Syllable-based Constraints on Properties of English Sounds

RLE Technical Report No. 555

May 1990

Mark A. Randolph

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

Syllable-based Constraints on Properties of English Sounds

RLE Technical Report No. 555

May 1990

Mark A. Randolph

Research Laboratory of Electronics
Massachusetts Institute of Technology
Cambridge, MA 02139 USA

This work was sponsored in part by the Office of Naval Research under Contract No. N00014-82-K-0727.

Syllable-based Constraints on Properties of English Sounds

by
Mark Anthony Randolph

Submitted to the Department of Electrical Engineering and Computer Science on August 31, 1989 in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Abstract

This thesis outlines a phonological representation and corresponding rule framework for modelling constraints on an utterance's acoustic-phonetic pattern. The proposed representation and framework of rules are based on the syllable and suggested as an alternative to other representations that are primarily segment-based. Specifically, the traditional notion of a segment is abandoned at the systematic-phonetic level of representation and replaced with a description of an utterance in terms of acoustic properties. Acoustic properties are related directly to distinctive features which comprise an utterance's underlying phonemic description. This relation is specified in the form of a grammar. The grammar incorporates realization rules that allow acoustic properties to be described in direct quantitative terms. Further, properties that are associated with bundles of features are not required to temporally align in the sound stream.

Constraints on patterns that comprise the acoustic representation of an utterance are stated as conditions on the realization of well-formed syllables, where an immediate constituent grammar is used to represent syllable structure at the phonemic level. The details of the proposed rule framework are provided, as well as an empirical justification for the use of the syllable in the organization of acoustic-phonetic, phonotactic, and lexical constraints. Justification is provided in the form of results of three sets of experiments. In the first set of experiments, the role of the syllable in describing the distributions of the allophones of stop consonants in American English was examined. The hypothesis is that the syllable simplifies this description. Binary regression trees were used to quantify the extent to which knowledge of a stop's syllable position in conjunction with other contextual factors aids in the prediction of its acoustic realization. The principle of Maximum Mutual Information was used to construct trees. In the second experiment, syllabic constraints on the underlying phonemic description were examined. Specifically, mutual information was used to quantify collocational constraints within the syllable. A hierarchical cluster analysis was used to determine a syllable internal structure that is consistent with mutual information statistics gathered from a 5500-token sample of syllables extracted from a 20,000-word dictionary. Finally, in the third set of experiments, a model of lexical representation is proposed and tested. The model is based on the notion of constraint satisfaction, where syllabic constraints are applied prior to accessing the lexicon to derive a partial phonemic specification of an utterance. In the partial phonemic specification, selected constituents of the syllable were specified for their feature content. Lexicon partitioning experiments were performed to gain intuition as to whether such a lexical representation is justified, and to determined which of the constituents within the syllable are most informative in identifying a given word candidate. Mutual information is used as a means of evaluating lexicon partitions.

The results of these experiments indicate that knowledge of a stop's syllable position is the single most important factor among those considered in predicting its acoustic realization. Furthermore, we note that acoustic-phonetic constraints interact with constraints on an utterance's phonemic representation. For example, a stop is almost certain to be released when placed in the syllable-onset position, whereas in the syllable coda, a stop tends to be unreleased. However, when placed in the coda, a stop exhibits greater acoustic variation (it may also be released, glottalized, or deleted). Given that place-of-articulation and voicing features are well represented for released stops, this latter result suggests that the syllable coda is a less reliable source of phonetic information than the syllable onset. Results from lexical partitioning experiments, however, suggest that the coda is the least informative of the syllable's constituents in terms of providing information regarding the lexical identity of a syllable.

Thesis Supervisor: Dr. Kenneth N. Stevens
Title: Clarence J. LeBel Professor of Electrical Engineering

Thesis Supervisor: Dr. Victor W. Zue
Title: Principal Research Scientist

Acknowledgements

I would like to acknowledge the encouragement and guidance of all members of my thesis committee: Victor Zue, Ken Stevens, Patti Price, Donca Steriade, and Jay Keyser. Each is deserving of special thanks for taking time to read and provide comments on drafts of this thesis, and for the role they played in the completion of my studies as a graduate student.

I would like to express my sincerest gratitude to Victor and Ken for their enthusiasm and patience, and for agreeing to the odd arrangement of co-supervision. While the thesis has undoubtedly suffered in places from this arrangement, I could not conceive of a more enlightening and satisfying experience than having these two speech scientists as mentors and role models.

A special thanks goes to Patti Price. Her morale-boosting correspondence over the past few months and willingness to read practically anything I managed to put down on paper were principal factors responsible for my getting through the process of writing this document.

Thanks to the phonologists on the committee: Jay Keyser and Donca Steriade – Jay, especially for stepping in at the last minute. I would like to particularly acknowledge their comments and suggestions on those aspects of the thesis concerning linguistic theory. I, however, take responsibility for any errors in the presentation of this material.

I would also like to acknowledge present and past members of the MIT speech group for their role in the research leading to this thesis, and for providing a most stimulating and enriching place to study speech. The outstanding nature of this working environment is clearly a tribute to the group's leaders: Ken, Victor, Dennis Klatt, Joe Perkell, and Stefanie Shattuck-Hufnagel, and the students and staff members who comprise it. In particular, I would like to acknowledge the technical expertise of many who have worked long hours in providing computer hardware and software support. Dave Shipman, Scott Cyphers, David Kaufman, Rob Kassel, Hong Leung, Dave Goodine, Charles Jankowski, and Keith North have provided a speech workstation that is simply unsurpassed. Thanks also to Rob and Dave Whitney who have made document preparation on the Macintosh practically effortless, and who have patiently answered my questions.

Great advisors and facilities provide only a fraction of what makes a graduate student experience successful. One needs the constant support and encouragement of persons willing to take time from their busy schedules to talk about matters personal and social, as well professional and technical. I have had the great fortune to have Marc Cohen, Nancy Daly, Jim Glass, Katy Kline, Lori Lamel, Hong Leung, Sharon Manuel, Stefanie Shattuck-Hufnagel, and Lorin Wilde as colleagues belonging to this category. In particular, Stefanie's constant enthusiasm was enough to make even the most dreary of graduate student existence somehow seem rewarding.

Thanks also to friends and colleagues outside of the Speech Group who have, at various times, taken me under their wings and showed me the ropes. These include fellow graduate students: Chris Rose and Karl Wyatt, as well as Ron Schafer and Ron Crochiere.

I would like to acknowledge and thank my immediate and extended families for providing me with love, support, encouragement, and at times help with babysitting. These persons include: Linda, Beverly, and Yvonne; Auntie and Uncle Bill; all of the Aunts and Uncles; Chris, Linda, and Donna. Thanks to "Mommy Mary" and John who provided me with a home away from home, and to Mom and Dad, who instilled in me the importance of education, and worked as hard as they could to ensure I received the best.

Finally, my most heartfelt appreciation goes to my wife Kathy and my daughter Allyson (the epitome of the phrase "bundle of joy"). To Kathy, for providing the love that has been a beacon in my life. During those times when I was uncertain whether the light was at the end of the tunnel, I always had her by my side to light the way. To Allyson, for making that final year of graduate school bearable.

This research has been sponsored through the generous support of the AT & T Cooperative Research Fellowship Program, and in part by DARPA through the Office of Naval Research.

To Kath

Contents

List of Figures

List of Tables

Chapter 1

Introduction

A spoken language is a relation or mapping between the physical properties of speech sounds and their meanings. Although the vocal apparatus seems capable of emitting an indeterminate variety of sound patterns, not all will have conceptual significance for any given language. A theory of spoken language has the task of specifying the parameters of this mapping: the primitive elements of an utterance's description, and the principles by which the primitives are to be combined. In addition, through the adopted means of description, a language theory attempts to provide an account of the constraints that admit certain sound patterns as being part of a language while excluding others. In this thesis, part of such a theory is developed for American English. In particular, it proposes a descriptive framework whereby the relationship between the acoustic properties of an utterance and its underlying surface-phonemic description is to be explicitly stated. On the basis of this descriptive framework, a theory of acoustic-phonological representation based on the syllable is proposed.

A fundamental issue that a theory of acoustic-phonological representation confronts is the problem of variability in speech. In our investigation, we have restricted our attention to variations in the acoustic realizations of phonemes that are dependent on the contexts in which they appear. Our assumption is that, to a large extent, this variability is grammatically determined. That is, it may be captured by rule. One of the fundamental problems that we address is the specification of an appropriate framework of rules.

Traditionally (cf., Oshika *et al.*, 1975; Cohen and Mercer, 1975), systematic contextual variation has been described using a formalism of context-sensitive *rewrite rules* of the form:

R 1.1

$$\alpha \rightarrow \beta \;/\; \gamma_1 \,\text{-}\, \gamma_2.$$

(R 1.1) is to be read "α is rewritten as β if γ_1 is immediately to its left and γ_2 is immediately to its right." In the use of this rule framework in phonology, the elements α, γ_1, and γ_2 are strings of phonemes (of possibly zero length), or specifications of features. The element β is α's phonetic realization.

The rule formalism given in (R 1.1) has as its historical basis the descriptive framework outlined in Chomsky and Halle's *Sound Pattern of English* (1968) (hereafter abbreviated *SPE*). In the *SPE* framework, an utterance's phonological derivation is to take place over a series of stages. Rules of the form given in (R 1.1) map an utterance's phonological representation at one stage onto a corresponding representation at some later stage. At each stage, an utterance's phonological representation consists of a linear arrangement of segments and boundaries (Chomsky and Halle, 1968). Each segment is associated with a binary distinctive feature specification (or feature bundle), where features are used in forming lexical contrasts and for grouping segments into *natural classes* for the purpose of forming generalizations. Boundaries, usually morpheme and word boundaries, serve to partition the string into substrings that constitute possible domains of phonological generalization.

From the perspective of specifying a theory of acoustic-phonological representation, there are at least three disadvantages associated with rules of the form given in (R 1.1). Foremost is the problem of phonetic segmentation. *SPE* posited the systematic-phonetic level of representation to intervene between an utterance's physical properties and its more abstract form of description. Like segments comprising the more abstract levels of representation, elements of the systematic-phonetic representation are approximately the size of phonemes. Theories of acoustic-phonological representation based on this descriptive framework, implemented primarily as algorithms for automatic speech recognition, assume that this level of representation is to be obtained by segmenting the waveform into non-overlapping regions (cf., Broad and Shoup, 1975; Zue and Schwartz, 1980). As a consequence of this assumption, we will argue in Chapter 3 that rules of the form given in (R 1.1) are at a theoretical disadvantage in describing a number of coarticulatory processes. A second disadvantage of this rule formalism is that its underlying theoretical framework fails to adequately provide the basis for representing constituents (e.g., syllables, metrical feet, and other alternative phonological units). Finally, this framework of rules implicitly assumes that contextual variations can be described in categorical terms, despite the fact that many acoustic-phonetic changes are inherently continuous.

In this thesis, we outline an alternative acoustic-phonological representation and corresponding rule formalism for relating an utterance's surface-phonemic representation to its acoustic form. It differs from more traditional descriptive frameworks by positing an explicit *acoustic realization component* as part of the grammar. Realization rules map features directly onto their corresponding acoustic properties.

