic phenomena. First, we will discuss two phenomena involving vowels: fronting and glottalization. We will then investigate allophony in coronal consonants, including deletion, flapping, glottalization and aspiration. For each phenomenon, we will attempt to provide a post-hoc analysis of the factors that appear to govern the network's phone choices. Where possible, we will provide details about other researchers' success at describing or modeling these phenomena.

6.4.1. Vowel fronting

We will first examine the two cases of vowel allophony involving fronting that were examined in acoustic space in Chapter 4. There we discussed the phoneme /u/, whose main allophonic reflexes are [u] and [ʉ], and /ə/, whose main reflexes are [ə] and [ɨ]. The environment we will be examining to explain the allophonic variation involves the feature coronal. In this regard, the work of Hume and Clements (e.g. Hume 1992, Clements and Hume 1995) demonstrating that front vowels possess the coronal feature is particularly helpful. Indeed, we have found that the presence of coronal segments in the

environment of /u/ and /ə/ usually favors the fronted allophones [ʉ] and [ɨ] over their unfronted counterparts [u] and [ə]. In accordance with work by Clements (1976), Zsiga (1995, 294-295) and others, we have found that [j] patterns with the coronals.

In addition, we have found that the dorsal component of syllable-final, dark or velarized /l/ tends to block fronting, despite its coronal component (cf. McLemore 1995, van Bergem 1994). According to Sproat and Fujimura (1993), /l/ involves a vocalic dorsal gesture and a consonantal apical gesture. In syllable-final position, the dorsal gesture is closer to the syllable nucleus than the apical gesture, while the reverse holds in syllable-initial position (Sproat and Fujimura 1993, 291). This accords with our analysis, in which we consider syllable-initial /l/ coronal, and syllable-final /l/ noncoronal with respect to environments surrounding vowel nuclei. As Sproat and Fujimura point out, intervocalic /l/, as in words like *foolish* or *frolic,* occupies an intermediate position articulatorily between syllable-initial and syllable-final /l/.

For the analyses below, we will consider [+coronal] to mean all traditional coronals with the addition of /j/ and excluding /l/'s that are unambiguously in the coda of their syllables (i.e. intervocalic /l/'s will be considered [+coronal]). [-coronal] will specify the complementary environment. We will be examining both left and right environments, ignoring the presence of word boundaries. We have excluded cases where either

environment consisted of a sentence boundary, as the speech database consists of isolated sentences.

We will first examine the patterning of the allophonic reflexes of /ə/. Figure 6-1 shows the realization of schwa allophones in the test data in both the original labeled speech and the output of the postlexical neural network. The unfronted [ə] allophone appears to be most favored when noncoronal environments are on the left and right, and secondarily with a noncoronal environment on the right alone. The fronted [ɨ] allophone appears to be most favored when coronal environments are on the left and right, and secondarily with a coronal environment on the right alone.

It may be interesting to examine the three exceptions to fronting in the environment surrounded by coronals in the original speech. In two of the exceptions, *pant**a**loons* and *subst**i**tutes* (underlying schwa in bold), the following coronal environment is not tautosyllabic with the vowel in question. As we suggested earlier with respect to McLemore's (1995) rule for predicting [ɪ]-like reduced vowels, tautosyllabicity with respect to surrounding coronals may be important. The remaining exception is in the word *jeal**ou**s*. The exceptionality of this word highlights a potential problem with our decision to consider intervocalic /l/ as coronal.

The neural network appears to have performed similarly in the cases of surrounding coronals or noncoronals, in addition to the following coronal alone case. It appears to

have performed less well in the case of preceding coronal alone case. It is possible that the neural network is relying on additional information beyond the coronality of the immediate environment.

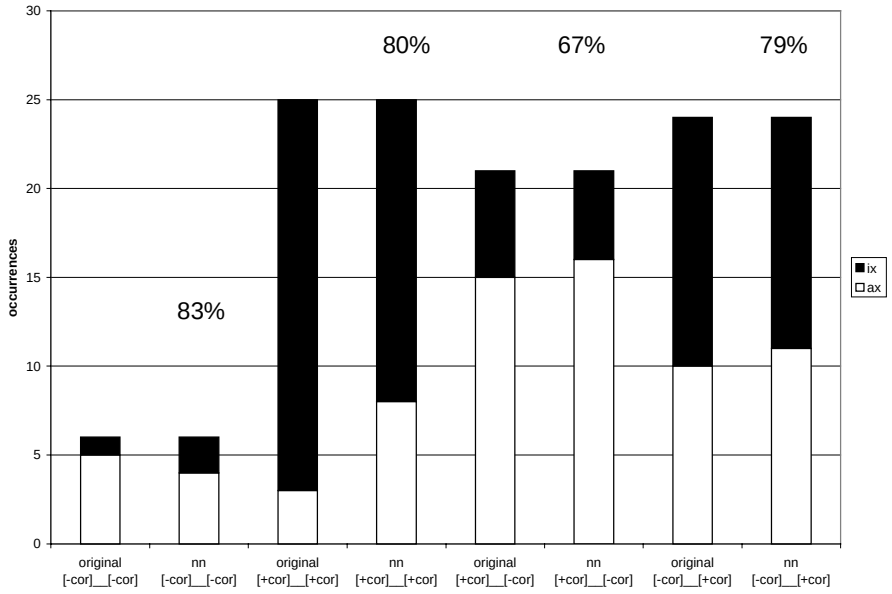

Figure 6-1: Distribution of schwa allophones

Each environment is represented by one bar representing original speech and another bar representing neural network (nn) output. Within each bar, the proportion of occurrences of each allophone is represented. Atop the bar representing the neural network output is placed a percentage reflecting the accuracy of the neural network's predictions with respect to the original data. The figure uses TIMIT labels.

We will now discuss other researchers' findings with respect to the varied realization of reduced vowels in English. From a theoretical standpoint, Jensen (1993, 212-213) proposes to redress what he claims to be an inadequate treatment of variation in reduced vowel qualities in SPE. SPE recognizes, as we do, one underlyingly reduced vowel, /ə/,

which may be realized with different qualities depending on low-level phonetic rules (SPE, 110). Jensen (1993, 213) proposes a rule of vowel reduction and assimilation by which reduced vowels before palatals become [ɪ]. He also proposes rules by which nonback nonlow underlying vowels reduce to [ɪ], while all others reduce to [ə]. Such rules rely on rather more abstract underlying representations than we have found convenient, as discussed in 3.1.

Byrd (1994), investigating the TIMIT database, found that [ɨ] was more than twice as common as [ə], across speakers. However, she found that women used [ə] less than men, and that speakers from the North, New York City and the West used [ə] less frequently, while speakers from the North Midland, South and South Midland used [ə] more frequently. Men used [ɨ] less frequently than would be predicted by random distribution. Speakers from the Northeast, New York City and the West used [ɨ] more frequently than expected.

Browman and Goldstein (1992) discuss the concept of a "targetless" schwa. Such a schwa takes on the characteristics of surrounding phonetic material, in contrast with views holding that schwa's target is in the center of the vowel space, with reduced vowels shifting towards it (e.g. SPE, 110). They cite evidence that schwa coloring is a function of neighboring stressed vowels in an examination of X-ray data of articulations by an American English speaker of utterances within the frame $pV^1pə'pV^2pə$, where V^1 and V^2

vary over /i, ɛ, a, ʌ, u/. Browman and Goldstein's (1992) findings are not easily comparable to our results, since they examined the influence of surrounding vowels on schwa realization, while we examined the influence of surrounding consonants.

Van Bergem (1994) studied the effects of both neighboring consonants and vowels on the realization of schwa in Dutch nonsense words, using the frames $C^1əC^2V$ and $VC^1əC^2$, where C^1 and C^2 varied over /p, t, k, f, s, χ, m, n, ŋ, r, l, j, ʋ/ and V varied over /i, aː, u/. As did Browman and Goldstein (1992), van Bergem (1994) noted that schwa lacks an articulatory target, underlining instead the importance of phonemic context for determining its acoustic realization.

Van Bergem (1994, 160) discusses schwa's tendency to strive for maximum "compatibility" with the surrounding phonemic context. He notes that the straightness of vowel F2 tracks between consonants is a reasonable measure of such compatibility. For example, /ɪ/ is compatible with the dentals /t/ and /n/ in the sense that they are both pronounced with a "fronting" tongue. Van Bergem (1994) demonstrates this spectrally by noting that /ɪ/'s F2 track is fairly straight across the transition from /t/ to /n/, in contrast to that of /ɔ/. He suggests that this explains the presence of /ɪ/-like schwa (our [ɨ]) in the environment of dentals (*round**e**d*, *hors**e**s*). In contrast, van Bergem (1994) notes that the F2 track of /ɔ/ is more compatible than /ɪ/ with the environment /ʋ_l/.

Consequently, he suggests this as an explanation for an /ɔ/-like schwa (our [ə]) in words like *vow**e**l* and *gru**e**l*.

In an acoustic analysis of three white Chicago speakers, Veatch (1991, 163) found that the vowel reduction target was high central [ɨ]. In accordance with Byrd's and Veatch's findings, [ɨ] is the most common reduced vowel for our speaker. In accordance with McLemore's (1995) hypothesis, coronal environment on the left and right is the most favorable environment for realization of [ɨ]. From the data in Figure 6-1, it appears that following [+coronal] environment is more favorable than preceding [+coronal]. The greater effect of following position on schwa realization was also found by Browman and Goldstein (1992, 56), and van Bergem (1994), though it is only described with respect to following vowels.

We will now discuss our findings with respect to the distribution of the fronted [ʉ] and unfronted [u] allophones of /u/ in the original speech and as predicted by the postlexical neural network, as shown in Figure 6-2. We have employed the same set of preceding and following environment classifications as we did with schwa. As we can see in the figure, the fronted variant is predicted most often in contexts with coronals on both the left and right. The preceding coronal environment alone is the next best predictor for fronted [ʉ].

Interestingly, following coronal does not appear to predict the fronted variant, although the data is scant in that environment. Recall that with respect to schwa, following coronal actually predicted the fronted [ɨ] more than preceding coronal. It is not clear why the effects of coronality with respect to direction should be reversed between these two cases. The postlexical neural network appears to perform best in the case where /u/ is surrounded by coronals. It does not appear to have noted that following coronal does not predict fronting in the original data.

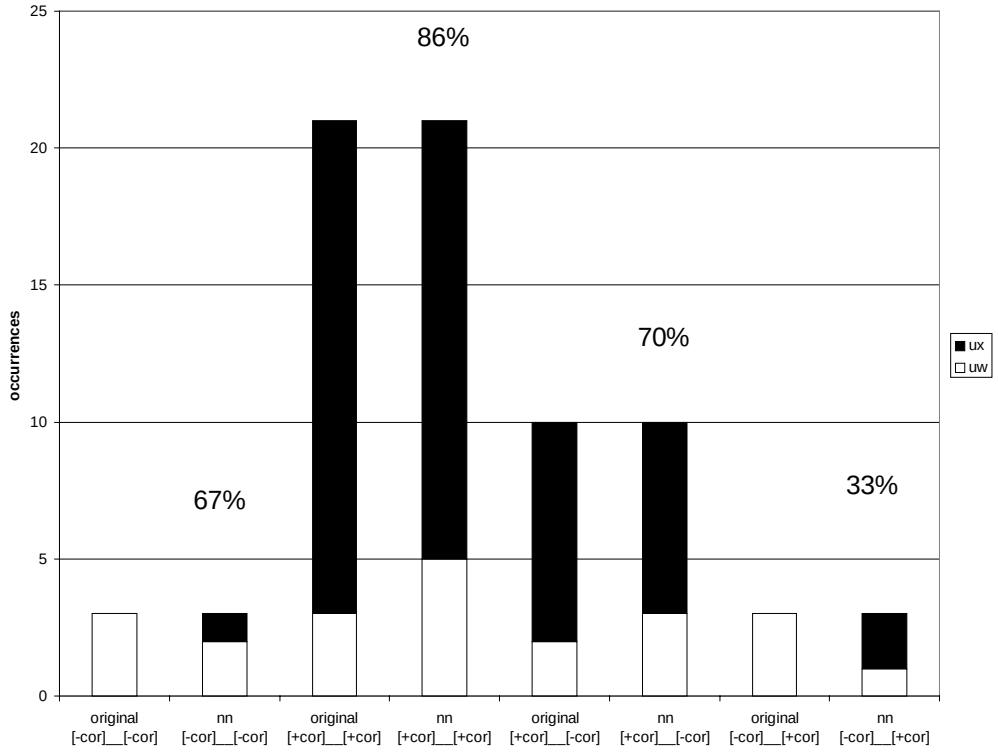

Figure 6-2: Distribution of /u/ allophones

Each environment is represented by one bar representing original speech and another bar representing neural network (nn) output. Within each bar, the proportion of occurrences of each allophone is represented. Atop the bar representing the neural network output is placed a percentage reflecting the accuracy of the neural network's predictions with respect to the original data. The figure uses TIMIT labels.