These rules are stated as conditions on well-formed English syllables. The position taken in this thesis is that acoustic-phonetic variability is constrained. Furthermore, we suggest that the acoustic shape of an underlying phonological form is determined by two sets of factors that are to be captured by constraints on syllable well-formedness. The first set of factors has to do with the intrinsic physical constraints imposed by the articulatory mechanism. Constraints on the articulatory process interact with the second set of factors, namely constraints that are imposed by the phonology of language. Within the set of language-specific factors, we distinguish three kinds of phonological constraints: constraints on the distributions of allophones, constraints on phoneme sequences (i.e., phonotactic constraints), and constraints on the lexicon.

1.1 Motivating the Syllable

Three reasons have traditionally been given in support of the syllable as being part of the phonological representation. The following set of statements is taken from Selkirk (1982):

> First, it can be argued that the most general and explanatory statement of phonotactic constraints in a language can be made only by reference to the syllable structure of an utterance. Second, it can be argued that only via the syllable can one give the proper characterization of the domain of application of a wide range of rules of segmental phonology. And third, it can be argued that an adequate treatment of suprasegmental phenomena such as stress and tone requires that segments be grouped into units which are the size of the syllable. The same three reasons leading to the postulation of the syllable can be shown to motivate the existence of privileged groupings of segments within the syllable which must be thought of as a constituent-like linguistic units themselves (p. 337).

The notion of syllable structure that has emerged from previous investigations is one of a hierarchical unit; an internally structured tree quite analogous to a tree representing syntactic structure. The constituent domains function to provide a set of structural conditions, over and above those of an utterance's morphological structure, for the concise statements of phonotactic and allophonic constraints.

More recently, efforts have been made to incorporate the syllable into frameworks that describe the phonetic representation of an utterance. Church (1983), for example, recodes context-sensitive rules such as those given in (R 1.1) into a context-free form. The resulting grammar encodes the syllable's immediate constituent structure, the leaf nodes of which are elements of the systematic-phonetic level of representation. Fujimura and Lovins (1978) and Fujimura (1981), as well as a number of other investigators (e.g., Browman and Goldstein, 1986) suggest the syllable as a unit of organization for articulatory gestures.

Our Contributions

Although these proponents of the syllable provide compelling arguments in favor of syllable-based phonological and phonetic representations, these theories are still in need of considerable qualitative and quantitative support. Quantitative support may come in many varieties. In this thesis we offer two kinds. First, for the class of stop consonants in American English, we provide results of experiments directed at correlating the surface acoustic realizations of these phonological segments with their positions within the syllable. Second, studies of large dictionaries and corpora of transcribed speech material are undertaken as a means of providing a quantitative basis for the topology of the syllable's internal structure and the role of syllabic constraints on the lexicon. Qualitative support is offered in the form of a grammatical framework incorporating the syllable. On the basis of this framework, constraints on an utterance's acoustic pattern are concisely stated and simplified.

1.2 Methodological Issues

There are two components to the methodology used in the present investigation – one addresses the phonological aspects of the syllable's representation, while the other is directed towards understanding the relationship between phonological syllables and their acoustic-phonetic realizations. In the former set of methods, an utterance is viewed as being the output of a highly constrained production system. The system of production incorporates physical constraints on articulation that are universal in nature, as well as allophonic, phonotactic, and lexical constraints that are language-specific. In certain phases of our investigation we have chosen to model this system statistically, as an *information source* that is to be characterized in terms of a probability measure. Principles of information theory (cf., Gallager, 1968) have provided a means of studying constraints, which in statistical terms manifest themselves as redundant information. A statistical approach is also used in understanding constraints on allophonic variation due to syllable structure. In Chapter 4, regression methods are developed for modelling the realization of a phoneme as a function of its syllable position, its local phonetic context, and its stress environment, as well as certain aspects of its internal feature specification. These regression methods have allowed us to arrive at a relative assessment of how valuable the syllable is in stating allophonic rules, as well as providing a basis for incorporating statistical models of the acoustic realization process as part of the grammar.

Our statistical approach is not without its limits, particularly when it has been applied to samples of speech materials of finite size. Technical limitations in the use of finite samples typically fall within the domain of statistical sampling theory, and are discussed at points in this thesis when particular methods are used.

There are theoretical limitations inherent in using a finite sample. A linguistic theory is intended to account for utterances that are possible in a language as well as those that are actually occurring. Although in the limit of an infinite sample, a statistical characterization of a language may be descriptively adequate, it may not be adequate in the sense of providing an explanation of a language's constraints. For this reason, this thesis has purposely sought to develop a theoretical framework for incorporating statistical observations.

There are limitations having to do with the use of acoustic data, as opposed to, for example, articulatory measurements. Acoustic descriptions are undertaken in this thesis for two principal reasons. First, we have a genuine interest in understanding the constraints on how acoustic properties of English are patterned. Second, as opposed to articulatory measurements, an important advantage of acoustic data is that they are plentiful. Large databases of speech material are becoming readily available (Lamel *et al.*, 1986) as well as computer software that allows them to be easily probed (see Appendix A). Therefore, in principle, it is possible to find enough data to build robust quantitative models.

The use of acoustic measurements as a basis for a theory requires a great deal of care. As Keating (1985) notes, the primitives of an acoustic description of an utterance may entail a range of candidates, not all of which will be linguistically relevant. Therefore, in order for a theory of acoustic-phonological representation to have scientific plausibility, one is required to enumerate a carefully selected set of acoustic properties that are representative of constraints on the speech production and perceptual mechanisms. Such an effort is made in Chapters 3 and 4, where we propose a tentative set of acoustic properties corresponding to a set of manner-of-articulation features that are assumed sufficient for the phonological description of American English.

1.3 Guide to the Thesis

The remainder of this thesis is organized as follows. Chapters 2 and 3 are theoretical in nature. Chapter 2 reviews previous theories of the syllable, and makes further suggestions of how syllabic constraints play a role in the description of the sound pattern of English. In Chapter 3, we describe our conception of the syllable and the role it plays in constraining an utterance's acoustic-phonological representation. These two chapters expand the position taken in Section 1.1 above.

Chapters 4 and 5 present experimental data and its analysis. In Chapter 4 allophonic variation is studied. First, regression techniques for the analysis of large databases of speech sounds are developed. Later in Chapter 4, these methods are applied in a case study involving stop consonants in American English. As described above, the experiments reported in Chapter 4 are directed towards assessing the role of syllable structure in the formulation of rules that predict acoustic-phonetic realizations of surface phonemes. Chapter 5 consists of lexical studies. Like its predecessor, Chapter 5 is divided into two major sections. First, methods are developed for studying constraints on the surface phonemic representation. These methods are then applied to assess the nature of the influence in stating constraints at the surface-phonemic level. Finally, in Chapter 6 we conclude.

Four appendices are provided. Appendix A provides the speech materials and databases used in Chapter 4 and 5. Appendix B provides some basic results of Information Theory. Appendix C presents in detail the syntax of our rule formalism. Finally, in Appendix D, we outline a parsing framework developed for testing the proposed framework of rules.

Chapter 2

Syllable Theory

This chapter provides background information on the phonetics and phonology of English syllable structure. The initial section (Section 2.1) examines a range of issues that a theory of syllable structure must consider. These issues include principles of syllable structure well-formedness, syllabification, and the role of the syllable in the statement of rules of phonetic realization. Sections 2.2 and 2.3 then address these issues, but from phonological and phonetic perspectives. In Section 2.2, the phonological conception of the syllable is illustrated. The phonological syllable is considered as a hierarchically structured object, in which phonemic segments are organized into an immediate constituent structure. Section 2.3 examines phonetic syllables. In particular, it focuses on the relationship between the syllable's phonological representation and its phonetic realization. Specifically, two models of phonetic realization are considered. The first, Church's proposal (1983), advocates the use of syllable structure to specify well-formedness conditions on the systematic phonetic level of representation. The second model is the framework for phonetic realization outlined by Fujimura and Lovins (1978). These authors propose the syllable as the principal means for organizing the temporal relations among articulatory gestures. Aspects of both of these proposals may be discerned in the acoustic theory of syllable structure proposed in this thesis and outlined in Chapter 3.

2.1 Issues in Syllable Phonology

The role of the syllable in the phonology of American English has been a topic of debate among linguists for several decades. Much of the controversy surrounding the status of this unit may be attributed to the ambiguous nature of the syllable's interaction with other aspects of linguistic description. In particular, it is not clear whether the syllable structure of an utterance should play a complementary role with respect to its underlying morphological specification, or whether these two forms of representation are in opposition. Kohler (1966), for example, argued that a division of an utterance into syllables was either unnecessary, since syllable boundaries often coincided with divisions between other grammatical formatives (e.g., words and morphemes), or arbitrary, since phonological and phonetic principles of syllabification had yet to be adequately defined. It has also been noted that the word "syllable" does not even appear in the index of *SPE.* As Anderson (1985) points out,

> ...The absence of syllables, etc., from phonological representations was not, as some have suggested, a matter of oversight or ignorance on Chomsky and Halle's part. Rather, it constituted a principled decision: insofar as all generalizations apparently requiring reference to units other than segments could be encoded in terms of segments alone without significant loss of generality, the more limited theory constituted a stronger empirical claim about the nature of language (p 347).

More recently, there have been a number of compelling arguments in favor of the syllable as a phonological unit (cf., Fudge 1969; Hooper 1972; Kahn 1976; Selkirk 1982; and Clements and Keyser 1983). Proponents of the syllable claim that words and morphemes are not the most appropriate domains of application for a large number of phonological processes. They argue, for example, that a more general and explanatory description of stress assignment and allophonic variation may only be obtained by reference to syllable structure. Furthermore, they suggest that the syllable is a necessary construct in the statement of phonotactic restrictions.

For the remainder of the present section, three issues raised in this debate will be examined in more detail. First, we address the problem of defining the syllable.

Although an adequate definition of the syllable is still a subject of active research, most theories of the syllable accept some notion of a sonority sequencing principle as a fundamental aspect of syllable-structure well-formedness (Clements, 1988). Conditions on syllable well-formedness constitute the basis of algorithms for syllabification. Syllabification is the second of the topics considered in this section. Finally, the role of syllable structure in the statement of rules of phonetic realization is examined.

2.1.1 Defining The Syllable

It will be useful in a discussion of the definition of the syllable to make a distinction between between *phonological* or *phonemic* syllables and *phonetic* syllables. Fudge (1969) also makes such a distinction; in his view, a phonological syllable assumes a segmental level of representation. That is, syllables are defined in terms of an organization of phonemes. In contrast, phonetic syllables are atomic.

2.1.1.1 Syllables as Concatenative Phonetic Units

Defining the syllable in strictly phonetic terms has proven difficult. Ladefoged (1975) reviews a number of objective definitions of the phonetic syllable in strictly physical terms, and summarizes the complexity of the problem:

> Although nearly everybody can identify syllables, almost nobody can define them. If I ask you how many syllables there are in "minimization" or "suprasegmental" you can easily count them and tell me. In each of these words there are five syllables. Nevertheless, it is curiously difficult to state an objective measure for locating the number of syllables in a word or a phrase (p. 218)."

Similar sentiments have been expressed by Price (1980).

Phonetic definitions of the syllable have ranged from those that specify the unit in terms of the muscular activities of the respiratory system, to perceptually based definitions based on a notion of *prominence* (usually defined in terms of signal "loudness"). Strictly acoustic definitions of the syllable have also been noted. Mermelstein (1975), for example, has attempted to provide a precise acoustic definition of signal loudness, by defining an algorithm that band-pass filters the speech signal, emphasizing the region of the spectrum between 500 and 4000 Hz. The resulting band-pass waveform is searched for "dips" and "peaks". All definitions which address the physical aspects of the syllable have thus far proven less than satisfactory.

The difficulty with viewing the syllable as a concatenative phonetic unit is that the definition itself must take into consideration a wide range of factors that are responsible for segmental (or within-syllable) phonetic variability. For example, the word *meal*, which many speakers consider a monosyllable, contains a high-front vowel followed by an /l/. Speakers often produce this word as two syllables with an intervening /y/. A /y/ may also be inserted in the words *neolithic* and *mediate*, where a vowel-vowel sequence exists and the first vowel is unstressed. On other occasions, speakers will merge this vowel-vowel sequence into one syllable.

At times, severe phonetic reductions involving syllables may occur. In some instances, the resulting surface phonetic form is a sequence of well-formed English syllables, where a well-formed syllable is one which begins and ends in valid phonetic sequences. For example, the phonetic reductions that sometimes occur in *multiply* (→ [mʌltplɑ^y]) and *police* (→ [pli^{y}s]) produce sequences of well-formed syllables. In contrast, the surface phonetic forms produced by the reductions occurring in *Toledo* (→ [tli^{y}do^w]) and *phonetic* (→ [fnɛɾɪk]) are invalid by the above principle.

Finally, Ladefoged (1975) notes an interaction between syllable reduction by speakers and the need to mark differences in the morphological or lexical structure of a sentence. For example, the monomorphemic word *hire* may be pronounced as two syllables except when speakers want to maintain a distinction between it and the bimorphemic word *higher*. Similarly, the word *err* will be be realized as one syllable if there is a particular need to maintain the distinction between it and the word *error*, which has similar meaning.

In summary, the definition of the phonetic syllable is based on the tacit assumption of invariance within the syllable's boundaries, and therefore is limited in its capability to account for a wide range of both within-syllable and across syllable boundary phonetic variability. This is not to imply syllables lack reality at the physical level. As an alternative to defining the syllable in direct physical terms, one could arrive at a definition of the phonetic syllable indirectly by first defining this unit at the abstract phonological level, and then carefully constructing a model of phonetic realization. This approach is pursued in this thesis and is based on the framework proposed by Fujimura and Lovins (1978) outlined in Section 2.3.

2.1.1.2 A Phonological Theory of the Syllable Based on Sonority

The phonological definition of the syllable is based on the notion of *sonority*: a phonological feature whose phonetic correlates are difficult to ascertain, but nonetheless, whose abstract definition provides a basis for an adequate definition of the syllable at the surface phonemic level.

According to Ladefoged (1975), "an *adequate* phonological definition of the syllable must account for words where there is agreement among listeners on the number of syllables that it contains. In addition, it must also explain words where there is disagreement." *Adequacy*, in this context has to do with explanatory adequacy. A theory of the syllable must explain the intuitions of native speakers of the language in determining the number of syllables there are in an isolated utterance. Furthermore, a definition of the syllable in terms of sonority, as will be seen below, allows one to posit a link between the phonological syllable, which is to be defined in abstract terms, and its physical realization, the representation of the syllable with which listeners and speakers of the language have direct contact.

Phonologists have used the concept of a *sonority hierarchy* or *scale* to define the syllable. In devising a sonority scale, the feature [*sonorant*] plays a dual role. On the one hand, it has its usual phonological connotation: that of a binary feature that distinguishes the class of "sonorants" from "obstruents". Sonorants are sounds produced with the source of vocal-tract excitation at the glottis and relatively little pressure built-up inside the vocal tract, while obstruents are produced with a significant obstruction to airflow. On the other hand, based on intuitions concerning the physics of speech production, sonority is also assumed to be a property that ranges over an ordinal scale having more than two levels. On the basis of this scale, phonological segments are ranked. An example of this ranking is given in (R 2.1).

R 2.1 (Sonority Scale)

Sounds:	*Sonority Value:*	*Examples:*
low vowels	*10*	/a, ɔ/
mid vowels	*9*	/e, o/
high vowels	*8*	/i, u/
flaps	*7*	/ɾ/
laterals	*6*	/l/
nasals	*5*	/m, n, ŋ/
voiced fricatives	*4*	/v, ð, z/
voiceless fricatives	*3*	/f, θ, s/
voiced stops	*2*	/b, d, g/
voiceless stops	*1*	/p, t, k/

The sonority scale given in (R 2.1) is taken from Hogg and McCulley (1987) and is typical of others found in the literature (e.g., Selkirk, 1984; Fujimura, 1975). In this particular scale, there are 10 sonority values. Segments that are at the top of the scale are vocalic sounds and are considered "most sonorant". Segments at the bottom are the "least sonorant". Among vocalics and obstruents, principles relating to the articulatory correlates of the feature [*sonorant*] are used to sort classes of sounds. For example, the vowels most sonorant are open (i.e., having the feature [+ *low*]). Within obstruents, the segments least sonorant are stops, realized with a complete closure and having the feature [- *continuant*]. Further, voiced obstruents are considered more sonorant than voiceless obstruents; vocalics that are continuant (i.e., liquids and glides) are considered more sonorant than non-continuants (i.e., nasals).

As we will discuss in Section 2.2, there are numerous ways of factoring principles of sonority into a phonological representation of the syllable. All these methods implement the following principle of sonority sequencing:

R 2.2 (Sonority Sequencing Principle)

Within any syllable, there is a segment constituting a sonority peak that is preceded and/or followed by a sequence of segments with progressively decreasing sonority values

The Sonority Sequencing Principle (SSP) is a well-formedness condition on syllables (Clements, 1988; Selkirk, 1984). It, along with a table of sonority values such as the one given in (R 2.1), allows one to analyze a phoneme string and determine the number of syllables that it contains. This analysis can also account for variation in the number of syllables heard by a listener, which is a desired aspect of the syllable's definition. For example, as noted above the words *suprasegmental* and *minimization* represent relatively simple cases for which listeners can determine the number of syllables. An analysis of their sonority patterns is given in (R 2.3)

R 2.3 /suprʌsɛgmɛn təl/ → *[3-8-1-7-9-3-8-2-5-9-5-1-9-6]*
/mɪnɪmɑ^{y}ze^{y}šən/ → *[5-8-5-8-5-10-4-9-3-9-5]*

In (R 2.3), the sequences of sonority values corresponding to the phonemic representations of these words are shown in square brackets. In constrast to the two cases above, determining the number of syllables in *hire* and *error* is more difficult.

R 2.4 /hɑ^{y}(y)r/ → *[6-10-(8)-9]*
/ɝ(r)ɝ(r)/ → *[9-(7)-9-(7)]*

The difficulty lies in the occasional insertion of a semivowel. In (R 2.4), the sonority value of this optional segment is shown in parenthesis.

2.1.1.3 Necessary Refinements to a Sonority-based Syllable Theory

Although sonority as a phonetic property is in need of further research and clarification, it nonetheless provides a substantial basis for a definition of the syllable at the phonological level. However, as one examines phonological syllables defined in terms of sonority a bit further, it becomes apparent that the SSP alone is an insufficient constraint on syllable well-formedness. That is, not all isolated sequences of phonemes that obey the SSP constitute possible syllables in English. In addition, without refinements to the sonority-based definition of the syllable, violations of SSP may rule out possible monosyllables in the language. For example, the SSP does not explain why native speakers of English are willing to accept the word *blick* [2-6-8-1] as part of English while rejecting *vnuk* [4-5-9-1], although both have valid sonority sequences. Furthermore, the sonority-based theory would reject the valid monosyllables *stick* [3-1-8-1] and *six* [3-8-1-3].

In the past, three refinements have been applied to a syllable theory in order to accommodate these examples. The first refinement has been to establish a basis for stating more detailed principles regarding permissible consonant sequences. Such principles would admit /bl/ as a valid syllable onset while ruling out /vn/. The second refinement has been to admit monosyllables such as *stick* as either exceptions to the theory or incorporating principles that would allow them to be considered well-formed. The latter approach is the one that is most typically adopted. For example, a number of theories consider /sp/, /st/, /sk/-clusters as single segments (e.g., Selkirk, 1982). Therefore, having these clusters as syllable onsets does not represent a violation of the SSP. In addition, the third refinement has been to posit *extrasyllabic* positions (sometimes called *affixes*) to handle cases such as *six* (/sɪks/) that ends in a consonant sequence that does not obey the SSP.

2.1.2 Principles of Syllabification

The syllable structure well-formedness conditions based on sonority explain a native speaker's ability to determine the number of syllables in an isolated utterance. Adding a set of more detailed phonotactic restrictions allows the theory to explain a native speaker's ability to determine whether an isolated word is possible in the language. In addition, syllable well-formedness conditions provide the basis of a set of principles for syllabification: the division of the sound stream into syllables and the internal representation of each syllable.

There are three aspects of the syllabification problem that a theory of syllable structure must somehow address. First, there is the problem of specifying the point in an utterance's derivation where syllabification occurs. Is it a lexical process, or does syllabification occur post-lexically? Second, evidence seems to suggest that principles of syllabification interact with rules of stress assignment (Selkirk, 1982; Clements and Keyser, 1983). A theory of syllable structure must determine the exact nature of this interaction. Finally, there are the mechanical details of the syllabification procedure itself, and the means by which the resulting syllabification of an utterance is to be displayed.

The first of these questions is largely beyond the scope of this thesis. The reader may consult Clements and Keyser (1983) and Mohanan (1985), along with the relevant references cited therein, for a discussion of this matter. We simply assume that rules that syllabify the lexical representation of an utterance apply at some point prior to its phonetic implementation. Therefore, rules of phonetic implementation, which we assume to be sensitive the locations of syllable boundaries, may apply without conflict.

In most theories of syllable structure, the interaction between stress assignment and syllabification is typically stated in terms of the following two principles.

R 2.5

a) *With the exceptions noted as (b), segments are assigned to syllables so that well-formedness conditions are satisfied, while maximizing syllable onsets.*

b) *Segments legally assigned to either of two syllables, are associated with the previous syllable if it is stressed.*

(R 2.5a) is known as the *maximal onset principle.* This principle is widely accepted as universal in its application in languages (Bell and Hooper, 1978). Rule (R 2.5b) is often referred to as the *principle of stress resyllabification.* It reflects a regularity in English, for example, for stressed vowels to attract post-vocalic consonants.

The two principles given in (R 2.5) explain the differences in syllabification between the word pairs, *[pa][trol]* vs. *[pet][rol]*, and *[or][chestral]* vs. *[orch][estra]*. The first syllable of the second word of each pair is stressed. (R 2.5b) explains the placement of the syllable boundary after the /t/ in the second word of the first word pair, and after the /k/ in the second word of the second word pair.

2.1.3 Syllabic Constraints on Phonetic Realization

The final set of issues that a syllable theory must address is the specification of the syllable's role in the statement of rules of phonetic realization. In this thesis, we distinguish two classes of realization rules: rules that describe extrinsic (allophonic) variation, and rules which describe intrinsic (coarticulatory) variation. Extrinsic variation in the realization of a phoneme is conditioned on its position within a larger phonological constituent. Therefore, it has been suggested by a number of investigators that the placement of allophones in the sound stream serve as prime markers of syllable, word, and larger morpho-syntactic or prosodic boundaries (Lehiste, 1960; Christie, 1974; Nakatani and Dukes 1977).