6.4.2. Glottalization of vowels

Dilley et al. (1996) examined several predictor factors for word-initial vowels surfacing with glottal onsets: whether the vowel occurs at the start of an intermediate or intonational phrase, whether the target syllable or word was pitch accented, whether the vowel was reduced, and preceding context. Of 125 word-initial vowels in the testing

portion of the data, 21% were glottalized in the original speech, while 14% were glottalized by the postlexical neural network. Among the vowels glottalized by the postlexical neural network, 72% were glottalized in the original speech. Dilley et al. (1996, 432) analyzed 5 speakers, whose glottalization rates ranged from 17% to 45%. Dilley et al. found that all speakers glottalized more when the syllable containing the vowel in question began an intermediate or intonational phrase. This was not true of our speaker. In our original data, the glottalization rate was still 21% at the start of intermediate phrases, and only 14% at the start of intonational phrases.

Both Dilley et al. (1996) and Pierrehumbert and Frisch (1997) found that vowels in accented syllables were more likely to be glottalized than those in unaccented syllables. Our speaker also demonstrated this behavior. 48% of accented word-initial vowels were glottalized in the original speech, while 45% were glottalized by the postlexical neural network. Among the vowels glottalized by the neural network, 79% were glottalized in the original speech.

6.4.3. Coronal allophones

In this section, we will examine postlexical neural network performance on coronals themselves, rather than their effects on other phones, as in the previous section. Coronals such as /t/ and /d/ are subject to a great deal of allophony in American English. In fact, according to the data in Figure 5-1, they have the most entropy of any consonants, and indeed any other phones, except for schwa. This entropy is reflected in the complexity of

phonological accounts of /t/ allophony, which often involve rule-ordering (Nespor and Vogel 1986, Kiparsky 1979). Since the postlexical neural network does not have any concept of rule-ordering, the success that it has at modeling /t/ allophony may indicate that a simpler constraint-based approach is more suitable. The processes involving /t/ that we will examine include aspiration, flapping, and glottalization. The segmental nature of our treatment of allophony in this section will also enable us to study *t,d* deletion.

With only one exception,[26] all syllable-initial /t/'s in the testing portion of the database surface with an aspirated release. The fact that our hand-syllabified database generally adhered to the principle of stress-conditioned resyllabification (see discussion in section 3.2) explains why intervocalic examples of flapping did not appear in syllable-initial position. Figure 6-1 summarizes the distribution of syllable-final /t/ allophones in the testing portion of the data. We have simplified the data by grouping together unreleased and unaspirated /t/ as "unaspirated". As can be seen, the postlexical neural network's general proportions of each of the phenomena appear to be in line with the original data. Glottalization and deletion appear to be harder for the network to learn than flapping and aspiration.

[26] The exception is the word *mountain* [maʊn-ʔn̩], whose hand-syllabification may be subject to debate.

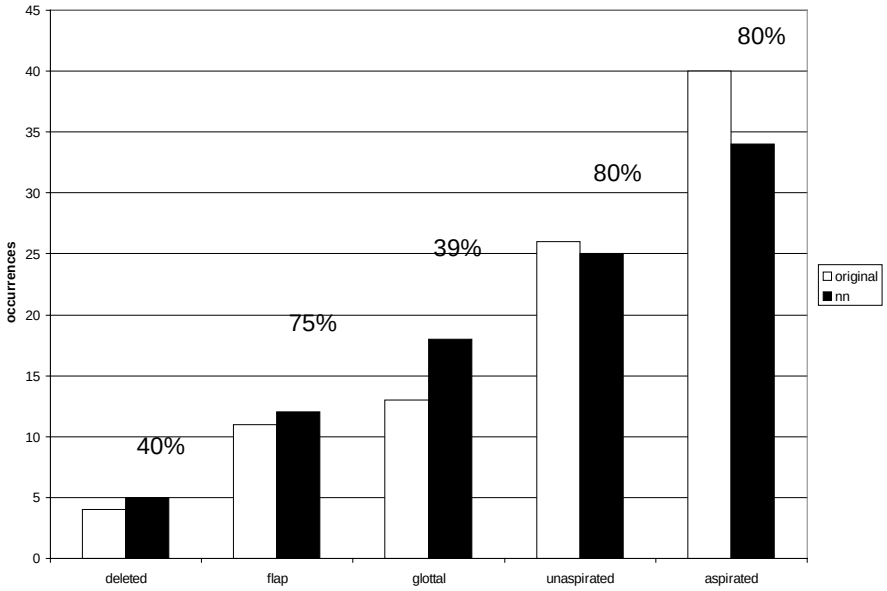

Figure 6-1: Distribution of syllable-final /t/ allophones

Each environment is represented by one bar representing original speech and another bar representing neural network (nn) output. Within each bar, the proportion of occurrences of each allophone is represented. Atop the bar representing the neural network output is placed a percentage reflecting the accuracy of the neural network's predictions with respect to the original data.

Let us try to assess the general constraints on the realization of these allophones. Deletion occurs only when /t/ is the final element of a consonant cluster in a syllable coda. Flapping occurs only in intervocalic environments, except in one case where it precedes [ɦ] across a word boundary in "that heavy". Glottalization occurs primarily before sonorants. Unreleased or unaspirated /t/ occurs before nonsonorants. Aspirated /t/ generally occurs at the ends of intonational phrases. We will look more closely at the factors explaining the variation in these allophones below.

English coronal stop deletion, or *t,d* deletion, has occupied sociolinguists for many years, as it provides a good example of inherent variability, as described by Weinreich, Labov and Herzog (1968). According to Guy and Boberg (1997, 151), "The process is universally variable in English— that is, every speaker that has been observed deletes some, but not all, of their final stops". The particular environment in which *t,d* deletion has been most studied is in tautosyllabic consonant clusters in which the /t/ or /d/ is the final element of the syllable coda. It is important to note that the concept of *t,d* "deletion" is somewhat of an abstraction. It is likely that a more precise analysis might find remnants of the /t/ or /d/ gesture in the remaining consonant (e.g. Browman and Goldstein 1990). Nevertheless, the notion of *t,d* deletion has proved useful in both sociolinguistics and speech technology (e.g. Randolph 1989).

Guy and Boberg propose an analysis whereby *t,d* deletion derives from the OCP (obligatory contour principle), which has been used to ban adjacent identical sequences on an autosegmental tier in languages, whether tones (Leben 1973), segments (McCarthy 1986), or features (Yip 1988). Guy and Boberg characterize /t/ and /d/ as [+coronal, -sonorant, -continuant]. They then characterize the other phones with which /t/ and /d/ might form clusters as in Table 6-1. Following Guy and Boberg, /r/ is marked with a question mark in the feature column. They note that while /r/ has traditionally been considered a coronal, the [ɚ] that might be found preceding /t/ is perhaps best considered noncoronal (Guy and Boberg 1997, 157).

Guy and Boberg observe that the OCP would favor *t,d* deletion after segments that have the closest match to /t/ and /d/ featurally. They note that in their data, there were no statistically significant differences among the segments which matched on two features, that is, /n/, the noncoronal stops and the sibilants. We have added the affricates /tʃ/ and /dʒ/ to the sibilant group, as the part of them that is adjacent to the /t/ or /d/ is sibilant. Guy and Boberg noted less deletion on the segments that match /t/ or /d/ on only one feature.

We examined all tautosyllabic coda consonant clusters ending with /t/ and /d/ in the testing data. Our results, as shown in Table 6-1, match those of Guy and Boberg at the gross level of the two-feature matches exhibiting less deletion than the one-feature matches. Interestingly, our speaker exhibited no deletion at all in the one-feature matches, which may have been a product of the somewhat formal speaking style of our recordings.

Table 6-1: *t,d* deletion by preceding phone

Preceding phone	features	N	% deleted original speech	% deleted neural network
n	[+cor,-cont]	33	42	33
p, b, k, g	[-son,-cont]	9	22	33
s, z, ʃ, ʒ, tʃ, dʒ	[+cor,-son]	26	8	4
f, v	[-son]	4	0	0
l	[+cor]	11	0	0
m, ŋ	[-cont]	2	0	0
r	?	13	0	0

Table 6-2 summarizes postlexical neural network performance on flapping lexical /t/. The correct rule applications were all in the classic flapping environment ý_ў (e.g. Kahn 1976). In contrast, the failures to apply flapping were in two cross-word environments: *tha**t** heavy* and *brough**t** out* (where the /t/ in question is in bold). The incorrect applications of flapping, as far as the original speech was concerned, were as follows: *shor**t** of*, *though**t** of*, and *jani**t**ors* (/t/ in question is in bold). In *short of*, an intonational phrase boundary separated the two words, and the speaker used a glottal stop. In *thought of*, an intermediate phrase boundary separated the two words and the speaker used an aspirated /t/.

Table 6-2: Postlexical neural network performance at flapping lexical /t/

correct application of rule	incorrect application of rule	failure to apply rule
9	3	2

The postlexical neural network's choice to flap in these phrases does not seem inappropriate, especially considering Nespor and Vogel's (1986) demonstration that flapping often occurs across intonational phrases in the phonological utterance, the largest domain in the prosodic hierarchy. In *janitors*, the speaker used a fully released /t/.

Strassel (1997) finds the non-post-tonic environment exemplified by the /t/ in *janitors* to be variable in both the Switchboard and HUB-4 corpora. Interestingly, our speaker is Northern, and this environment was found by Strassel to be flapped categorically by speakers of dialects other than AAVE (African American Vernacular English), Southern and South Midland.

Glottalization in English occurs in two main environments: as an allophone of voiceless stops, as in *button* [bʌʔn̩], and as an onset to vowels as in *ask Emma* [æsk ʔɛmə], which was discussed above. We will now investigate glottalization of /t/. Pierrehumbert and Frisch (1997, 12) found that, contrary to reports indicating that wider following environments were possible, glottalization of /t/ was found before sonorants only. In contrast to their findings with respect to vowel glottalization, they found that whether a syllable was accented did not play a role in /t/ glottalization. Table 6-3 shows the distribution of phones following glottalized /t/'s in both the original speech and the postlexical neural network. In addition to preceding sonorants, as predicted by Pierrehumbert and Frisch, one glottal stop occurred before [f] in the original speech. It is interesting to note that this glottal stop occurs at the end of an intermediate phrase. The presence of several glottal stops at the ends of utterances provides a further clue to the importance of prosodic boundaries to /t/ glottalization.

Table 6-3: /t/ glottalization by following phone

following phone	N	original glottalizations	neural net glottalizations	neural network % correct
utterance final	8	2	6	0
n, m	5	4	2	20
r	2	2	2	100
j, w	4	4	4	100
vowels	4	1	3	25
ɦ	1	0	1	0
f	1	1	0	0

Table 6-4 presents data on /t/ glottalization by prosodic position, that is, whether the /t/ in question occurs at end of an intermediate or intonational phrase, or whether it occurs non-phrase-finally. It appears that glottalization surfaces in all three environments.

Table 6-4: /t/ glottalization by prosodic position

N	prosodic status	original glottalization	neural net glottalizations	neural net % correct
13	non-phrase-final	9	10	46
2	end of intermediate phrase	1	2	50
10	end of intonational phrase	4	6	0

We will now look more closely at the phrase-final environments, to see which /t/ allophones are most likely to surface there. Figure 6-2 displays /t/ allophony in intermediate phrase final position. Note that in our system, all intonational phrase

boundaries are also considered intermediate phrase boundaries. As we can see, deletion and flapping are less likely here, compared with the display of syllable-final /t/ allophony in Figure 6-1. Note that the neural network accuracy percentage refers to the number of correct predictions of an allophone among the predictions of a particular allophone by the neural network. For example, of the 27 aspirated /t/'s predicted by the neural network, 25 were associated with aspirated /t/'s in the original speech.