Intrinsic variation, on the other hand, encompasses a range of phenomenon otherwise known as coarticulation. It is a consequence of the multi-dimensional nature of the articulatory mechanism. That is, vocal tract articulators (the lips, tongue, jaw, glottis, and velum) are loosely connected. Although to a large extent their movements are coordinated, articulatory gestures may vary with respect to one another in their spatio-temporal orientation. As a result of this variability, the phonetic information pertaining to one phonological segment in the sound stream may overlap with that of surrounding segments. The suggestion that is offered below, and in the following chapter, is that certain types of coarticulatory phenomenon (e.g., vowel nasalization) often seen as problematic in models of phonetic realization, may be seen in a different light when analyzed in terms of a syllabic organization of articulatory dimensions.

2.1.4 Summary of Issues

In summary, well-formedness conditions on phonological syllables may be precisely stated. In constrast, phonetic syllables as atomic units are ill-defined. Thus, a theory of the syllable should consist of two components: a *phonological component* and a *realization component*. The phonological component would provide the basis for stating well-formedness conditions on syllables at the surface-phonemic level of representation. These conditions would specify constraints on sonority sequencing as well as phonotactic constraints. The phonological component would also specify: 1) when in an utterance's derivation syllabification occurs, and 2) the interaction between rules that assign syllable structure to an utterance and rules that assign stress. The realization component would specify constraints on the articulatory mechanism and structure-dependant allophonic constraints.

2.2 The Phonological Representation of Syllable Structure

In this section, we will discuss, by way of illustration, one view of the phonological syllable, in particular, that which has been advanced in Clements and Keyser (1983) (cf., Steriade, 1988 for a critical review of this theory).

2.2.1 CV Theory

In recent years, there has been a trend in phonology towards the view of an utterance having several separate, but simultaneous, levels of phonological representation, each placed on its own individual "tier". This *non-linear* approach to phonological description contrasts with the view held in *SPE* of segments concatenating to form morphemes and words, which in turn concatenate to form syntactic constituents. The modern conception of the phonological representation has served as the principal basis for stating phonological theories of the syllable. CV theory, for example, is the conception of the syllable proposed by Clements and Keyser (1983). We will use CV theory as a vehicle for illustrating how the phonological syllable is defined.

CV theory posits four principal tiers for representing the syllable: 1) the *CV*-tier, 2) the *segmental*-tier, 3) the *nucleus*- or ν-tier, and 4) the *syllable*- or σ-tier. Clements and Keyser view their theory as one of syllabic organization over the elements of the *CV*-tier. The elements *C* and *V* of this tier serve a dual purpose. First, they comprise the skeleton of this phonological representation. Second, they denote a major class division of the elements of the segmental tier. That is, segments having the feature value [+ *syllabic*] are assigned to *V* elements, and segments that are [- *syllabic*] are assigned to elements denoted *C*. Except for phonotactic restrictions, CV theory claims that well-formedness conditions on syllables which reference syllable internal structure need only distinguish these the two categories of phonological segments.

The remaining two tiers are linked to members of the *CV*-tier according to a specified set of well-formedness conditions. In the case of the ν-tier, its elements are related to the CV-tier according to the association principles given in (R 2.6):

R 2.6

$$\begin{array}{ccc} \nu & \leftrightarrow & V \\ & \leftrightarrow & VV \\ & \leftrightarrow & VC \end{array}$$

In (R 2.6), the arrows indicate the elements of these two tiers that may legitimately be linked.

One may note from the association principles stated in (R 2.6) that a single element of the ν-tier may be associated with more than one element of the CV-tier.[1] In particular, these principles state that ν-tier elements may be linked to at least one, but no more than two, CV-tier elements. If linked to the ν-tier, the first CV-tier element must be a V.

In the case where an element of the ν-tier is linked to two CV elements, it is said that the syllable's nucleus is "branching". This situation occurs when a syllable contains a long vowel (denoted *VV*) or when the syllable is closed by a post-vocalic consonant (denoted *VC*). In either of these situations, the syllable may bear primary stress.

Well-formedness conditions specifying the link between the σ-tier and the CV-tier are stated as *syllable structure conditions.* Syllable structure conditions are the mechanism in this framework for encoding phonotactic restrictions, and as alluded to above, make reference to both a syllable's CV-tier and its *segmental*-tier. CV theory posits two types of these conditions. *Positive syllable structure conditions*

[1] This is a general property of nonlinear phonological representations: the mapping between tiers may link any number of elements of one tier to any number of elements of another as long as the lines of association do not cross (cf., Goldsmith, 1979; van der Hulst and Smith, 1982).

specify phoneme sequences that are permissible in the language, and *negative syllable structure conditions* specify phoneme sequences that are invalid. We will illustrate below by providing some examples.

The first example is what is referred to as the *Basic Onset Condition.* It is presented here as (R 2.7):

R 2.7 (Basic Onset Condition)

$$
{}_{\sigma}\left[\begin{array}{cc} \mathrm{C} & \mathrm{C} \\ | & | \\ \left[- \ \text{sonorant} \right] & \begin{bmatrix} + & \text{sonorant} \\ - & \text{nasal} \end{bmatrix} \end{array} \right.
$$

The notation used to state syllable structure conditions is to partially enclose elements of both the *CV*-tier and the *segmental*-tier in a left square bracket. Below and to left of the square bracket, the symbol σ indicates that the condition applies to the syllable as opposed to some other category (e.g., the nucleus).

Rule (R 2.7) implements the sonority sequencing principle. It does so by requiring a two consonant syllable-onset sequence to have a rising sonority contour. That is, it admits /tr/, /sw/, and /fl/, as well-formed syllable onsets, but, for example, not */mg/, */rt/, or /st/ (where the * preceding a phoneme sequence in this instance denotes "impermissible").

Syllable-initial clusters involving /s/ in CV theory are handled by a separate syllable structure condition, specifically, that which is given in (R 2.8):

R 2.8 (s-Clusters)

$$
{}_{\sigma}\left[\begin{array}{cc} \mathrm{C} & \mathrm{C} \\ | & | \\ \begin{bmatrix} + & \text{strident} \\ + & \text{coronal} \\ + & \text{anterior} \end{bmatrix} & \left[- \ \text{continuant} \right] \end{array} \right.
$$

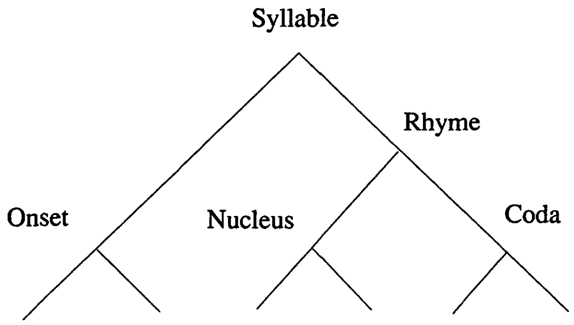


Figure 2.1: Syllable Template (after Selkirk, 1982).

One may note that /s/-stop clusters are not treated as single segments.

The syllable onset conditions given in (R 2.7) and (R 2.8), only restrict sonority patterns. Thus, they overgenerate the set of permissible syllables. CV theory posits negative syllable structure conditions like that given in (R 2.9) to act as "filters":

R 2.9

$$ {}^{*}\left[\begin{array}{cc} \text{C} & \text{C} \\ | & | \\ \begin{bmatrix} + & \text{coronal} \\ - & \text{strident} \end{bmatrix} & \begin{bmatrix} + & \text{lateral} \end{bmatrix} \end{array} \right]_{\sigma} $$

This particular condition rules out coronal-lateral sequences in the syllable onset. The fact that it rules out such sequences is indicated by the presence of the "*" to the left and above the left bracket. In Chapter 3 we will extend the use of the term filter to cover both positive and negative syllable structure conditions, and we will introduce our own notation for stating filters.

2.2.2 More Enriched Hierarchical Structures

The number of tiers used to represent the internal structure of syllables is a question that is yet to be completely settled (cf., Steriade, 1988, and Chapter 5). Where

CV theory posits a much flatter internal structure between the CV-tier and the σ-tier, the syllable structure depicted in Figure 2.1 is considerably more enriched. In subsequent discussion, we will adopt the terms *immediate constituent structure* to refer to the syllable internal organization shown in Figure 2.1, and the term *template* to refer to the grammar describing its structure.

The template shown in Figure 2.1 is taken from Selkirk (1982). Its immediate constituent is primarily motivated by principles of phonotactics. Specifically, the labels ONSET, CODA, and RHYME shown as constituents in the hierarchical structure denote domains over which restrictions on phoneme sequences apply. Based on notation from Clements and Keyser (1983), the conditions stated as (R 2.10) and (R 2.11) are examples of restrictions applying to the syllable's onset and coda respectively:

R 2.10

$$ {}^{*}\left[\begin{array}{cc} \text{C} & \text{C} \\ | & | \\ \begin{bmatrix} + & \text{coronal} \\ - & \text{strident} \end{bmatrix} & \begin{bmatrix} + & \text{lateral} \end{bmatrix} \end{array}\right]_{\text{Onset}} $$

R 2.11

$$ \left[\begin{array}{cc} \text{C} & \text{C} \\ | & | \\ \begin{bmatrix} + & \text{nasal} \\ \alpha & \text{labial} \\ \beta & \text{coronal} \end{bmatrix} & \begin{bmatrix} - & \text{continuant} \\ \alpha & \text{labial} \\ \beta & \text{coronal} \end{bmatrix} \end{array}\right]_{\text{Coda}} $$

Rule (R 2.10) is a negative condition on the syllable onset that filters coronal-lateral sequences. Rule (R 2.11) is a positive condition on the syllable coda that enforces the so-called *nasal-stop homogamic rule* in English.[2]

[2] Rule 2.11 is different from conditions previously seen in that it incorporates variable coefficients for features. This is a typical linguistic notation for specifying conditions on feature agreement.

2.2.3 Discussion

We view the issue of whether to adopt constituents such as the ONSET, the CODA, and the RHYME as an empirical question. The basis for making a decision whether to adopt a constituent as part of the syllable's representation will stem from considerations related to *grammatical realization*: the effort to incorporate a grammar into a model of language use, for example, a parser. In a parsing model, one may assign a measure of computational complexity to a grammar by determining the number of processing steps required to assign structure to an utterance. Specifically, parsing in the case of a grammar of the syllable could be performed using a paradigm of "overgenerate and filter" (see Appendix D). That is, structures that correspond to syllabic constituents are constructed and then tested against the syllable structure conditions stated as filters. Such an analysis would favor a grammar whose rules consist primarily of local constraints: constraints that apply over a relatively small number of phonemes. The justification for favoring local constraints would be that if a substring were considered ill-formed, then less effort would have been required in constructing it. In this particular case, a syllable structure representation incorporating the ONSET, CODA and RHYME would be favored, if they could be shown to be linguistically significant. In Chapter 5, we will address this problem.

2.3 Phonetic Syllable Well-formedness

It has been argued that syllables, as concatenative phonetic units, are ill-defined because they are unable to capture a range of within-syllable segmental phonetic alternations. Thus, the approach that has been advanced is to specify the syllable as an abstract, phonological object, and then construct a model of phonetic realization. In this section, two approaches to specifying phonetic syllables in this manner are reviewed: a framework proposed by Church (1983), and a model advanced by Fujimura and Lovins (1978).