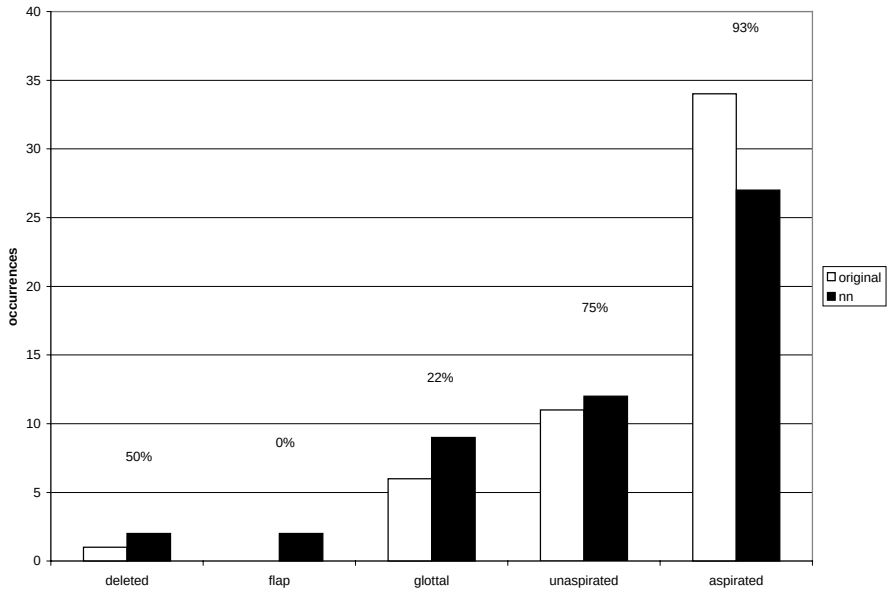

Figure 6-2: /t/ allophones at intermediate phrase ends

Each environment is represented by one bar representing original speech and another bar representing neural network (nn) output. Within each bar, the proportion of occurrences of each allophone is represented. Atop the bar representing the neural network output is placed a percentage reflecting the accuracy of the neural network's predictions with respect to the original data.

Figure 6-3 shows the distribution of /t/ allophones at intonational phrase ends. Note that the preference for aspirated /t/ closures is even greater than at intermediate phrase ends. Byrd (1994, 45) examined sentence-final /t/ releases in "that" in one of the sentences spoken by all speakers in the TIMIT corpus and found the following allophone distribution: 23% released, 67% unreleased, and 9% glottalized. Although Byrd did not find a significant effect of dialect on the release of oral stops in general, it is not clear whether there was such an effect just with respect to /t/.

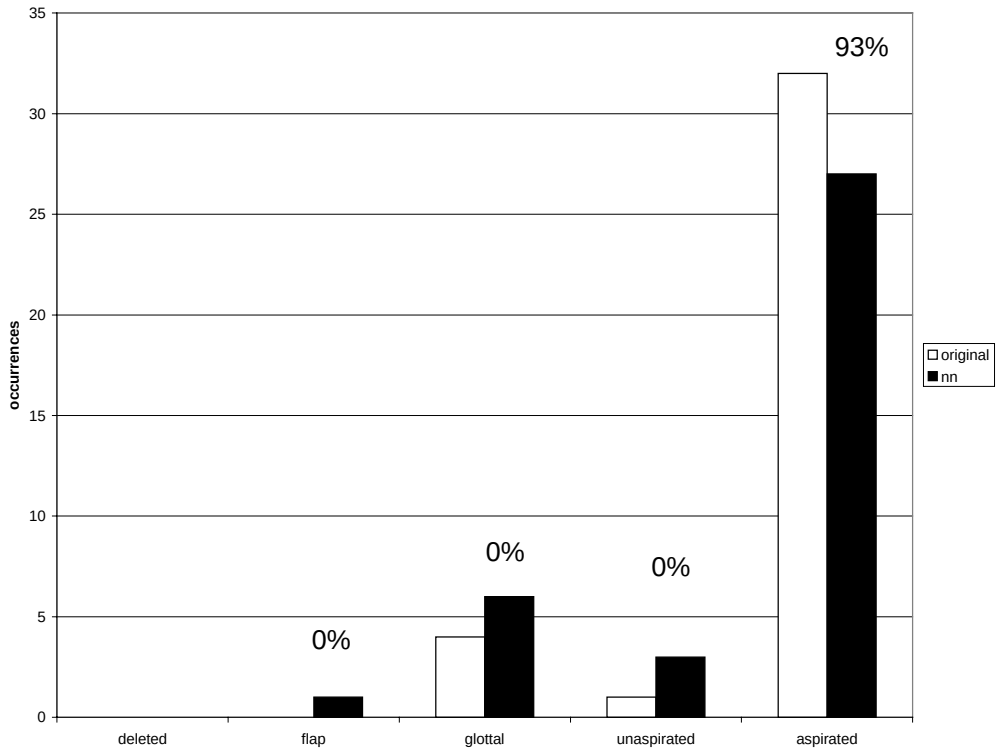

Figure 6-3: /t/ allophones at intonational phrase ends

Each environment is represented by one bar representing original speech and another bar representing neural network (nn) output. Within each bar, the proportion of occurrences of each allophone is represented. Atop the bar representing the neural network output is placed a percentage reflecting the accuracy of the neural network's predictions with respect to the original data.

In order to investigate Pierrehumbert and Frisch's (1997) claim that syllable accentuation does not affect /t/ glottalization, we organized the data in Table 6-5 by whether or not the syllable containing the /t/ in question was accented. While a higher proportion of accented syllables are glottalized, Fisher's exact test showed this not be significant. It is interesting to note that the neural network's performance on glottalization was generally poor, except when examined in the case of preceding /r/, /j/ and /w/, where it was 100%.

Table 6-5: /t/ glottalization by syllable accentuation

N	syllable accented	original glottalization	neural net glottalizations	neural net % correct
9	yes	7	5	33
16	no	7	13	25

6.5. Vowel reduction in function words

In this section, we will analyze various examples of vowel reduction in function words. We are considering these phenomena apart from the general processes of allophony discussed above because they appear to be mainly limited to certain classes of words, namely function words. This limitation is true of postlexical reduction; however, many analyses of English lexical phonology (e.g. SPE, Halle and Mohanan 1985) rely on vowel reduction processes that occur throughout the vocabulary.

We believe that some instances of vowel reduction are best analyzed as phonologically-conditioned allomorphy. We will demonstrate that such an analysis best explains variation in the pronunciation of *the* in our corpus. In addition, we will show how such an analysis may not be appropriate for explaining reductions in other function words, despite their phonological similarity.

Keating et al. (1994) investigate vowel variation in the word *the* in the multi-speaker TIMIT database. They found that [i] was most common before consonants, while [ə] and [ɨ] were most common before vowels. Keating et al. (1994, 136) note that [ə] and [ɨ] are reduced with respect to [i] in duration and quality. They found that some age-grading was involved in *the* realization, and that younger speakers are failing to make the distinction along vowel length grounds, preferring [ə] in all instances, with glottal stop insertion before vowels.

As shown in Figure 6-1, our speaker has maintained the traditional distinction of pronouncing *the* before consonants with a lax vowel (either [ə] or [ɨ]) and before vowels with a tense [i]. The postlexical neural network appears to have learned this distinction well, on the basis of a lexical representation of /ði/.

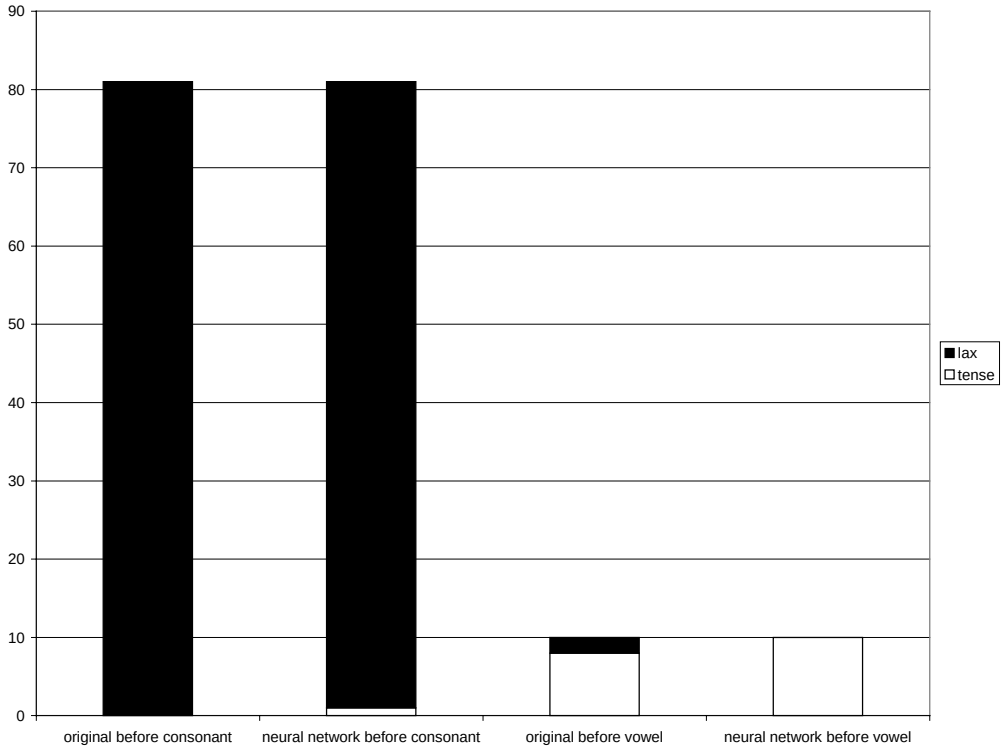

Figure 6-1: *the* allomorphy

Now we do not consider variation in *the* to be a case of allophony, such as those examined in section 6.4, because it involves a process (/i/ → [ə] /__C), that is not general across the vocabulary. Now it might be argued to be part of a general case of vowel reduction in function words. In order to examine this, we will consider other vowel-final function words, to see if the reduction of final vowels is correlated with the following segment in a similar way to *the*.

Table 6-1 examines all of the vowel-final function words that occur in the testing portion of the speech corpus for the possibility that the variation between reduced and unreduced

vowels is correlated with following consonant or vowel as in the case of *the*. In this table, neural network correct percentages refer to correctness with respect to reduction; an exact match of reduced vowel was not required. Unless otherwise specified, reduced vowels are [ə] and [ɨ]. Only two of the words in the table reduce to [ə] and [ɨ] at all, namely *a* and *to*. Since *a* only occurs in this dialect before a consonant, we will have to look for an explanation for its variation between [e] and [ə]/[ɨ] elsewhere. Since the only token of *to* occurring before a vowel surfaced in reduced form, explaining reduction along the lines of *the* does not seem promising.

Table 6-1: Reduced and unreduced vowel-final function words before consonants and vowels

word	N	neural net % correct	preceding consonant		preceding vowel	
			reduced	unreduced	reduced	unreduced
a	19	37	10	9	0	0
be	1	100	0	1	0	0
by	10	100	0	8	0	2
he	2	100	0	2	0	0
I	1	0	1 ([ɑ])	0	0	0
my	2	100	0	1	0	1
through	1	100	0	1	0	0
to	7	29	4	2	1	0
we	1	100	0	1	0	0
you	3	100	0	3	0	0

In order to further investigate the possibility that *to* reduction might be related to following vowel or consonant, we chose to examine the entire labeled corpus, despite the fact that neural network hypotheses would not be available for all tokens. The results of this examination are in Table 6-2. Interestingly, before vowels, *to* is reduced half as often as it is unreduced. Before consonants, *to* is reduced more than twice as much as it is unreduced. On the surface, this behavior appears to be similar to *the*, though not as categorical. With *the*, the unreduced [ðɪ] form occurs almost exclusively before vowels, while the reduced forms, [ðə] or [ðɨ], occur almost exclusively before consonants. Unfortunately, the data is rather skewed in the preceding consonant direction, making statistical analysis somewhat tentative. However, Fisher's exact test showed the difference between the preceding consonant and preceding vowel cases to be significant at $p < .05$.

Table 6-2: Reduced and unreduced *to* before consonants and vowels in complete corpus

N	preceding consonant		preceding vowel	
	reduced	unreduced	reduced	unreduced
103	66	28	3	6

If we look further into *a*, one possible pattern emerges: if *a* occurs at the beginning of an intermediate or intonational phrase, it likelihood of not being reduced is 75%. However,

Fisher's exact test showed that partitioning *a* into reduced and unreduced sets on the basis of the word's starting an intermediate or intonational phrase, as shown in Table 6-3, was not quite significant ($p = .07$). In order to have more data, we examined the complete corpus without regard to neural network hypotheses in Table 6-4. As can be seen the preference for unreduced a phrase-initially is clearer when more data is considered. In fact, a chi-squared test of independence showed the difference between the phrase initial and non-phrase-initial categories to be significant at $p < .01$.