Both of these frameworks treat the syllable at the phonological level as a hierarchically structured object. However, these two frameworks have differing conceptions of the representation of an utterance at the phonetic level. Church posits a syllabic organization over sequences of systematic phonetic segments (or allophones). In other words, allophones comprise the terminal categories of the syllable's immediate constituent grammar. In contrast, Fujimura and Lovins propose the syllable as a framework for the organization of articulatory gestures at the physical level, completely bypassing any segmental form of phonetic description. The spirit of the latter approach is similar to the framework adopted in this thesis.

2.3.1 Syllabic Constraints on the Systematic Phonetic Representation

Several studies (Lehiste, 1960; Christie, 1974; and Nakatani and Dukes,1977) have shown that certain forms of allophonic variation are not only acceptable, but provide useful information about the structure of a message. These processes include stop aspiration, the realizations of the various allophones of the phoneme /l/, and the insertion of glottal stops at the onsets of phonetic syllables that begin with vowels. A more complete review of these studies is provided in Chapter 4 when the subject of allophonic variation is again raised.

Church (1983) offers a concrete proposal for dealing with these facts, and for using allophonic variation as a source of constraint on phonological parsing and the lexical retrieval tasks of an automatic speech recognition system. He does so by positing two kinds of phonetic features as part of the systematic phonetic representation: 1) features which he assumes to be relatively invariant (e.g., manner, place, and voicing), and 2) features that appear to be susceptible to considerable contextual variation (e.g., stop aspiration). The suggestion is that both kinds of features play a role in phonological parsing and lexical retrieval. Specifically, he argues that invariant features are useful for lexical retrieval while variant features mark the prosodic structure of an utterance including its syllable structure. In general, we support Church's position. However, we take issue with some specific aspects of his approach.

Church's suggestion is to make modifications in the phonological representation so that allophonic constraints may be readily exploited. These modifications amount to 1) factoring features into the variant and invariant categories, and 2) positing constituents in the grammar that serve as domains over which allophonic processes may occur. These constituents are then arranged into syllable and metrical structure by the rules of an immediate constituent grammar much like the one already seen. The following example illustrates how allophonic processes involving stop consonants are treated in this framework.

Rules (R 2.12) and (R 2.13) provide the transformational account of stop aspiration in American English.

R 2.12

$$\left\{\begin{array}{c} p\\ t\\ k\end{array}\right\} \longrightarrow \left\{\begin{array}{c} p^h\\ t^h\\ k^h\end{array}\right\} / \ \$\underline{\hspace{1.5em}}\mathrm{V}\cdots\$$$

R 2.13

$$\left\{\begin{array}{c} p\\ t\\ k\end{array}\right\} \longrightarrow \left\{\begin{array}{c} \mathrm{p}^{\square}\\ \mathrm{t}^{\square}\\ \mathrm{k}^{\square}\end{array}\right\} / \ \$\cdots\underline{\hspace{1.5em}}\$.$$

The *left-hand sides* of the rules are surface phonemes; the elements on the rules' *right-hand sides* are elements of the systematic phonetic representation (or allophones). In particular, the diacritics ("^{h}") and ("$^{\square}$") denote that a stop is aspirated and is unreleased respectively. In words, (R 2.12) states that stops are aspirated when syllable initial, whereas (R 2.13) states stops are unreleased when syllable final.

Church, making use of suggestions made earlier by Kahn (1976), recognizes the syllable-initial and syllable-final contexts as specified by the conditions of the context-sensitive rules given in (R 2.12) and (R 2.13) as corresponding exactly to the ONSET and CODA constituents of the syllable's hierarchical structure. He then formulates the above rules into the following immediate constituent grammar:

R 2.14

$$\begin{array}{lcl}
\textsc{Utterance} & \longrightarrow & \textsc{Syllable}^{*} \\
\textsc{Syllable} & \longrightarrow & \textsc{Onset Rhyme} \\
\textsc{Rhyme} & \longrightarrow & \textsc{Peak Coda} \\
\textsc{Onset} & \longrightarrow & p^{h}|t^{h}|k^{h} \\
\textsc{Nucleus} & \longrightarrow & \textsc{Vowel} \\
\textsc{Coda} & \longrightarrow & p^{\square}|t^{\square}|k^{\square}
\end{array}$$

The rules of the grammar given in (R 2.14) correspond to the elements of the syllable template shown in Figure 2.1.

Of immediate concern is that Church has defined the syllable at the systematic phonetic level of representation. The principle disadvantage of doing so is that the definition of the syllable now relies on phonetic segments that are smaller and more difficult to precisely define.

Alternatively, one could conceive of the syllable as providing an organization over phonological features which themselves correspond to phonetic properties. The indirect relationship between phonetic properties and the syllable is the second method by which the phonetic syllable may be defined.

2.3.2 A Phonetic Syllable Definition Based on Features

The principal proponents of a feature-based phonetic representation have been Stevens and his colleagues (cf., Stevens, 1986). The potential advantages of a feature-based description that they cite are its relative parsimony in the description of the phonetic realization of an utterance as well as its ability to simplify the explanation of certain low-level segmental phonological processes. In particular, with respect to the description of allophonic processes, the claim is that although phonological segments may undergo drastic phonetic alternations, most of the crucial features remain intact.

A feature-based acoustic-phonetic representation incorporating principles of syllable structure has yet to be fully explored. In Chapter 3 we will outline our conception of how such a representation might be organized. Some of the basic principles may be discerned in a framework proposed by Fujimura and Lovins (1978).

Although directed more towards a description of the speech production process, Fujimura and Lovins suggest that there are three components to a feature-based syllable framework: 1) an underlying set of distinctive features, 2) a set of ordering constraints, and 3) a set of realization rules. The role of distinctive features, like their role in phonology in general, is to provide enough information to distinguish one syllable from another. These features may be partitioned into three subsets, C_o, V, and C_c, corresponding to the syllable onset, nucleus, and coda respectively. Features within these groups are unordered; the ordering among groups is specified by phonotactic restrictions and the principle of sonority sequencing (or in the terms that Fujimura and Lovins label as *vowel affinity*). Finally, the realization rules specify the sequential ordering of articulatory gestures within a syllable. They are assumed to be language specific and their temporal aspect also reflects vowel affinity. An example given by Fujimura and Lovins (1978) is the feature specification of the monosyllabic word *limp*,

[+ LATERAL; + HIGH, + FRONT; +NASAL, + LABIAL, + TENSE],

where the semicolons in this specification separate feature groups.[3]

There are several interesting points to be noted about this representation. First, syllable phonotactics allow a number of simplifications in the underlying feature specification. In this example, for instance, only one place feature need be specified for the C_c position. This is a consequence of the nasal-stop homogamic rule in English, which maintains that both the nasal and the stop must agree in place of articulation (see above). Secondly, it is assumed that the realization rules are sensitive to syllable structure. They will ensure, for example, the production of the syllable-initial allophone of the [+ LATERAL] segment in the C_o position, as opposed to the syllable-final

[3] The use of the feature [*tense*] reflects the articulatory basis of this phonological description.

allophone which would result if this feature were assigned to the C_c position. Finally, the realization rules may also be used to capture intrinsic allophonic variation due to feature spreading. For example, studies (e.g., Malecot, 1960) have shown that in sequences in which post-vocalic nasals are followed by consonants that are in the C_o position, the vowel will be heavily nasalized if the consonant is [+ TENSE].

Fujimura and Lovins have noted a number of theoretical advantages provided by this framework in addition to those already cited. For example, at the phonetic level, it describes the complicated intra-syllable integration of articulatory and acoustic events (e.g., feature spreading). In addition, although it is not a phonemic representation, it does allow phonemes to be accounted for through the syllable's feature specification and the nonredundant statement of sequential constraints. On the other hand, there are still a number of details that either have been omitted from this discussion or require resolution. For example, Fujimura and Lovins do not elaborate the syllable concatenation process. In addition, the nature of the realization rules has yet to be specified.

2.4 Summary

We have provide background information on the phonetics and phonology of English syllable structure. Specifically, we have examined a range of issues that a syllable theory must consider. These issues include principles of syllable structure well-formedness, syllabification, and the role of the syllable in the statement of rules of phonetic realization. The syllable is conceived as both a phonological and phonetic object. Moreover, we have suggested that an explicit realization component be included in a grammatical description of the syllable for relating these two aspects of its representation. Our framework for specifying and implementing this grammatical framework are given in the remaining chapters, as is justification for the approach that we have taken.

Chapter 3

The Proposed Framework

Having sketched the goals of a syllable theory within a more general phonological framework in the previous chapter, we now shift attention to the proposed theory of the syllable. In the current proposal, an utterance's surface phonemic representation is comprised of distinctive features; features are binary-valued and align into columns to form phonological segments. Restrictions on possible surface phonemic forms are stated in terms of constraints on well-formed phonological syllables. The proposed theory of the syllable allows the expression of constraints on an utterance's acoustic representation. At the acoustic level, an utterance is comprised of a pre-defined set of properties, where acoustic properties are a direct realization of features and consist of three broadly defined types. The first class of properties are termed *acoustic autosegments*. Autosegments reflect manner of articulation and are represented on autonomous tiers. Properties in the second class are present in the sound stream at the transitions of autosegments. For the most part, transitional properties indicate place of articulation. Finally, the third class of properties represents the timing of events in the sound stream, and reflects in some cases the voicing characteristics of phonological segments, and in certain other instances, an utterance's prosodic structure. We propose that the mapping between an utterance's surface phonemic representation and its acoustic form is grammatically determined. In this chapter, we propose a framework of realization rules that determine this mapping and suggest how the syllable plays a role in their statement.

This chapter is organized into two main sections, each corresponding to a component of the theory. Section 3.1 describes the nature of syllabic constraints on an utterance's phonological representation. In this section, we present the syllable template that has been adopted and introduce rules for assigning syllable structure to the surface-phonemic representation of an utterance. Section 3.2 discusses the details of the proposed acoustic representation and how conditions of acoustic syllable structure well-formedness arise from constraints on phonological syllables.

3.1 Phonological Syllables

The suggestion made in Chapter 2 was that phonetic syllables are to be defined as phonological objects first, and then mapped onto an utterance's phonetic representation. This is the position taken in this thesis: the present chapter sketches the logical structure of a syllable theory at both the phonological and acoustic-phonetic levels. Phonological syllables are defined in terms of distinctive features. Conditions that describe a well-formed syllable at this level of representation are stated in terms of an *augmented context-free grammar.* The core of this grammar is a set of rules that define the syllable's immediate constituent structure. These rules are augmented with constraints on sonority and phonotactics. For the remainder of this section, the form of this grammar and its substance are presented. The notation used to state constraints on syllable structure well-formedness is introduced. Using this notation, we describe the details of the theory itself. Towards the end of this section, the problem of syllabification is addressed, and a framework for assigning syllable structure to an utterance is outlined.

3.1.1 The Grammar's Core: the Syllable Template

As a first approximation, constraints on an utterance's surface phonemic representation may be concisely stated in terms of the production:

R 3.1

$$\textsc{Utterance} \longrightarrow \sigma^{+},$$

which states that an utterance is one or more well-formed syllables. In (R 3.1), the category "SYLLABLE" is denoted by the symbol "σ". The *Kleene star* ("+") symbol is to be read "one or more".

Rule (R 3.1) is deemed an "approximation" to constraints on the surface phonemic form of an utterance because there is a range of phenomena that syllable structure is not well suited to describe. Constraints on length, sonority, and phonotactics, however, fall well within the domain of a syllable theory. The rule given in (R 3.1), accompanied by the additional set of productions given in (R 3.2) for describing syllable internal structure, form a concrete basis for stating such a theory.

R 3.2

$$\begin{array}{rcl}
\sigma & \longrightarrow & \textsc{Core}\ (\textsc{Affix}) \\
\textsc{Core} & \longrightarrow & (\textsc{Onset})\ \textsc{Rhyme} \\
\textsc{Affix} & \longrightarrow & \textsc{Affix-1}\ (\textsc{Affix-2})\ (\textsc{Affix-3}) \\
\textsc{Onset} & \longrightarrow & \textsc{Outer-Onset} \mid \textsc{Inner-Onset} \mid \\
 & & \textsc{Outer-Onset}\ \textsc{Inner-Onset} \\
\textsc{Rhyme} & \longrightarrow & \textsc{Nucleus}\ (\textsc{Coda}) \\
\textsc{Coda} & \longrightarrow & \textsc{Inner-Coda} \mid \textsc{Outer-Coda} \mid \\
 & & \textsc{Inner-Coda}\ \textsc{Outer-Coda}
\end{array}$$

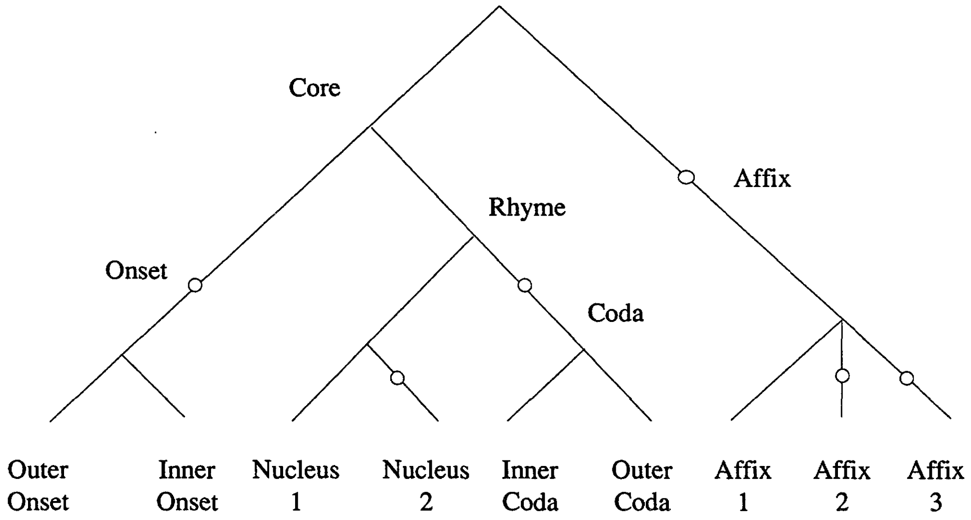


Figure 3.1: Syllable template corresponding to the grammar given in (R 3.1)

In (R 3.2), optional constituents are surrounded by parenthesis, and the symbol ("—") is to be read "or".

The rules given in (R 3.2) are a completely equivalent form of representation to the template shown in Figure 3.1. This template is similar to the one introduced in Chapter 2. Constituents of the grammar represented as "left-hand sides" of rules in (R 3.2) are the non-terminal nodes in the template's tree-like structure. Optional constituents are represented by having "open circles" placed on branches that connect them to their dominating constituents.

This template is principally the one proposed by Fudge (1969). It is also similar to others cited in the literature.[1] The syllable category in the grammar (i.e, σ) dominates a CORE and an AFFIX. These constituents eventually dominate a set of terminal categories, each of which is assigned a label (e.g., OUTER-ONSET, INNER-ONSET, AFFIX-1, etc.). In the proposed framework, it is useful to think of the

[1] See, for example, the template proposed by Selkirk in Chapter 2. Following Fujimura and Lovins (1978), a syllable AFFIX position is posited as well as a syllable CORE.

terminal categories of this grammar as "slots", i.e., placeholders that are filled by feature specifications. We have assigned labels to terminals so that we may readily refer to their contents in the statement of phonological rules.

Features at the surface phonemic representation are *simultaneous* in the sense that they may be viewed as "bundling" together to form phonological segments for the purpose of maintaining lexical contrasts and stating constraints on the length of consonant clusters. However, features are not unordered. In particular, they group into manner-of-articulation, place-, and voicing categories as described below. In addition, the terminal slots of the syllable have no physical significance (e.g., intrinsic duration). In fact, the features associated with the terminal positions of the syllable are not required to maintain their associations at the physical level.

3.1.2 Features and the Statement of Syllable Structure Conditions

In order to present the current proposal for representing the syllable we will need to define a representative set of features. The intent is not to present the current inventory as definitive, but simply to have a concrete set with which to work. On the basis of this inventory of features, we will illustrate how the current descriptive framework states constraints on sonority sequencing and phonotactics.

3.1.2.1 Inventory of Features

The inventory of distinctive features that we assume is summarized in (R 3.3).

R 3.3

Place of Articulation	*Manner of Articulation*	*Voicing*
high	*continuant*	*voiced*
low	*sonorant*	
back	*strident*	
round	*nasal*	
tense		
labial		
coronal		
anterior		
velar		
lateral		
retroflex		
distributed		

As previously mentioned, features are grouped into three categories (i.e., according to manner, place, and voicing).

The basis for the categorization of features given in (R 3.3) is two-fold. The first pertains to the role that features play in the statement of syllable structure conditions at the surface phonemic level. The second has to do with the manner in which features are phonetically realized. For the specification of phonological syllables, manner-of-articulation features are used for stating constraints on sonority, while place and voicing features play a more significant role in stating more detailed constraints on phonotactics. Furthermore, one may also point to the "more universal nature" of constraints on sonority sequencing, in that the majority of the world's languages obey something akin to the sonority sequencing principle (Clements, 1988), whereas phonotactic constraints tend to be more language specific. As we will describe in Section 3.2, we assume a difference among the various categories of features in terms of their implementation in the sound stream. Specifically, manner features correspond to *regions* of the sound stream and are associated with modes of speech production at which the vocal tract assumes some degree of constriction. In contrast, place features may be associated with properties that are more instantaneous in nature, i.e., implemented at specific points in time. Finally, the voicing feature is manifested in terms of the durations between acoustic events at which the degree of constriction changes.[2]

3.1.2.2 Sonority Sequencing Principles

As pointed out in the previous chapter, the sonority sequencing principle (SSP) is central to the definition of the syllable. In particular, this principle states that, within any syllable, there is a segment constituting a sonority peak that is preceded and/or followed by a sequence of segments with progressively decreasing sonority values as one moves away from the peak. The sonority peak is the syllable's NUCLEUS, which in the grammar given above is the only obligatory terminal constituent.

The SSP prescribes a specific assignment of features to the terminal constituents of the syllable's grammar. Using the notation for stating production rules, the assignment principles adopted in the current theory are stated in (R 3.4):

R 3.4

OUTER-ONSET ⟶ STOP | STRONG FRICATIVE |
WEAK FRICATIVE | AFFRICATE

INNER-ONSET ⟶ NASAL | SEMIVOWEL

NUCLEUS ⟶ VOWEL(VOWEL)

INNER-CODA ⟶ NASAL | SEMIVOWEL

OUTER-CODA ⟶ STOP | STRONG FRICATIVE |
WEAK FRICATIVE | AFFRICATE |
NASAL | SEMIVOWEL

In (R 3.4) the categories STOP, FRICATIVE, VOWEL, etc., are phonological categories that pertain to the surface-phonemic level of representation. The are simply substitu-

[2] We should add that certain aspects of our thinking concerning features and their acoustic properties are similar to ideas advanced by Stevens (1986).

tions for bundles of manner-of-articulation features. The values of features to which these categories correspond are given in the following table:

R 3.5

Categories			*Features*		
	Syllabic	*Sonorant*	*Continuant*	*Strident*	*Nasal*
Vowel	+	+	+	–	–
Semivowel	–	+	+	–	–
Nasal	–	+	–	–	+
Strong Fricative	–	–	+	+	–
Weak Fricative	–	–	+	–	–
Stop	–	–	–	–	–
Affricate	–	–	–	+	–

By making reference to the phonological categories defined in (R 3.5), the rules listed in (R 3.4) elaborate principles such as the *basic onset condition* included in the CV theory (Clements and Keyser, 1983; see Chapter 2). In particular, obstruents are restricted to the syllable's Outer-onset and Outer-coda positions. Sonorant consonants are assigned the positions next to the syllable's nucleus (i.e., the syllable's Inner-onset and Inner-coda positions). Finally, these constraints require that the Nucleus be a vowel.

As prescribed by the SSP, sonority increases and falls in the Onset and Coda of the syllable. On the other hand, affixes, corresponding to morphological suffixes, and /sp/, /st/, /sk/ clusters that may occur at the beginnings of words represent exceptions to the SSP. The assignment of segments to the affix positions is specified by the rule stated in (R 3.6):

R 3.6

$$\text{Affix} - i \longrightarrow \text{Stop} \mid \text{Fricative}.$$

(R 3.6) states that the affix positions must contain obstruents that may be either stops or fricatives.

The assignment of a segment to the affix represents a violation of the SSP. However, the occupants of the affix are constrained in other ways (see the discussion on phonotactic principles given below). The other exceptions to the SSP, /sp/, /st/, and /sk/ clusters, are treated as single segments as is the practice in other syllable theories (cf., Selkirk, 1982) and are allowed to occupy the OUTER-ONSET and OUTER-CODA slots.[3]

3.1.2.3 Phonotactic Filters

Phonotactic restrictions represent the second type of syllable structure condition included in the proposed framework. These conditions on sound sequences at the surface-phonemic level are stated as collocational constraints (i.e., restrictions on the terminal occupants of the syllable template). In the proposed framework, these conditions are stated as *filters*, where filters apply to constituents within the syllable or to the SYLLABLE itself.[4]

Formally, filters are implemented as *implications*, where each individual filter is stated as an *attribute-value matrix*[5] consisting of four attributes: 1) the name of the syllabic constituent that the filter applies to, 2) an *antecedent*, 3) a *consequent*, and 4) a *polarity*. An example is provided in (R 3.7):

[3] We should note that the assignment of /st/ clusters to the OUTER-CODA implies that the proposed grammar is ambiguous. The ambiguity stems from the fact that the /t/ may also occupy the AFFIX-1 position in certain instances. Although one may conceive of strategies for resolving this ambiguity, our efforts to do so will not be addressed.

[4] Most of the filters required to define the phonological syllable will apply to either the ONSET, the CODA, or the RHYME. Exceptions include collocational restrictions between the OUTER-CODA and the AFFIX-1.

[5] The details of the formalism used to state constraints in the grammar are summarized in Appendix C. Readers unfamiliar with the use of attribute-value structures may find a summary there.