Table 6-3: Reduced and unreduced *a* depending on phrase initial status in test data

N	phrase-initial		non-phrase-initial	
	reduced	unreduced	reduced	unreduced
19	2	6	8	3

Table 6-4: Reduced and unreduced *a* depending on phrase-initial status in complete data

N	phrase-initial		non-phrase initial	
	reduced	unreduced	reduced	unreduced
201	17	57	75	52

It is clear that understanding the segmental and prosodic factors explaining vowel reduction in function words other than *the* is more complex than the allophonic and

allomorphic phenomena we have analyzed previously. It is hoped that by analyzing increased data and more complex forms of discourse, a more satisfactory understanding of this problem will emerge.

6.6. Dialect

We will now examine several dialect phenomena learned by the postlexical neural network. In these cases, the postlexical neural network has adapted the Generalized American dialect in the lexical database to the forms it has encountered in the labeled recorded database. Among the phenomena we will investigate are /ʍ/ voicing, vowel variation before /r/, the low back merger and stem-final tensing.

The lexical database makes a distinction between voiced and voiceless *wh* as in *which* [ʍɪtʃ] *witch* [wɪtʃ], however our speaker does not. There were only two cases of orthographic *wh* in the test data, but the neural network correctly learned to use [w] instead of [ʍ] in both instances. While in the past a distinction between /w/ and /ʍ/ was widespread in the Northern and Southern United States (Kurath and McDavid 1961), according to the Phonological Atlas of North America,[27] it appears to be a recessive feature in most contemporary American dialects.

As is well known, there is a great deal of variation in vowels before /r/ in American English. As was discussed in section 3.1, we chose to represent distinctions rather than

mergers in such cases, since mergers would be easier to derive from distinctions than vice versa given only phonological information. For example, Lexorola lists *marry* /mæri/, *Mary* /meri/ and *merry* /mɛri/. Now, these distinctions are only common in the Eastern United States (e.g. Hartman 1985), somewhat correlated with currently or formerly *r*-less areas (Miller 1993).

Table 6-1 shows neural network performance on three vowels before /r/ where the lexicon differs from the original speech. Due to the categorical nature of these differences, neural network performance is generally good. The one error in merging a lexical /ær/ sequence was in the word *arrow*, which was also glottalized. It should be noted that the lexical database does not encode the distinction between *morning* /mɔrnɪŋ/ and *mourning* /mornɪŋ/, since that is rather recessive in most regions of the United States (Wells 1982, 161). The choice between lexical /or/ and /ɔr/ might be seen as fairly arbitrary, though /or/ would be preferential for dialects that do not otherwise need /ɔr/ due to the ɑ/ɔ merger .

[27] www.ling.upenn.edu/phono_atlas/maps/Map8.html.

Table 6-1: Vowel mergers before /r/

Lexical	N	% original /ɛr/	% neural network /ɛr/
/ær/	3	100	67
/er/	5	100	100
Lexical	N	% original /ɔr/	% neural network /ɔr/
/or/	18	100	100

We have discussed the low back merger in American English in words like *caught* and *cot* with respect to both the lexical database in section 3.1, and the allophony experiments in Chapter 4. In the testing subset of the postlexical neural network data, there were only two cases of mismatch between /ɑ/ and /ɔ/. In both cases, *frolic* and *swan*, the lexical pronunciation had /ɔ/ and the original postlexical pronunciation had /ɑ/. The postlexical neural network did not successfully learn to imitate the original pronunciation in either case. As mentioned in section 3.1, consistency and reality with respect to varying dialects is very hard to achieve in the lexical transcription of the low back vowels. However, we believe that a future approach in which the lexicon actually merges the low back vowels may be promising, at least in dialects like the present one where the difference between /ɑ/ and /ɔ/ has an allophonic character, as shown in the experiments of Chapter 4.

Halle and Mohanan (1985, 59) discuss the phenomenon of stem-final tensing, where nonlow vowels become tense without simultaneous diphthongization or lengthening. They provide examples where underlying /ɪ/ becomes [i] at the ends of various

constituents: *cit**y*** (word finally), *cit**ie**s* (before inflection), *cit**y** hall* (stem-finally in compounds), and *happ**i**ness* (before some stress-neutral suffixes). Halle and Mohanan note that the presence of stem-final tensing in each of these environments is dialect-dependent, though they do not identify the dialects with which the environments are associated. We have noted a generalization of this process in our speaker in certain nonneutral prefixes (cf. discussion in Selkirk 1984, Chapter 3; Kreidler 1989, 227-230).

Table 6-2 shows all words with nonneutral *be-, pre-* and *re-* prefixes in the original data. Lexical and postlexical transcriptions are provided, with postlexical neural network hypotheses where available. As can be seen, in the majority of cases, lexical schwa becomes tense [i] in this environment. Although we believe this to be a dialect phenomenon, it may also be influenced by the formal style of the recordings. In addition, many of these words are somewhat rare, and, as Fidelholtz (1975) has shown, likelihood of schwa in content words increases with the familiarity of the words.

It is also possible that the speaker is forming an analogy on the neutral, productive *re-* prefix, as in *reconnect, refreeze,* etc., which is usually pronounced [ri] in dialects where the nonneutral prefixes in Table 6-2 are pronounced with [ə].[28] Though we have scant data, it is interesting to note that stem-final tensing was learned by the neural network in the case of *beheld*, though not in *resume*.

[28] Dialects that distinguish between the two *re-* prefixes would have minimal pairs in words like *recall* 'remember' vs. *recall* 'take back' and *recount* 'tell' vs. *recount* 'count again'.

Table 6-2: Stem-final tensing in prefixes

orthography	lexical representation	original speech	neural network
bedraggled	bə	bi	
began	bə	bi	
beheld	bə	bi	bi
bemoaned	bə	bi	
bemused	bə	bi	
defense[29]	də	di	
preceding	prə	pri	
recalled	rə	ri	
received	rə	ri	
recording	rə	ri	
rejected	rə	ri	
repelled	rə	ri	
resources	rə	ri	
respected	rə	ri	
respectful	rə	ri	
restores	rə	ri	
resumed	rə	ri	rɨ
revenge	rə	ri	
revised	rə	ri	
revived	rə	ri	
revolve	rə	ri	
besieged	bə	bə	
precocious	prə	prɨ	
preferred	prə	pəˑ	
preparing	prə	pəˑ	
reduce	rə	rɨ	
removed	rə	rɨ	
repaired	rə	rɨ	
reportedly	rə	rɨ	
responsible	rə	rɨ	

Note: Prefix is in bold in orthography column. Other columns only transcribe prefix. Neural network data only available for testing subset of data.

[29] According to Merriam-Webster (1993), when this word is used as an antonym of *offense*, it is often pronounced [ˈdifɛns].

Chapter 7. Conclusion

In this dissertation, we have presented a methodology for attaining appropriate speaker-specific context-dependent pronunciations for speech synthesis. The method is applicable in an economical fashion to a wide range of dialects, due to the use of a dialect-generalized dictionary and automatic training procedures requiring labeled speech corpora of individuals.

In contrast to pronunciation modeling work in speech recognition, we have emphasized the importance of modeling individual variation in an effort to provide a synthetic voice that is as natural as possible. Aiming to model individual variation has not always been considered a worthwhile or even a possible endeavor. According to Labov:

> The construction of complete grammars for "idiolects", even one's own, is a fruitless and unrewarding task; we now know enough about language in its social context to realize that the grammar of the speech community is more regular and systematic than the behavior of any one individual. Unless the individual speech pattern is studied within the overall system of the community, it will appear as a mosaic of unaccountable and sporadic variation. (Labov 1972, 124)

In contrast to this view, we hope to have shown that individual phonological grammar can indeed be profitably modeled. Of course, we have gained a great deal by being able to compare and contrast our results with speech community studies.

We have introduced a postlexical neural network, capable of learning postlexical processes in symbolic terms. This network mediates between a generalized phonemic

dictionary and a phonetic implementation network, which is responsible for generating acoustic parameters that will be synthesized into speech. In Chapter 4, we discussed a number of acoustic experiments designed to assess the extent to which some of the kinds of allophony learned by the postlexical network could actually be learned directly by the phonetic implementation network, that is, without symbolic mediation.

Since the results of the acoustic experiments were promising in the sense that allophony in reduced vowels and /u/ was able to be learned without symbolic mediation, albeit not as well as with symbolic mediation, it is worth considering what prospects exist for future speech synthesis systems that aim to recreate individual variation with the kind of fidelity for which we have aimed here.

One of the major labor bottlenecks for the approach described here is the hand-labeling of speech corpora. This task is time-consuming, often tedious, and requires substantial training and expertise. In addition, attempts to divide the labor among several individuals often results in high degrees of labeler disagreement (Fulop and Keating 1996, Lander et al. 1995) which must be accounted for and dealt with appropriately.

As an alternative to hand-labeling, several researchers have described automatic labeling techniques using automatic speech recognition (e.g. Wightman and Talkin 1997). Such techniques offer the promise of fast labeling of great amounts of data, at the possible expense of labeling precision. It is possible that the greater amount of data permitted by automatic labeling might compensate for any loss in accuracy. However, there are certain

speech events such as the closure and release of stops whose tolerance for inaccuracy may be minimal.

One of the ways in which accuracy in automatic labeling is increased, and indeed the common way by which segmental divisions are achieved, is by "force-aligning" the speech with a transcription of the speech. That transcription may originally be in orthographic form, in which case a phone transcription is obtained from a dictionary, letter-to-sound system, or input by hand. In this situation, a speech-recognition dictionary which potentially offers several sociolinguistically-variant pronunciations for words (in contrast to part of speech or semantic homographs; consider the discussion in section 3.1) may be available.

Let us consider the different possibilities for a speech recognition dictionary to be used in conjunction with an automatic transcription system. First, we will consider the implications of dictionary selection on the standard speech recognition problem, where the transcription of the utterance is unavailable. Riley and Ljolje (1996) contrast three different lexical representation systems with respect to the performance achieved on a speech recognition task. As shown in Table 7-1, Riley and Ljolje demonstrated that using multiple phonetic pronunciations with associated likelihoods (derived from corpus analysis or alignment techniques such as those discussed above) can result in improved word accuracy.

Table 7-1: Comparison of lexical representation system with speech recognizer performance

lexical system	percentage word accuracy on Resource Management task
single phonemic	93.4
single phonetic	94.1
multiple phonetic with likelihoods	96.3

Source: Riley and Ljolje (1996).

Consider first the situation where a dictionary with a single phonemic transcription (similar to the synthesis dictionary described in section 3.1) is used. In this case, it is probable that mismatches will occur between the pronunciation used in the dictionary and the pronunciation used by the speaker. For example, a speaker might pronounce *marry* [mɛri], while the dictionary has /mæri/. In this case, the automatically transcribed database will have a wider array of sounds labeled [æ] than might occur in a hand-labeled database. Acoustic models for synthesis generated from such an automatically labeled database will consequently have a more diffuse model of [æ].

Now, it is possible that the regularity of the [æ]/[ɛ] "allophony" will be learnable and accurately reproduced by the phonetic implementation neural network or its equivalent. Indeed, we have shown that our phonetic implementation network can learn allophonic distinctions without symbolic mediation in Chapter 4. A speech synthesis system using

an automatically transcribed (or hand-labeled) database at a dialect-generalized phonemic level will be able to dispense with a postlexical processor as described in this dissertation.

However, as we showed in Chapter 4, use of allophonic symbols results in increased fidelity between the original speech and that generated by the phonetic implementation network. This fact, coupled with the improved recognition results using allophonic symbols shown by Riley and Ljolje (1996), indicates that use of allophonic symbols may result in better automatic labeling as well as better phonetic models for synthesis.

Such a system would require two separate dictionaries, one containing multiple pronunciations for use with the automatic transcription system, and one containing dialect-generalized phonemic pronunciations for speech synthesis. A postlexical processor would be necessary to mediate between the synthesis dictionary and the allophonic transcription labels, much as in the current system. Rather than training the postlexical processor on hand-labeled speech as in the current system, it could be trained on aligned allophonic transcriptions from the automatic transcriber and dialect-generalized phonemic transcriptions from the synthesis dictionary.

Riley and Ljolje (1996) actually describe two methods for generating multiple phonetic transcriptions. One involves training decision trees on the hand-labeled TIMIT corpus (6300 sentences), and the other involves using the same procedure on 37,000 automatically transcribed sentences from the North American Business (NAB) task. It was hoped that the allophonic information from the larger NAB database might prove of

greater value for the recognition dictionary, despite any loss of accuracy accrued from use of an automatic method. In fact, word accuracy for the NAB task using pronunciation networks generated from automatic transcription was .8% better than that achieved using the TIMIT-based networks. Such results provide a promising indication that future postlexical networks could be trained using automatically transcribed data.

Speech recognition and synthesis applications attempt to simulate the human activities of speaking and listening. We believe that successful speech applications must strive for human-like competency in order to be accepted by humans. We also believe that attaining human-like competency cannot be achieved without adopting human-like methods in the conversion between text and linguistic representation.

We have explored the different levels at which linguistic information can be provided to a speech synthesizer, including the lexicon, speech labeling and a postlexical module. We hope to have presented a procedure for intelligently allocating linguistic resources at both the symbolic level and the level of acoustic-phonetic realization. We hope that this procedure results in speech that is considered natural and acceptable by listeners, in addition to being comprehensible when employed in situations where the message content is complex.

Appendix

Table A-1: TIMIT/IPA Correspondences

TIMIT	IPA	TIMIT	IPA	TIMIT	IPA	TIMIT	IPA
p	p	th	θ	hh	h	ao	ɔ
t	t	dh	ð	w	w	ow	o
k	k	z	z	ah	ʌ	uw	u
b	b	jh	dʒ	ae	æ	uh	ʊ
d	d	ch	tʃ	aa	ɑ	ay	aɪ
g	g	l	l	eh	ɛ	oy	ɔɪ
y	j	r	r	ey	e	aw	aʊ
v	v	m	m	ih	ɪ	sh	ʃ
s	s	n	n	iy	i	ax	ə
en	n̩	el	l̩				

References

Adamson, M., and Robert I. Damper. 1996. A recurrent network that learns to pronounce English text. ICSLP.

Ainsworth, W., and Warren N. 1992. Applications of multilayer perceptrons in text-to-speech synthesis systems. *Neural networks for vision, speech, and natural language,* ed. R. Linggard, D. J. Meyers, and C. Nightingale, 256-288. London: Chapman & Hall.

Albano, Eleonora Cavalcante, and Agnaldo Antonio Moreira. 1996. Archisegment-based letter-to-phone conversion for concatenative speech synthesis in Portuguese. ICSLP.

Allen, Jonathan, Sheri Hunnicutt, and Dennis Klatt. 1987. *From text to speech: The MITalk system.* Cambridge: Cambridge University Press.

Alleva, F. and K. F. Lee. 1989. Automatic new word acquisition: Spelling from acoustics. *Proceedings of the DARPA speech and natural language workshop.*

Archangeli, D. 1988. Aspects of underspecification theory. *Phonology* 5:183-207.

Barry, W. J., and A. J. Fourcin. 1992. Levels of labeling. *Computer Speech and Language* 6:1-14.

Baum, L. E. 1972. An inequality and associated maximization technique in statistical estimation of probabilistic functions of a Markov process. *Inequalities* 3:1-8.

Beckman, Mary, and Gayle Ayers Elam. 1997. Guidelines for ToBI Labeling, version 3. Manuscript.

Bell, Alan. 1984. Language style as audience design. *Language in Society* 13: 145-204.

Bird, Steven. 1995. *Computational phonology: A constraint-based approach.* Cambridge: Cambridge University Press.

Bladon, Anthony, Rolf Carlson, Björn Granström, Sheri Hunnicutt, and Inger Karlsson. 1987. A text-to-speech system for British English and issues of dialect and style. *European conference on speech technology,* ed. J. Laver and M. Jack, volume 1.

Blaauw, Eleonora. 1994. The contribution of prosodic boundary markers to the perceptual difference between read and spontaneous speech. *Speech Communication* 14: 359-375.

Blevins, Juliette. 1995. The syllable in phonological theory. In *The handbook of phonological theory,* ed. John A. Goldsmith. Cambridge, Mass.: Blackwell.

Bolinger, Dwight. 1981. Two kinds of vowels, two kinds of rhythm. Manuscript reproduced by the Indiana University Linguistics Club.

Booij, G., and J. Rubach. 1987. Postcyclic vs. postlexical rules in lexical phonology. *Linguistic Inquiry* 18.

Bradlow, Ann R., Gina M. Torretta and David B. Pisoni. 1996. Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech Communication* 20:255-272.

Breiman, Leo, Jerome H. Friedman, Richard A. Olshen, and Charles J. Stone. 1993. *Classification and regression trees*. New York: Chapman & Hall.

Broe, Michael. 1993. Specification theory: The treatment of redundancy in generative phonology. Ph.D. diss., University of Edinburgh.

Broe, Michael. 1996. A generalized information-theoretic measure for systems of phonological classification and recognition. *ACL Workshop on Computational Phonology in Speech Technology,* 17-24.

Broe, Michael. 1997. Stochastic constraints in phonological analysis. CLS monthly talk, University of Chicago.

Bronstein, Arthur J. 1960. *The pronunciation of American English*. New York: Appleton-Century-Crofts.

Browman, Catherine P., and Louis Goldstein. 1986. Towards an articulatory phonology. In *Phonology Yearbook* 3, ed. C. Ewen and J. Anderson, 219-252. Cambridge: Cambridge University Press.

Browman, Catherine P., and Louis Goldstein. 1990. Tiers in articulatory phonology, with some implications for casual speech. In *Papers in Laboratory Phonology* I, ed. John Kingston and Mary E. Beckman, 341-376. Cambridge: Cambridge University Press.

Browman, Catherine P., and Louis Goldstein. 1992. "Targetless" schwa: an articulatory analysis. In *Gesture, Segment, Prosody, Papers in laboratory phonology* II, ed. Gerard J. Docherty and D. Robert Ladd, 26-56. Cambridge: Cambridge University Press.

Byrd, Dani. 1994. Relations of sex and dialect to reduction. *Speech Communication* 15: 39-54.

Byrne, B., M. Finke, S. Khudanpur, J. McDonough, H. Nock, M. Riley, M. Saraclar, C. Wooters, and G. Zavaliagkos. Pronunciation modeling for conversational speech recognition: A status report from WS97. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings,* ed. Sadaoki Furui, B.-H. Juang, and Wu Chou, 26-33. Piscataway, N.J.: IEEE.

Charniak, Eugene. 1993. *Statistical language learning.* Cambridge, Mass.: MIT Press.

Chomsky, Noam. 1965. Aspects of the theory of syntax. Cambridge, Mass.: MIT Press.

Chomsky, Noam. 1981. Principles and parameters in syntactic theory. In *Explanation in Linguistics: The Logical Problem of Language Acquisition,* ed. Norbert Hornstein and David Lightfoot, 32-75. London: Longman.

Chomsky, Noam, and Morris Halle. 1968. *The sound pattern of English.* Cambridge, Mass.: MIT Press.

Church, Kenneth W. 1983. *Phrase-structure parsing: a method for taking advantage of allophonic constraints.* Ph.D. diss., MIT, 1983. Distributed by Indiana University Linguistics Club. Also published as *Phonological parsing in speech recognition.* 1987. Boston: Kluwer Academic Publishers.

Church, Kenneth W. 1986. Stress assignment in letter to sound rules for speech synthesis. ICASSP 86.2423-2426.

Church, Kenneth W. 1989. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text. IEEE.

Clements, George N. 1976. Palatalization: linking or assimilation? CLS 12, 96-109.

Clements, George N. 1985. The geometry of phonological features. *Phonology Yearbook* 2.225-252.

George N. Clements, and Elizabeth V. Hume. 1995. The internal organization of speech sounds. In *The handbook of phonological theory*, ed. John A. Goldsmith, 245-306. Cambridge, MA: Blackwell.

Cohen, Andrew. 1995. Developing a non-symbolic phonetic notation for speech synthesis. *Computational Linguistics* 21:567-575.

Cohen, Andrew. 1997. Embedding prior phonetic knowledge in neural network based text-to-phonetics conversion. Technical report, University of Reading.

Cohen, Michael Harris. 1989. Phonological structures for speech recognition. Ph.D. diss., University of California, Berkeley.

Coile, Bert Van. 1990. Inductive Learning of Grapheme-to-Phoneme rules. ICSLP.19.1.1-19.1.4.

Coile, Bert Van. 1991. Inductive learning of pronunciation rules with the DEPES system. IEEE.

Coker, Cecil H., Kenneth. W. Church, and Mark Liberman. 1990. Morphology and rhyming: Two powerful alternatives to letter-to-sound rules for speech synthesis. Coling.

Coleman, John. 1993. English word-stress in unification-based grammar. In *Computational Phonology, Edinburgh Working Papers in Cognitive Science* 8, ed. T. Mark Ellison and James Scobbie, 97-106.

Coleman, John. 1995a. Phonology and computational linguistics— a personal overview. In *Linguistics and Computation*, ed. Jennifer Cole, Georgia M. Green and Jerry L. Morgan, 223-254. Stanford: CSLI Publications.

Coleman, John. 1995b. Declarative lexical phonology. In *Frontiers of Phonology*, ed. Jacques Durand and Francis Katamba, 333-382. London: Longman.

Coleman, John, and Janet Pierrehumbert. 1997. Stochastic phonological grammars and acceptability. Third Meeting of the ACL Special Interest Group in Computational Phonology, 49-56.

Coltheart, M. 1978. Lexical access in simple reading tasks. *Strategies of Information Processing*, ed. G. Underwood. New York: Academic.

Corrigan, Gerald. 1996. Speaker understandability as a function of prosodic parameters. Ph.D. diss., Northwestern University.

Corrigan, Gerald, Noel Massey and Orhan Karaali. 1997. Generating Segment Durations in a Text-to-Speech System: A Hybrid Rule-Based/Neural Network Approach. Eurospeech '97.

Covington, Michael A. An algorithm to align words for historical comparison. *Computational Linguistics* 22:481-496.

Cremelie, Nick, and Jean-Pierre Martens. 1996. Generation of word pronunciation networks from automatically learned inter-word coarticulation rules. Proceedings of ProRISC/IEEE Workshop on Circuits, Systems and Signal Processing, 89-94.

Daelemans, Walter, Steven Gillis, and Gert Durieux. 1994. The acquisition of stress: a data-oriented approach. *Computational Linguistics* 20:421-451.

Daelemans, Walter M. P., and Antal P. J. Van den Bosch. 1997. Language independent data-oriented grapheme-to-phoneme conversion. In *Progress in speech synthesis*, ed. Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, 77-89. New York: Springer.

Damper, Robert I. 1995. Self-learning and connectionist approaches to text-phoneme conversion. *Connectionist models of memory and language*, ed. by J. Levy, D. Bairaktaris and P. Cairns, 117-144. London: UCL Press.

Dedina, M. J., and Howard C. Nusbaum. 1991. PRONOUNCE: A program for pronunciation by analogy. *Computer Speech and Language* 5:55-64.

Delattre, Pierre. 1968. Le jeu de l'*e* instable intérieur en français. In *Studies in French and comparative phonetics*, 17-22. The Hague: Mouton.

Deshmukh, Neeraj, Mary Weber, and Joseph Picone. 1996. Automated generation of N-best pronunciations of proper nouns. ICSLP.

Dilley, L., S. Shattuck-Hufnagel, and M. Ostendorf. 1996. Glottalization of word-initial vowels as a function of prosodic structure. *Journal of Phonetics* 24:423-444.

Divay, Michel, and Anthony J. Vitale. 1997. Algorithms for grapheme-phoneme translation for English and French: Applications. *Computational Linguistics* 23: 495-524.

Dresher, B. Elan, and Jonathan D. Kaye. 1990. A computational learning model for metrical phonology. Cognition 34: 137-195.

Dutoit, Thierry. 1997. *An introduction to text-to-speech synthesis*. Dordrecht: Kluwer.

Egan, James P. 1944. Articulation testing methods, II. OSRD Report No. 3802.

Egan, James P. 1948. Articulation testing methods. *Laryngoscope* 58: 955-991..

Ellison, T. Mark. 1994. Phonological derivation in optimality theory. Coling 94.1007-1113.

Elman, Jeffrey L. 1990. Finding structure in time. *Cognitive Science* 14:179-211.

Elman, Jeffrey L. and James L. McClelland. 1986. Exploiting lawful variability in the speech wave. In *Invariance and variability in speech processing*, ed. J. S. Perkell and D. H. Klatt, 360-385. Hillsdale, N.J.: Erlbaum.

Elovitz, H.S., R. Johnson, A. McHugh, and J.E. Shore. 1976. Letter-to-sound rules for automatic translation of English text to phonetics. IEEE ASSP-24.446-459.

Eskénazi, Maxine. 1992. Changing speech styles: strategies in read speech and casual and careful spontaneous speech. ICSLP 92, 755-758.

Faust, Lioba. 1995. What do spectral and perceptual analyses reveal about spontaneous speech in dialogues of different style? ICPhS 95: Vol. 4, 236-239.

Fisher, William M., and Ira J. Hirsh. 1976. Intervocalic flapping in English. CLS, 183-198.