R 3.7 (Strident Fricative Before Nasal)

$$\left[\begin{array}{ll}
\textit{constituent:} & \textsc{Onset} \\
\textit{antecedent:} & \left[\begin{array}{l}
\left[\begin{array}{ll}\textsc{Outer-onset} & \left[\begin{array}{ll}\textsc{Continuant} & + \\ \textsc{Strident} & +\end{array}\right]\end{array}\right] \\
\left[\begin{array}{ll}\textsc{Inner-Onset} & \left[\begin{array}{ll}\textsc{Nasal} & +\end{array}\right]\end{array}\right]
\end{array}\right] \\
\textit{consequent:} & \left[\begin{array}{ll}\textsc{Outer-onset} & \left[\begin{array}{ll}\textsc{Coronal} & + \\ \textsc{Anterior} & + \\ \textsc{Voiced} & -\end{array}\right]\end{array}\right] \\
\textit{polarity:} & +
\end{array}\right]$$

The filter given in (R 3.7) applies to the syllable's onset, where the filter's antecedent specifies a set of "pre-conditions" which the onset must satisfy if the filter is to apply. For example, (R 3.7) applies if the ONSET dominates inner and outer positions, if the OUTER-ONSET is a strident continuant, and if the INNER-ONSET is a nasal. In such a case, as indicated by the *consequent*, the filter requires that the OUTER-ONSET be coronal, anterior, and unvoiced (i.e., an /s/). The polarity of this filter is positive (indicated by the "+") which means that, when the above conditions apply to a constituent, the constituent is considered well-formed.

One may note some similarity between the framework for stating filters proposed here and the syllable structure conditions that are part of CV theory. In particular, the filter stated in (R 3.7) is precisely CV theory's *positive syllable structure condition* for s-clusters (Clements and Keyser, 1983; p 46). Aside from the obvious notational differences, there are two more substantial differences between the two frameworks. The first is in the level of specificity. Given the richer internal structure attributed to the syllable in the proposed theory, filters may apply to smaller syllabic constituents, such as, in this example the ONSET. Second, CV theory posits the categories *C* and *V* to partition the set of phonological segments that filters apply to into consonants and vowels. This categorization is analogous to the manner-of-articulation categories posited in the proposed framework, and occupies the value of the filter's antecedent.

Moreover, the proposed number of categories currently adopted is five instead of just two.

As in the Clements and Keyser framework, there are also negative syllable structure conditions in the proposed theory. For example, the filter given in (R 3.8) restricts syllable onsets from containing coronal-lateral sequences (i.e., *tl, *dl, *θl):

R 3.8 (No Coronal Lateral)

$$\begin{bmatrix} \textit{constituent:} & \textsc{Onset} \\ \textit{antecedent:} & \begin{bmatrix} \begin{bmatrix} \textsc{Outer-Onset} & \begin{bmatrix} \text{STRIDENT} & - \end{bmatrix} \end{bmatrix} \\ \begin{bmatrix} \textsc{Inner-Onset} & [\mathrm{T}] \end{bmatrix} \end{bmatrix} \\ \textit{consequent:} & \begin{bmatrix} \begin{bmatrix} \textsc{Outer-Onset} & \begin{bmatrix} \text{CORONAL} & + \end{bmatrix} \\ \textsc{Inner-Onset} & \begin{bmatrix} \text{LATERAL} & + \end{bmatrix} \end{bmatrix} \end{bmatrix} \\ \textit{polarity:} & - \end{bmatrix}$$

The symbol [T] indicates that a phoneme of any feature specification applies in this particular part of this rule.

The filters included in this section are intended as illustrations and are representative of the kinds of constraints on phonological syllables employed in the current framework. Before discussing how filters are applied during syllabification, we will provide an additional filter as an example, this time illustrating one of the consequences of adopting a hierarchical representation of the syllable and employing filters that apply to constituents.

In the discussion of the syllable affix above, it was mentioned that the segments assigned to the affix positions were not required to obey principles of sonority sequencing but were restricted in other ways. One of the restrictions is that each member of the affix must be either a coronal fricative or stop. In addition, each affix must agree in voicing with the Outer-coda, if the Outer-coda is an obstruent. A filter enforcing this constraint is given in (R 3.9):

R 3.9

$$\left[\begin{array}{ll}
\textit{constituent:} & \textsc{Syllable} \\
\textit{antecedent:} & \left[\begin{array}{l} \left[(\textsc{Core Rhyme Coda Outer-Coda}) \; \left[\textsc{sonorant} \quad - \right]\right] \\ \left[(\textsc{Affix Affix-1}) \; [T] \right] \end{array}\right] \\
\textit{consequent:} & \left[\begin{array}{l} \left[(\textsc{Core Rhyme Coda Outer-Coda}) \; \left[\textsc{voiced} \quad \alpha \right]\right] \\ \left[(\textsc{Affix Affix-1}) \; \left[\textsc{voiced} \quad \alpha \right] \right] \end{array}\right] \\
\textit{polarity:} & +
\end{array}\right]$$

There are two aspects concerning this filter that are worth noting: the first is the filter contains the variable α. Variables are a descriptive mechanism first employed in Chapter 2 for stating agreement constraints. (R 3.9) is a constraint on voicing agreement that applies between the Outer-coda and Affix-1. The second aspect is that the filter applies to the Syllable. The basis for incorporating constituents into the grammar is that they serve as domains over which certain processes occur, agreement being an example. The Syllable is the smallest constituent in the grammar that dominates both the Outer-coda and Affix-1 positions. Therefore, it is the constituent to which the agreement constraint between these two terminal slots is to apply. The syllable constituent, however, does not *immediately dominates* either of these two terminals. Therefore, a mechanism called a *path* is employed. In the antecedent, the path,

(Core Rhyme Coda Outer-coda),

is employed to state that the Outer-coda must contain an obstruent; the path points to the feature value [- *sonorant*]. The second path used in the antecedent,

(AFFIX AFFIX-1)

ensures that the first affix position is occupied. The consequent of this filter employs variables to ensure that the values for the feature [*voiced*] that is assigned to both constituents is the same.

The length of a path may be viewed as a metric. A grammar consisting entirely of rules requiring long paths like the one that is employed for stating agreement among the AFFIX-1 and the OUTER-CODA positions suggests a rearrangement of constituents in the grammar. If it were not the case that there are more constraints that apply to the CODA as a constituent than those that apply to the OUTER-CODA and AFFIX-1, then a new constituent immediately dominating these latter two positions would need to be created and employed. In Chapter 5 this point is addressed again but in more quantitative terms.

3.1.3 Phonological Syllabification

We conclude our discussion of phonological syllables with a description of the process of syllabification: how syllable structure is assigned to an utterance, and how an utterance's syllable structure is displayed. The syllable structure of an utterance is constructed by a process whereby constraints on phonological syllables is satisfied. In Section 3.2 we introduce constraints on phonetic syllables. There, the present set of ideas are extended to the problem of syllabification when the input to the procedure is an utterance's acoustic representation.

3.1.3.1 The Syllabic Sketch

It is useful to think of syllable structure as being constructed in stages. At the end of each stage, an intermediate form of description for an utterance is formed. We coin the term “syllabic sketch” to describe these intermediate forms of description, of which two are distinguished as having major significance.

The *initial* syllabic sketch is the first of these two significant forms of an utterance’s description. It is formed through a process which satisfies constraints on sonority sequencing. For example, the initial syllabic sketch for the monosyllable *strengths* is given in (R 3.10):

R 3.10

$$
\left[\begin{array}{ll}
\textit{constituent:} & \textsc{Syllable} \\
\textit{conditions:} & \left[\begin{array}{ll}
\textsc{Core} & \left[\begin{array}{ll}
\textsc{Onset} & \left[\begin{array}{ll}
\textsc{Outer-onset} & \left[\begin{array}{l}
\left[\begin{array}{ll} \textsc{Continuant} & + \\ \textsc{Strident} & + \end{array}\right] \\
\left[\begin{array}{ll} \textsc{Continuant} & - \\ \textsc{Strident} & - \end{array}\right]
\end{array}\right] \\
\textsc{Inner-onset} & \left[\begin{array}{ll} \textsc{Continuant} & + \end{array}\right]
\end{array}\right] \\
\textsc{Rhyme} & \left[\begin{array}{ll}
\textsc{Nucleus} & [T] \\
\textsc{Coda} & \left[\begin{array}{ll}
\textsc{Inner-coda} & \left[\begin{array}{ll} \textsc{Continuant} & - \\ \textsc{Nasal} & + \end{array}\right] \\
\textsc{Outer-coda} & \left[\begin{array}{ll} \textsc{Continuant} & - \\ \textsc{Strident} & - \end{array}\right]
\end{array}\right]
\end{array}\right]
\end{array}\right] \\
\textsc{Affix} & \left[\begin{array}{ll}
\textsc{Affix-1} & \left[\begin{array}{ll} \textsc{Continuant} & + \\ \textsc{Strident} & - \end{array}\right] \\
\textsc{Affix-2} & \left[\begin{array}{ll} \textsc{Continuant} & + \\ \textsc{Strident} & + \end{array}\right]
\end{array}\right]
\end{array}\right]
\end{array}\right]
$$

One may observe from the form given in (R 3.10) that the syllabic sketch is also implemented as an attribute-value matrix, much like the representation used for phonotactic filters. We call an attribute-value matrix used to represent an utterance its *functional description.* A formal syntax for functional descriptions is given in Appendix C.

Attribute-value matrices that describe the surface-phonemic representation of an utterance consist of a *constituent* attribute whose value is the name of the constituent that the entire structure represents, and a *conditions* attribute whose value describes the constituent's internal structure. The value of the *conditions* attribute is itself an attribute-value matrix. Further, the ability to embed attribute-value matrices inside one another provides the means for representing an utterance's hierarchical structure on an functional description.

The functional description of the word *strengths* shown in (R 3.10) has a number of characteristics worth noting. First, the word itself contains both an Affix and an /st/ cluster at the word's onset. Both are exceptions to the SSP. The value for the Affix attribute of the structure given in (R 3.10) is comprised of two attribute-value matrices, one describing the contents of the Affix-1, and the other describing the contents of Affix-2. The values corresponding to these attributes are specifications for the features [*continuant*] and [*strident*]. Both the /s/ and /t/ of the word-initial /st/ cluster are assigned to the Outer-onset position. This is reflected by having the value for the Outer-onset contain two feature bundles. One bundle is a feature specification for the /s/, the other for the /t/.

In general, the feature specifications included in an utterance's functional description will contain the minimal number of features. This is to avoid redundancy. For example, in the Outer-onset value of (R 3.10), only the features [*continuant*] and [*strident*] need be specified. The manner-of-articulation feature [*sonorant*], for example, is implied. Its value is minus (–) in this syllable position. Similarly, the feature [*continuant*] is not specified in the syllable's Inner-coda position. The feature value [+ *nasal*] is specified in that position, and nasals in English are redundantly [- *continuant*]. Finally, in the effort to rid the syllabic representation of redundant information, the Nucleus is specified as [T], since it is implicitly [+ *syllabic*].

At the phonological level, the initial syllabic sketch of an utterance is a underspecified representation reflecting only its manner of articulation. Put another way, it is the utterance's functional description satisfying restrictions on phonological syllables specified by rules (R 3.2) and (R 3.4) given above. The word *strengths* shown in (R 3.10), for example, satisfies principles of sonority sequencing and thus it comprises a well-formed syllable at this stage of analysis. To our knowledge, it is the only word satisfying the feature specification shown in (R 3.10). Typically, the number of syllables having the feature specification comprising an utterance's initial syllabic sketch will be significantly greater (see Chapter 5).