Fisher, W. M., V. Zue, J. Bernstein, and D. S. Pallett. 1987. An acoustic-phonetic database. *Journal of the Acoustic Society of America*, Supplement 1: 81.

Fitt, Susan. 1995. The pronunciation of unfamiliar native and non-native town names. *Eurospeech* 95, 2227-2230.

Fitt, Susan. 1997. The generation of regional pronunciations of English for speech synthesis. *Eurospeech* 97, 2447-2450.

Fosler, Eric, Mitch Weintraub, Steven Wegmann, Yu-Hung Kao, Sanjeev Khudanpur, Charles Galles, and Murat Saraclar. 1996. Automatic learning of word pronunciation from data. ICSLP '96.

Fulop, Sean A., and Patricia A. Keating. 1996. Pronunciation variability in the Switchboard corpus. Poster presented at the Third Joint Meeting of the Acoustical Society of America and the Acoustical Society of Japan.

Gaskell, M. Gareth, Mary Hare, and William D. Marslen-Wilson. 1995. A connectionist model of phonological representation in speech perception. *Cognitive Science* 19:407-439.

Gerson, Ira, Noel Massey, Gerald Corrigan, and Orhan Karaali. 1996. Neural network speech synthesis. Sixth Australian International Conference on Speech Science and Technology.

Giegerich, Heinz J. 1992. *English phonology: an introduction.* Cambridge: Cambridge University Press.

Gildea, Daniel, and Daniel Jurafsky. 1996. Learning bias and phonological-rule induction. *Computational Linguistics* 22:497-530.

Glushko, R. J. 1981. Principles for pronouncing print: The psychology of phonography. Interactive processes in reading, ed. by C. A. Perfetti. Hillsdale, N.J.: Lawrence Erlbaum.

Godfrey, John J., Edward C. Holliman, and Jane McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. ICASSP '92, I-517-I-520.

Golding, Andrew. 1991. Pronouncing names by a combination of rule-based and case-based reasoning. Ph.D. diss., Stanford University.

Goldsmith, John. 1990. *Autosegmental and metrical phonology*. Oxford: Basil Blackwell.

Goldsmith, John. 1992. Local modeling in phonology. *Connectionism: theory and practice*, ed. Steven Davis, 229-246. New York: Oxford University Press.

Goldsmith, John. 1993. Harmonic Phonology. In *The last phonological rule*, ed. John Goldsmith, 21-60. Chicago: University of Chicago Press.

Golston, Chris. 1996. Direct optimality theory: Representation as pure markedness. *Language* 72:713-748.

Greenberg, Steven, Joy Hollenback, and Dan Ellis. 1996. Insights into spoken language gleaned from phonetic transcriptions of the Switchboard corpus. ICSLP 96.

Grimes, Joseph E. 1988. Information dependencies in lexical subentries. In *Relational models of the lexicon*, ed. Martha Walton Evens. Cambridge, England: Cambridge University Press.

Gupta, Prahlad, and David S. Touretzky. 1994. Connectionist models and linguistics theory: investigations of stress systems in language. *Cognitive Science* 18: 1-50.

Guy, Gregory. 1980. Variation in the group and the individual: the case of final stop deletion. In *Locating language in time and space*, ed. William Labov, 1-36. New York: Academic Press.

Guy, Gregory. 1991. Explanation in variable phonology: An exponential model of morphological constraints. *Language Variation and Change* 3:1-22.

Guy, Gregory, and Charles Boberg. 1997. Inherent variability and the obligatory contour principle. *Language Variation and Change* 9:149-164.

Haegeman, Liliane. 1991. *Introduction to government and binding theory.* Oxford: Blackwell.

Halle, Morris, and K.P. Mohanan. 1985. Segmental phonology of modern English. *Linguistic Inquiry* 16: 57-116.

Hammond, Michael. 1995. Syllable parsing in English and French. ROA-58-0000.

Hammond, Michael. 1997. Vowel quantity and syllabification in English. *Language* 73:1-17.

Hare, Mary. 1990. The role of similarity in Hungarian vowel harmony: a connectionist account. *Connection Science*. 2: 123-150.

Harris John, and Geoff Lindsey. 1995. The elements of phonological representation. *Frontiers of Phonology: atoms, structure, derivations*, ed. Jacques Durand and Francis Katamba, 34-79. London: Longman.

Hartman, James W. 1985. Guide to pronunciation. In *Dictionary of American Regional English, Vol. 1,* ed. Frederic G. Cassidy, xli-lxi. Cambridge, Mass.: Belknap/Harvard.

Hayes, Bruce. 1982. Extrametricality and English stress. *Linguistic Inquiry* 13:227-276.

Hayes, Bruce. 1990. Precompiled phrasal phonology. In *The phonology-syntax connection*, ed. Sharon Inkelas and Draga Zec, 85-108. Chicago: University of Chicago Press.

Hayes, Bruce. 1995. *Metrical stress theory: principles and case studies.* Chicago: University of Chicago Press.

Herold, Ruth. 1990. Mechanisms of merger: the implementation and distribution of the low back merger in eastern Pennsylvania. Ph.D. diss., University of Pennsylvania.

Hertz, Susan, and Marie K. Huffman. 1992. A nucleus-based timing model applied to multi-dialect speech synthesis by rule. *ICSLP '92,* 1171-1174.

Hetherington, Irvine Lee. 1995. A characterization of the problem of new, out-of-vocabulary words in continuous-speech recognition and understanding. Ph.D. diss., MIT.

Hieronymus, James L. 1993. ASCII Phonetic Symbols for the World's Languages: Worldbet. Bell Laboratories manuscript.

Hochberg, J., S. Mniszewski, T. Calleja, and A. Papcun. 1991. A default hierarchy for pronouncing English. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13.957-964.

Hock, Hans Heinrich. 1986. *Principles of historical linguistics.* Berlin: Mouton de Gruyter.

Hopcroft, John E., and Jeffrey D. Ullman. 1979. *Introduction to automata theory, languages, and computation.* Reading, MA: Addison-Wesley.

House, A. S., C. E. Williams, M. H. L. Hecker and K. D. Kryter. 1965. Articulation testing methods: consonantal differentiation with a closed response set. *Journal of the Acoustic Society of America* 37: 158-166.

Huggins, A. W. F., and Yogen Patel. 1996. The use of shibboleth words for automatically classifying speakers by dialect. ICSLP.

Hume, Elizabeth V. 1992. Front vowels, coronal consonants and their interaction in nonlinear phonology. Ph.D. diss., Cornell University.

Hunnicutt, Sheri, Helen Meng, Stephanie Seneff, and Victor Zue. 1993. Reversible letter-to-sound sound-to-letter generation based on parsing word morphology. Eurospeech, 1993.

IEEE. 1969. IEEE Recommended practice for speech quality measurements.

Jakobson, Roman, C. G. M. Fant, and Morris Halle. 1952. Preliminaries to speech analysis: the distinctive features and their correlates. Tech. Rep. No. 13, Acoustics Laboratory, MIT.

Jannedy, Stefanie, and Bernd Möbius. 1997. Name pronunciation in German text-to-speech synthesis. ANLP '97, 49-56.

Jenkins, J. J., and L. D. Franklin. 1981. Recall of passages of synthetic speech. Paper presented at the Psychonomics Society meeting, November.

Jensen, John T. 1993. *English phonology*. Amsterdam: John Benjamins.

Jiang, Li, Hsiao-Wuen Hon and Xuedong Huang. 1997. Improvements on a trainable letter-to-sound converter. Eurospeech '97, 605-608.

Jordan, M. 1986. Serial order: a parallel distributed processing approach. ICS Report No. 8604. UC San Diego.

Junqua, Jean-Claude, and Jean-Paul Haton. 1996. *Robustness in automatic speech recognition: fundamentals and applications.* Boston: Kluwer.

Kahn, Daniel. 1976. Syllable-based generalizations in English phonology. Ph.D. diss., MIT.

Kaisse, Ellen. 1985. *Connected speech.* San Diego: Academic Press.

Kaisse, Ellen. 1990. Toward a typology of postlexical rules. In *The phonology-syntax connection,* ed. Sharon Inkelas and Draga Zec, 127-143. Chicago: University of Chicago Press.

Kaplan, Ronald M., and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics* 20:331-378.

Karaali, Orhan. 1989. A very high performance neural network system architecture using grouped weight quantization. Ph.D. diss., Florida Atlantic University.

Karaali, Orhan. 1994. Fast random training method for recurrent backpropagation neural networks. *Motorola Technical Developments* 21:48-49.

Karaali, Orhan, Gerald Corrigan, and Ira Gerson. 1996. Speech Synthesis with Neural Networks. World Conference on Neural Networks.

Karaali, Orhan, Gerald Corrigan, Ira Gerson and Noel Massey. 1997. Text-to-Speech Conversion with Neural Networks: A Recurrent TDNN Approach. Eurospeech '97.

Karaali, Orhan, Gerald Corrigan, Ira Gerson, Noel Massey, Corey Miller, Otto Schnurr and Andrew Mackie. Forthcoming. Application of multiple neural networks for high quality text-to-speech synthesis.

Karaali, Orhan, Gerald Corrigan, Noel Massey, Corey Miller, Otto Schnurr, and Andrew Mackie. 1998. A high quality text-to-speech system composed of multiple neural networks. ICASSP.

Kaye, Jonathan. 1989. *Phonology: A Cognitive View.* Hillsdale, N.J.: Lawrence Erlbaum.

Keating, Patricia A., Dani Byrd, Edward Flemming, and Yuichi Todaka. 1994. Phonetic analyses of word and segment variation using the TIMIT corpus of American English. *Speech Communication* 14:131-142.

Kenstowicz, Michael. 1994. *Phonology in generative grammar.* Cambridge, MA: Blackwell.

Kenyon, John Samuel, and Thomas Albert Knott. 1953. *A pronouncing dictionary of American English.* Springfield, Mass.: G. & C. Merriam Company.

Kintsch, W., and T. A. van Dijk. 1978. Toward a model of text comprehension and production. *Psychological Review* 85:363-394.

Kiparsky, Paul. 1979. Metrical structure assignment is cyclic. *Linguistic Inquiry* 10: 421-441.

Kiparsky, Paul. 1985. Some consequences of lexical phonology. *Phonology Yearbook* 2: 85-138.

Kiparsky, Paul. 1995. The phonological basis of sound change. In *The handbook of phonological theory,* ed. by John A. Goldsmith, 640-670. Cambridge, Mass.: Blackwell.

Klatt, Dennis. 1980. Scriber and Lafs: two new approaches to speech analysis. In *Trends in speech recognition,* ed. Wayne A. Lea. Englewood-Cliffs, N.J.: Prentice-Hall.

Knight, Kevin, and Jonathan Graehl. 1997. Machine transliteration. In *Proceedings of the 35th Annual Meeting of the Association for Computionational Linguistics,* 128-135.

Kohonen, Teuvo. 1988. The 'neural' phonetic typewriter. IEEE Computer 21.11-22.

Kohonen, Teuvo. 1989. *Self-organization and associative memory.* Berlin: Springer.

Koopmans-Van Beinum, Florien J. 1991. Spectro-temporal reduction and expansion in spontaneous speech and read text: Focus words versus non-focus words. In *ETRW: Phonetics and Phonology of Speaking Styles,* 36.1-36.5.

Koopmans-Van Beinum, Florien J. 1992. The role of focus words in natural and synthetic continuous speech: Acoustic aspects. *Speech Communication* 11:439-452.

Koskenniemi, Kimmo. 1983. Two-level morphology: a general computational model of word-form recognition and production. Publication number 11, Department of Linguistics, University of Helsinki.

Krause, Jean Christine. 1995. The effects of speaking rate and speaking mode on intelligibility. Master's thesis, MIT.

Kreidler, Charles W. 1989. The pronunciation of English. Oxford: Blackwell.

Kruskal, Joseph B. 1983. An overview of sequence comparison. In *Time warps, string edits, and macromolecules*, ed. by Joseph B. Kruskal and David Sankoff, 1-44. Reading, MA: Addison-Wesley.

Kurath, Hans, and Raven I. McDavid, Jr. 1961. *The pronunciation of English in the Atlantic States*. Ann Arbor: University of Michigan Press.

Laan, Gitta P. M. 1997. 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.

Labov, William. 1966. *The social stratification of English in New York City*. Washington, D.C.: Center for Applied Linguistics.

Labov, William. 1972. *Language in the inner city: Studies in the Black English vernacular*. Philadelphia: University of Pennsylvania Press.

Labov, William. 1989. The limitations of context. Chicago Linguistic Society, part 2, 171-200.

Labov, William. 1991. The three dialects of English. In *New ways of analyzing sound change*, ed. Penelope Eckert, 1-44. San Diego: Academic Press.

Labov, William. 1994. *Principles of linguistic change*. Oxford: Blackwell.

Labov, William. 1996. The organization of dialect diversity in North America. ICSLP.

Labov, William. 1997. Resyllabification. In *Variation, change, and phonological theory*, ed. Frans Hinskens, Roeland Van Hout, W. Leo Wetzels. Amsterdam: Benjamins.

Labov, William, Mark Karan, and Corey Miller. 1991. Near-mergers and the suspension of phonemic contrast. *Language Variation and Change* 3:33-74.

Labov, William, Malcah Yaeger, and Richard Steiner. 1972. A quantitative study of sound change in progress. Report on National Science Foundation Contract NSF-GS-3287, University of Pennsylvania.

Ladd, D. Robert. 1996. *Intonational phonology*. Cambridge: Cambridge University Press.

Ladefoged, Peter. 1982. A course in phonetics. 2nd ed. San Diego: Harcourt Brace Jovanovich.

Lahiri, A., and W. D. Marslen-Wilson. 1991. The mental representation of lexical form: a phonological approach to the recognition lexicon. *Cognition* 38:245-294.

Lander, Terri. 1997. CSLU labeling guide. Center for Spoken Language Understanding. Oregon Graduate Institute.

Lander, Terri, Beatrice Oshika, Ronald A. Cole, and Mark Fanty. Multi-language speech database: creation and phonetic labeling agreement. ICPhS 95: 166-169.

Eric. 1997. Rational transductions for phonetic conversion and phonology. In *Finite-state language processing*, ed. Emmanuel Roche and Yves Schabes, 407-429. Cambridge: MIT Press.

Lawrence, S. G. C., and G. Kaye. 1986. Alignment of phonemes with their corresponding orthography. *Computer Speech and Language* 1:153-165.

Lea, W. A. 1980. Prosodic aids to speech recognition. In *Trends in speech recognition*, ed. W. A. Lea. Englewood Cliffs, N.J.: Prentice-Hall.

Leben, Will. 1973. Suprasegmental phonology. Ph.D. diss., MIT.

Liberman, Mark. 1993. Optionality and optimality. Manuscript.

Liberman, Mark. 1994a. Phonological optionality in Latin clitics. *The Penn Review of Linguistics* 18: 87-102.

Liberman, Mark. 1994b. Computer speech synthesis: its status and prospects. In *Voice Communication between humans and machines*, ed. David B. Roe and Jay C. Wilpon. Washington: National Academy of Sciences.

Liberman, Mark, and Kenneth W. Church. 1992. Text analysis and word pronunciation in text-to-speech synthesis. In *Advances in speech signal processing*, ed. S. Furui and M. Sondhi. New York: Marcel Dekker.

Liberman, Mark, and Janet Pierrehumbert. 1984. Intonational invariance under changes in pitch range and length. In *Language sound structure*, ed. Mark Aronoff and Richard T. Oehrle. Cambridge, Mass.: MIT Press.

Linguistic Data Consortium. 1995. COMLEX English pronouncing lexicon. Trustees of the University of Pennsylvania, version 0.2.

Lippmann, Richard P. 1996. Recognition by humans and machines: miles to go before we sleep. *Speech Communication* 18:247-248.

Ljolje, Andrej, Julia Hirschberg, and Jan P. H. van Santen. 1997. Automatic speech segmentation for concatenative inventory selection. In *Progress in speech synthesis*, ed. Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, 305-311. New York: Springer.

Logan, J.S., Greene, B.G., and D.B. Pisoni. 1989. Segmental intelligibility of synthetic speech produced by rule. JASA 86:566-581.

Lucas, S. M., and R. I. Damper. 1992. Syntactic neural networks for bi-directional text-phonetics translation. Talking machines: theories, models and applications, ed. by G. Bailly and C. Benoît. Amsterdam: Elsevier/North-Holland, 127-141.

Lucassen, J., and R. Mercer. 1984. An information-theoretic approach to the automatic determination of phonemic baseforms. ICASSP, 42.5.1-42.5.4.

Luce, P. A. 1981. Comprehension of fluent synthetic speech produced by rule. *Research on perception progress report no.* 7, 229-241. Bloomington, Indiana: Speech Research Laboratory, Psychology Department, Indiana University.

Luk, R. W. P., and Robert I. Damper. 1991. Stochastic transduction for text-to-phoneme conversion. Eurospeech 91, 779-782.

Luk, R. W. P., and Robert I. Damper. 1993. Inference of letter-phoneme correspondences using generalized stochastic transducers. ICSLP 93.

Luk, Robert W. P., and Robert I. Damper. 1991. A novel approach to inferring letter-phoneme correspondences. ICASSP.

Luk, Robert W. P., and Robert I. Damper. 1992. Inference of Letter-Phoneme Correspondences by Delimiting and Dynamic Time Warping Techniques. IEEE.

Luk, Robert W. P., and Robert I. Damper. 1992. Inference of letter-phoneme correspondences with pre-defined consonant and vowel patterns. IEEE.

Magen, Harriet S. 1997. The extent of vowel-to-vowel coarticulation in English. *Journal of Phonetics* 25: 187-205.

McCarthy, John. 1986. OCP effects: Gemination and antigemination. *Linguistic Inquiry* 17:207-263.

McCarthy, John, and Alan Prince. 1993. Generalized Alignment. *Yearbook of Morphology*.

McCarthy, John, and Alan Prince. 1995. Faithfulness and reduplicative identity. Manuscript.

McCulloch, M., M. Bedworth, and J. Bridle. 1987. NETspeak— a re-implementation of NETtalk. *Computer Speech and Language* 2: 289-302.

McHugh, A. 1976. Listener preference and comprehension tests of stress algorithms for a text-to-phonetic speech synthesis program. Naval Research Laboratory Report 8015.

McLemore, Cynthia. 1995. Pronlex transcription. Distributed with Comlex English pronouncing lexicon, Linguistic Data Consortium.

Meng, Helen M., Stephanie Seneff, and Victor Zue. 1996. Reversible letter-to-sound/sound-to-letter generation based on parsing word morphology. *Speech Communication* 18:47-63.

Meng, Helen M.. 1995. Phonological parsing for bi-directional letter-to-sound/sound-to-letter generation. Ph.D. diss., MIT.

Merriam-Webster. 1984. *Webster's Ninth New Collegiate Dictionary.* Springfield, Mass.: Merriam-Webster.

Merriam-Webster. 1989. *The New Merriam-Webster Dictionary.* Springfield, Mass.: Merriam-Webster.

Merriam-Webster. 1996. Merriam-Webster's Collegiate Dictionary. 10th ed. Springfield, Mass.: Merriam-Webster.

Miller, Corey. 1993. American English /r/ and ambisyllabicity. Paper presented at Linguistic Society of America Annual Meeting, Los Angeles.

Miller, Corey, Orhan Karaali, and Noel Massey. 1997. Variation and synthetic speech. Paper presented at NWAVE 26, Quebec.

Miller, Corey, Orhan Karaali, and Noel Massey. 1998. Learning postlexical variation in an individual. Ppaer presented at Linguistic Society of America Annual Meeting, New York.

Miller, Corey, Noel Massey, and Orhan Karaali. 1998. Exploring the nature of postlexical processes. Paper presented at 22nd Penn Linguistics Colloquium, Philadelphia.

Miller, David R., and James Trischitta. 1996. Statistical dialect classification based on mean phonetic features. ICSLP.

Mohanan, K. P. 1982. Lexical phonology. Ph.D. diss., distributed by Indiana University Linguistics Club.

Mohanan, K. P. 1986. *The theory of lexical phonology*. Dordrecht: Reidel.

Moody, T. S., and M. G. Joost. 1986. Synthesized speech, digitized speech and recorded speech: A comparison of listener comprehension rates. *Proceedings of the Voice Input/Output Society,* Alexandria, Virginia.

Nagy, Naomi, and William T. Reynolds. 1996. Accounting for variable word-final deletion within optimality theory. In *Sociolinguistic variation: data, theory, and analysis: selected papers from NWAV 23 at Stanford*. Stanford: CSLI.

Nerbonne, John, and Wilbert Heeringa. 1997. Measuring dialect distance phonetically. In *Proceedings of the third meeting of the ACL special interest group in computational phonology*, ed. John Coleman, 11-18.

Nespor, Marina, and Irene Vogel. 1986. *Prosodic phonology*. Dordrecht: Foris.

Nusbaum, Howard C., Alexander L. Francis, and Anne S. Henly. 1995. Measuring the naturalness of synthetic speech. *International Journal of Speech Technology* 1: 7-19.

Nusbaum, Howard C., Alexander L. Francis, and Tracy Luks. 1995. Comparative evaluation of the quality of synthetic speech produced at Motorola. Research report, Spoken Language Research Laboratory, University of Chicago.

Nye, P. W., F. Ingemann, and L. Donald. 1975. Synthetic speech comprehension: A comparison of listener performances with and preferences among different speech forms. Haskins Laboratories: Status report on speech perception SR-41. 117-126. New Haven, CT: Haskins Laboratories.

O'Shaughnessy, Douglas. 1976. Modeling fundamental frequency, and its relationship to syntax, semantics, and phonetics. Ph.D. diss., MIT.

Paradis, Carole. 1986. On constraints and repair strategies. *The Linguistic Review* 6.71-97.

Paradis, Carole. 1995. Derivational constraints in phonology: evidence from loanwords and implications. *CLS* 31.360-374.

Parfitt, S. H., and R. A. Sharman. 1991. A bidirectional model of English pronunciation. *Eurospeech '91*, 2.800-804.

Picheny, Michael A, Nathaniel I. Durlach, and Louis D. Braida. 1985. Speaking clearly for the hard of hearing I: intelligibility differences between clear and conversational speech. *Journal of speech and hearing research* 28:96-103.

Picheny, Michael A., Nathaniel I. Durlach, and Louis D. Braida. 1986. Speaking clearly for the hard of hearing II: acoustic characteristics of clear and conversational speech. *Journal of speech and hearing research* 29:434-446.

Pierce, John R. 1980. *An introduction to information theory: symbols, signals and noise.* New York: Dover.

Pierrehumbert, Janet B., and Mary Beckman. 1988. *Japanese tone structure.* LI Monograph Series No. 15. Cambridge, Mass.: MIT Press.

Pierrehumbert, Janet B, and Stefan Frisch. 1997. Synthesizing allophonic glottalization. In *Progress in speech synthesis*, ed. Jan P.H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, 9-26. New York: Springer.

Pierrehumbert, Janet B., and David Talkin. 1992. Lenition of /h/ and glottal stop. In *Gesture, Segment, Prosody, Papers in laboratory phonology* II, ed. Gerard J. Docherty and D. Robert Ladd, 90-117. Cambridge: Cambridge University Press.

Pisoni, David B. 1993. Long-term memory in speech perception: Some new findings on talker variability, speaking rate and perceptual learning. *Speech Communication* 13:109-125.

Pisoni, David B. 1997. Perception of synthetic speech. In *Progress in speech synthesis*, ed. Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, 541-560. New York: Springer.

Pisoni, David B., and Sheri Hunnicutt. 1980. Perceptual evaluation of MITalk: The MIT unrestricted text-to-speech system. ICASSP, 572-575.

Poplack, Shana. 1980. The notion of the plural in Puerto Rican Spanish: Competing constraints on /s/ deletion. In *Locating language in time and space*, ed. William Labov, 55-86. New York: Academic Press.

Portele, Thomas. 1996. *Ein phonetisch-akustisch motiviertes Inventar zur Sprachsynthese deutscher Äußerungen.* Tübingen: Max Niemeyer Verlag.

Portele, Thomas. 1997. Reduktionen in der einheitsbasierten Sprachsynthese. Fortschritte der Akustik. DAGA '97, Kiel.

Press, William H., Saul A. Teukolsky, William T. Vetterling, and Brian P. Flannery. 1992. *Numerical recipes in C.* Cambridge: Cambridge University Press.

Price, Patti. 1996. Combining linguistic with statistical methods in automatic speech understanding. In *The balancing act*, ed. Judith L. Klavans and Philip Resnik, 119-134. Cambridge: MIT Press.

Prince, Alan, and Paul Smolensky. 1993. Optimality theory: constraint interaction in generative grammar. Manuscript.

Rabiner, Lawrence R., and Biing-Hwang Juang. 1993. *Fundamentals of speech recognition.* Englewood Cliffs, NJ: PTR Prentice Hall.

Ralston, James V., David B. Pisoni, and John W. Mullennix. 1995. Perception and comprehension of speech. In *Applied Speech Technology*, ed. A. Syrdal, R. Bennett and S. Greenspan, 233-287. Boca Raton: CRC. 223-287.

Ralston, James. V, David B. Pisoni, Scott E. Lively, B. G. Greene, John W. Mullennix. 1991. Comprehension of synthetic speech produced by rule: word monitoring and sentence-by-sentence listening times. *Human Factors* 33:471-491, 1991.

Randolph, Mark A. 1989. Syllable-based constraints on properties of English sounds. Ph.D. diss., MIT.

Randolph, Mark A. 1990. A data-driven method for discovering and predicting allophonic variation. ICASSP 90.1177-1180.

Rentzepopoulous, Panagiotis A., and George K. Kokkinakis. 1996. Efficient multilingual phoneme-to-grapheme conversion based on HMM. *Computational Linguistics* 22:351-376.

Reynolds, William T. 1994. Variation and phonological theory. Ph.D. diss., University of Pennsylvania.

Reynolds, William T., and Naomi Nagy. 1994. Phonological variation in Faetar: an optimality account. *Chicago Linguistics Society* 30-2.

Reynolds, William T., and Hadass Sheffer. 1994. Variation and optimality. *Penn Working Papers in Linguistics* 1.

Riley, Michael. 1989. Some applications of tree-based modeling to speech and language. In *Proceedings of the speech and natural language workshop,* 339-352. DARPA, Morgan Kaufmann.

Riley, Michael, 1991. A statistical model for generating pronunciation networks. ICASSP 91.S11.1-S11.4.

Riley Michael, and Andrej Ljolje. 1996. Automatic generation of detailed pronunciation lexicons. In *Automatic speech and speaker recognition: advanced topics,* ed. Chin-Hui Lee, Frank K. Soong, and K.K. Paliwal, 285-301. Boston: Kluwer.

Ristad, Eric Sven, and Peter N. Yianilos. 1997. Learning String Edit Distance. Princeton University Computer Science Department, Research Report CS-TR-532-96.

Roche, Emmanuel, and Yves Schabes. 1995. Deterministic part-of-speech tagging with finite-state transducers. *Computational Linguistics* 21.

Rodd, Jennifer. 1997. Recurrent neural network learning of phonological regularities in Turkish. In *CoNLL97: Computational Natural Language Learning,* ed. T. M. Ellison, 97-106. ACL.

Rumelhart, D. E., and J. L. McClelland. 1986. On learning the past tenses of English verbs. In *Parallel distributed processing,* Vol. 2, ed. James L. McClelland, David E. Rumelhart and the PDP Research Group, 216-271. Cambridge, Mass.: MIT Press.

Sankoff, David. 1988. Variable rules. In *Sociolinguistics: An international handbook of the science of language and society,* ed. Ulrich Ammon, Norbert Dittmar, and Klauss J. Mattheier. Berlin: de Gruyter.

Schmidt, M., S. Fitt, C. Scott and M. Jack. 1993. Phonetic transcription standards for European names (ONOMASTICA). *Eurospeech 93,* 279-282.

Schmidt-Nielsen, Astrid. 1995. Intelligibility and acceptability testing for speech technology. In *Applied Speech Technology,* ed. A. Syrdal, R. Bennett and S. Greenspan, 195-231. Boca Raton: CRC. 194-231.

Schwab, E.C., Howard .C. Nusbaum, and David B. Pisoni. 1985. Some effects of training on the perception of synthetic speech. *Human Factors* 27:395-408.

Scobbie, James M. 1995. What do we do when phonology is powerful to imitate phonetics? Comments on Zsiga. In *Phonology and phonetic evidence, papers in laboratory phonology* IV, ed. Bruce Connell and Amalia Arvaniti, 303-314. Cambridge: Cambridge University Press.

Seidenberg, Mark. 1989. Visual word recognition and pronunciation: a computational model and its implications. In *Lexical representation and process*, ed. William Marslen-Wilson. Cambridge: MIT Press.

Sejnowski, T., and C. Rosenberg. 1987. NETtalk: a parallel network that learns to pronounce English text. *Complex Systems* 1:145-168.

Selkirk, Elizabeth O. 1984. *Phonology and syntax.* Cambridge, Mass.: MIT Press.

Sells, Peter, John Rickford, and Thomas Wasow. 1996. Variation in negative inversion in AAVE: an optimality theoretic approach. In *Sociolinguistic variation: data, theory, and analysis: selected papers from NWAV 23 at Stanford.* Stanford: CSLI.

Seneff, Stephanie, and Victor Zue. 1988. Transcription and alignment of the TIMIT database. Manuscript.

Shillcock, Richard, Joe Levy, Geoff Lindsey, Paul Cairns, and Nick Chater. 1993. Connectionist modeling of phonological space. In *Edinburgh Working Papers in Cognitive Science* 8: *Computational Phonology*, ed. T. Mark Ellison and James M. Scobbie.

Sorin, Christel. 1991. Some observations on the processing of mute "e" in a French diphone-based speech synthesis system. *Journal of Phonetics* 19: 147-159.

Spiegel, Murray, and Marian Macchi. 1990. Development of the Orator synthesizer for network applications: name pronunciation accuracy, morphological analysis, customization for business listings, and acronym pronunciation. AVIOS.

Sproat, Richard. 1994. English noun-phrase accent prediction for text-to-speech. *Computer Speech and Language*, April, 1994: 79-94.

Sproat, Richard, and Osamu Fujimura. 1993. Allophonic variation in English /l/ and its implications for phonetic implementation. *Journal of Phonetics* 21:291-311.

Sproat, Richard, Bernd Möbius, Kazuaki Maeda, and Evelyne Tzoukermann. Multilingual text analysis. In *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*, ed. Richard Sproat, 31-87. Boston: Kluwer.

Sproat, Richard, and Michael Riley. 1996. Compilation of weighted finite-state transducers from decision trees. *34th Annual Meeting of the Association for Computational Linguistics*, 215-222.

Stanfill, C., and D. Waltz. 1986. Towards memory-based reasoning. *Communications of the ACM* 29: 1213-1228.

Strassel, Stephanie M. 1997. Variation in North American English flapping. Paper presented at NWAVE 26, Quebec.

Street, R. J., Jr. and H. Giles. 1982. Speech accommodation theory: a social and cognitive approach to language and speech behavior. In *Social cognition and communication*, ed. M. Roloff and C. Berger. Beverly Hills: Sage.

Ström, Nikko. 1997. Speaker Modeling for Speaker Adaptation in Automatic Speech Recognition. In *Talker Variability in Speech Processing,* ed. Keith Johnson and John W. Mullennix, 167-190. San Diego: Academic Press.

Syrdal, Ann K. 1995. Text-to-speech systems. In *Applied Speech Technology,* ed. A. Syrdal, R. Bennett, and S. Greenspan, 99-126. Boca Raton: CRC Press.

Tajchman, Gary, Daniel Jurafsky, and Eric Fosler. 1995. Learning phonological rule probabilities from speech corpora with exploratory computational phonology. ACL-95, 1-8.

Tesar, Bruce. 1995. Computing optimal forms in optimality theory: basic syllabification. Technical Report CU-CS-763-95, Dept. Of Computer Science, University of Colorado, Boulder.

Thomas, Charles Kenneth. 1958. *An introduction to the phonetics of American English*. 2nd ed. New York: Ronald Press.

Torkkola, Kari. 1993. An efficient way to learn English grapheme-to-phoneme rules automatically. IEEE.

Trager, George L., and Bernard Bloch. 1941. The syllabic phonemes of English. *Language* 17: 223-246.

Trager, George L., and Henry Lee Smith, Jr. 1956. *An outline of English structure.* Studies in linguistics: occasional papers, no. 3. 2nd printing. Washington, D.C.: American Council of Learned Societies.

Turkel, Bill. 1994. The acquisition of optimality theoretic systems. ROA-11-0000.

Urbanczyk, Suzanne C., and Stephen J. Eady. 1989. Assignment of syllable stress in a demisyllable-based text-to-speech synthesis system. IEEE.

van Bergem, Dick. R. 1991. Acoustic and lexical vowel reduction. In *ETRW: Phonetics and Phonology of Speaking Styles,* 10.1-10.5.

van Bergem, Dick R. 1993. Acoustic vowel reduction as a function of sentence accent, word stress, and word class.

van Bergem, Dick R. 1994. A model of coarticulatory effects on the schwa. *Speech Communication* 14:143-162.

Van de Velde, Hans, and Roeland van Hout. 1997. Dangerous aggregations: A case study of Dutch /n/ deletion. Paper presented at NWAVE 26, Quebec.

Veatch, Thomas C. 1991. English vowels: their surface phonology and phonetic implementation in vernacular dialects. Single-spaced version. Ph.D. diss., University of Pennsylvania.

Vitale, Tony. 1991. An algorithm for high accuracy name pronunciation by parametric speech synthesizer. *Computational Linguistics* 17.257-276.

Voiers, W. D. 1983. Evaluating processed speech using the diagnostic rhyme test. *Speech Technology* Jan./Feb.: 30-39.

Vorstermans, A., and J. P. Martens. 1994. Automatic labeling of corpora for speech synthesis development. Proceedings ProRisc-94, 261-266.

Waibel, Alexander. 1988. *Prosody and speech recognition*. London: Pitman; San Mateo: Morgan-Kaufmann.

Walther, Markus. OT SIMPLE - a construction kit approach to optimality theory. ROA-152-1096.

Ward, Grady. 1996. Moby Pronunciator II.

Weide, Robert L. 1995. The Carnegie Mellon Pronouncing Dictionary. cmudict.0.4.

Weinreich, Uriel, William Labov, and Marvin Herzog. 1968. Empirical foundations for a theory of language change. In *Directions for historical linguistics*, ed. W. Lehmann and Y. Malkiel, 95-195. Austin: University of Texas Press.

Wells, J. C. 1982. *Accents of English* I. Cambridge: Cambridge University Press.

Wightman, Colin W., and David T. Talkin. 1997. The Aligner: text-to-speech alignment using Markov models. In *Progress in speech synthesis*, ed. Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, 314-323. New York: Springer.

Wilks, Yorick, and Mark Stevenson. 1996. The grammar of sense: Is word-sense tagging much more than part-of-speech tagging? cmp-lg/9607028.

Williams, Briony, and Stephen Isard. 1997. A keyvowel Approach to the synthesis of regional accents of English. *Eurospeech '97*, 2435-2438.

Williams, Sheila M. 1994. Lexical phonology and speech style: Using a model to test a theory. In *Proceedings of the first meeting of the ACL special interest group in computational phonology*, 43-57.

Withgott, M. Margaret, and Francine R. Chen. 1993. *Computational Models of American Speech.* Stanford: CSLI.

Wood, Sidney A. J. 1996. Assimilation or coarticulation? Evidence from the temporal coordination of tongue gestures for the palatalization of Bulgarian alveolar stops. *Journal of Phonetics* 24: 139-164.

Woods, Anthony, Paul Fletcher, and Arthur Hughes. 1986. *Statistics in language studies.* Cambridge: Cambridge University Press.

Yarowsky, David. 1997. Homograph disambiguation in text-to-speech synthesis. In *Progress in speech synthesis*, ed. Jan P. H. van Santen, Richard W. Sproat, Joseph P. Olive, and Julia Hirschberg, 157-172. New York: Springer.

Yip, Moira. 1988. The obligatory contour principle and phonological rules: A loss of identity. *Linguistic Inquiry* 19:65-100.

Young, Steve, Joop Jansen, Julian Odell, Dave Ollason, and Phil Woodland. 1995. *The HTK Book*. Version 2.0. Washington: Entropic Research Laboratory.

Yvon, François. 1996. Grapheme-to-phoneme conversion using multiple unbounded overlapping chunks. Proceedings of NemLap '96.

Zhao, Yunxin. 1997. Overcoming speaker variability in automatic speech recognition: the speaker adaptation approach. In *Talker variability in speech processing*, ed. Keith Johnson and John W. Mullennix, 191-210. San Diego: Academic Press.

Zsiga, Elizabeth C. 1995. An acoustic and electropalatographic study of lexical and postlexical palatalization in American English. In *Phonology and phonetic evidence, papers in laboratory phonology* IV, ed. Bruce Connell and Amalia Arvaniti, 282-302. Cambridge: Cambridge University Press

Zsiga, Elizabeth C. 1997. Features, gesture and Igbo vowels. *Language* 73:227-274.

Zubritskaya, Katya. 1994. Sound change in optimality theory and constraint tie. *CLS* 30-2.

Zue, Victor W., and Martha Laferriere. 1979. Acoustic study of medial /t,d/ in American English. *Journal of the Acoustical Society of America* 66:1039-1050.