The *fully-specified syllabic sketch* is the second of the two significant levels of description mentioned above. For phonological syllables, the fully-specified syllabic sketch is formed by satisfying phonotactic constraints. Phonotactic filters may apply to any of the syllable's constituents. For example, the nonpermissible monosyllable /θlæd/ whose fully-specified syllabic sketch is shown in (R 3.11) will be considered ill-formed by the ONSET constraint given in (R 3.8):

R 3.11

$$\begin{bmatrix} \textit{constituent:} & \textsc{Syllable} \\ \textit{conditions:} & \begin{bmatrix} \textsc{Core} & \begin{bmatrix} \textsc{Onset} & \begin{bmatrix} \textsc{Outer-onset} & \begin{bmatrix} \textsc{Continuant} & + \\ \textsc{Strident} & - \\ \textsc{Coronal} & + \\ \textsc{Voiced} & - \end{bmatrix} \\ \textsc{Inner-onset} & \begin{bmatrix} \textsc{Continuant} & + \\ \textsc{Lateral} & + \end{bmatrix} \end{bmatrix} \\ \textsc{Rhyme} & \begin{bmatrix} \textsc{Nucleus} & \begin{bmatrix} \textsc{High} & - \\ \textsc{Tense} & - \\ \textsc{Low} & + \\ \textsc{Back} & - \end{bmatrix} \\ \textsc{Coda} & \begin{bmatrix} \textsc{Outer-coda} & \begin{bmatrix} \textsc{Continuant} & - \\ \textsc{Strident} & - \\ \textsc{Labial} & - \\ \textsc{Coronal} & + \\ \textsc{Anterior} & + \\ \textsc{Voiced} & + \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix} \end{bmatrix}$$

Once again, the general principle employed in producing a fully-specified syllabic sketch such as the one shown in (R 3.11) is that it will contain the minimal number of features that are non-redundant. In this particular case, other features have been left unspecified in order to shorten the description. The features that are most pertinent, for example, are the place features [+ *coronal*] in the OUTER-ONSET and [+ *lateral*] in the INNER-ONSET. These are the characteristics of the syllable's onset that make it ill-formed. For the sake of completeness, other place features have also been specified.

The actual process of syllabification consists of constructing representations such as (R 3.11) and testing them against all of the constraints in the grammar. In this particular case, the test would have failed.

3.1.4 Discussion

In concluding this section on phonological syllables, there are a number of points of a more formal nature that one should consider. Theoretically, the core grammar, although written in context-free form, describes a *finite-state language.* This is another way of saying that the set of well-formed syllables in English is enumerable. Such an enumeration, however, would fail to capture in its description the nature of syllable internal structure and the regularities that syllables in the language possess.[6] Second, because many of the grammar's rules are capable of deriving more than one sequence of categories, the grammar is nondeterministic. With regards to its usefulness in stating the relations among the syllable's internal constituents, nondeterminism does not represent a problem. During parsing, however, one may wish to encode certain preferences on the rules (perhaps in the form of probabilities) in order to facilitate efficient processing. Furthermore, the grammar as stated is *ambiguous.* For example, according to (R 3.4) and (R 3.6) stop consonants may be assigned to one of several terminal categories. Since the grammar states well-formedness conditions on phoneme sequences, it does not possess a way of determining which of these syllable positions is preferred. In Chapter 4, constraints on allophones that may be incorporated into the grammar are described. These additional constraints permit ambiguities to be resolved in many cases.

[6] There is a rich literature concerned with the theory of formal languages. The reader may consult Lewis and Papadimitriou (1981), as a general reference that describes, for example, differences between *finite-state* and *context-free* languages.

3.2 The Grammar's Realization Component

In the current section, the acoustic *realization* of phonological syllables is described. In the proposed framework, it is through the realization component of the grammar that the syllable, at the acoustic level of representation, is defined. We propose that the syllable is an organization over both an utterance's acoustic properties, and its distinctive features. Having spelled out the nature of the syllabic organization of distinctive features in the previous section, we now turn to a description of syllabic constraints on the pattern of acoustic properties.

The assumption that underlies the organization of distinctive features at the surface phonemic level is that feature values align into columns, and the resulting feature columns are assigned to terminal positions of the syllable's hierarchical structure. If, however, acoustic properties are associated with features, one then may observe that such an alignment is not always to be found at the acoustic level. In particular, following Stevens (1986), the suggestion is that only certain features are implemented simultaneously in the sound stream, while others are asynchronous.

Our approach to specifying the acoustic correlates of features is to begin by making a formal distinction among the various types of features. The distinction is based on how features are implemented in the sound stream. In particular, we postulate that manner-of-articulation features, with the exception of [*continuant*], are associated entirely with "acoustic autosegments". As the name implies, acoustic autosegments correspond to regions of the waveform, and are represented on autonomous tiers. Furthermore, acoustic autosegments, in general, are allowed to overlap temporally. Examples of this type of acoustic property are presented below.

Place-of-articulation features are implemented simultaneously, principally at the transitions (i.e., onsets and offsets) of acoustic autosegments, or other pre-defined landmarks. For example, the place-of-articulation features for a vowel may be ascertained by examining its spectrum at the midpoint of a syllabic region. Thus, the term *transitional properties* is used to denote the class of properties associated with place of articulation . Finally, it is suggested that the feature [*voiced*] is manifested in terms of the durations between pairs of autosegment transitions, where the two corresponding autosegments are not necessarily represented on the same tier. For example, the Voice Onset Time (or VOT) of a pre-stressed syllable-initial stop is often associated with the phonological feature [*voiced*] (Lisker and Abramson, 1964). It is the time between the release of a stop, an event represented on one autosegmental tier, and the onset of the periodicity in the waveform associated with the voicing of a following sonorant, an event on a separate tier.

Realization Rules

In this section, an additional grammatical construct is introduced for stating constraints on the realization of an utterance's surface phonemic form, namely, the *realization rule.* Realization rules describe the association between features and their corresponding acoustic properties.

Realization rules are grouped into three classes according to the type of features they specify. The first class of realization rules describes constraints on the temporal orientation of acoustic autosegments and thus state constraints on the acoustic realization of manner-of-articulation features. The second class of realization rules specifies acoustic properties that correspond to place of articulation and their locations. Included in this second class of rules is one that partially describes the realization of the feature [*continuant*] which is associated with a measure of "abruptness" that is found at either the onset or the release of an occlusion. Finally, the third class of realization rules describes the implementation of the feature [*voiced*].

In the remainder of this section, each type of realization rule is described in more detail. The section is concluded with a discussion of how realization rules are used to assign syllable structure to an utterance beginning with its acoustic description.

3.2.1 The Timing of Acoustic Autosegments

As was outlined in Chapter 1, the traditional means by which the acoustic realization of an utterance is described is in terms of transformational rules. These rules are of the general form,

R 3.12

$$\alpha \rightarrow \beta \;/\; \gamma_1 \_\_\_ \gamma_2,$$

and describe the realization of the phoneme α as β, a symbol belonging to the systematic phonetic level of representation. In the standard theory, these rules describe the canonical realization of phonemes, and also attempt to capture systematic variability in phonetic realizations. The tacit assumption is that variability may be described as a process in which phonetic segments are either inserted, deleted, or substituted in a phonetic sequence with alternate phonetic symbols.

Although this rule formalism is capable of producing an accurate description of a language's sound pattern, it is questionable whether it may provide a fully satisfying account of the *nature* of variability, particularly once one considers the multidimensional nature of speech production (cf., Zue (1983), p. 183, for a similar discussion). For example, there are assimilatory processes that the above rule framework describes as symbol substitutions. Assimilation, however, is more naturally described in terms of acoustic properties that "overlap", not segments that are "substituted".

This type of overlap, in this case nasal assimilation, is illustrated in Figure 3.2, which shows a spectrogram of the word *gently* spoken by a male talker. The focus of our discussion will be the word-internal vowel-nasal-stop sequence. The production of this sound sequence requires the coordination of gestures involving the tongue, the velum, and the glottis. It is suggested that, in general, the timing of articulatory gestures in speech is not very precise, although it is necessary to meet certain timing commitments in order to transmit meaningful phonological sequences. An

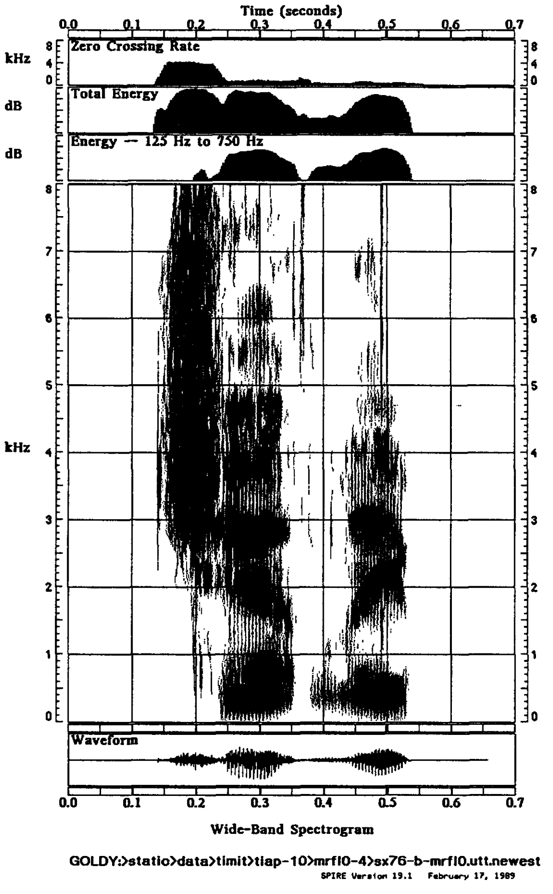


Figure 3.2: Spectrogram of the word *gently*.

example where this hypothesis applies is the sequence of a vowel followed by a nasal in American English, where nasalization in vowels is not distinctive.

The acoustic realization of a vowel-nasal-stop sequence may be qualitatively described as a sonorant region followed by a period of reduced amplitude in the signal corresponding to the oral stop's occlusion. Within the sonorant region, one *may* observe a nasal murmur, but certainly a period of sustained nasalization within the vowel will be observed. In other words, the acoustic properties associated with the features that specify the nasal segment in the abstract representation of this sound sequence are not implemented simultaneously, but instead, may be associated with the preceding vowel at the physical level.

The traditional account of the above process is summarized in terms of the following two transformational rules:

R 3.13

$$\varepsilon \rightarrow \tilde{\varepsilon} \,/\, \_\_ \begin{bmatrix} C \\ +\text{NASAL} \end{bmatrix} \tag{3.1}$$

R 3.14

$$n \rightarrow \emptyset \,/\, \_\_ \begin{bmatrix} C \\ -\text{CONTINUANT} \\ -\text{VOICED} \end{bmatrix} \tag{3.2}$$

Rule (R 3.13) denotes a substitution, specifically, the vowel ε by its nasalized allophone. Rule (R 3.14) denotes a deletion of a nasal consonant.

These two rules fail to reflect the fact that listeners seem to be quite tolerant of this kind of variability in English. Furthermore, this fact may lead one to speculate whether nasalization ought to be considered variability at all (cf., Zue and Sia, 1982). One might hypothesize, for example, in a model of speech recognition, that humans are aware of some set of *obligatory timing constraints* among gestures which serve as a principle of *relative invariance*. In this case, these constraints would express the fact that the gestures involving the velum and the tongue would have to be timed such that a period of sustained nasalization in the sonorant region between the two obstruents /ǰ/ and /t/ is present in order to successfully transmit the presence of an underlying nasal consonant. The rules given in (R 3.13) and (R 3.14) only implicitly capture this intuition.

Over the years, a number of investigators have suggested a multidimensional account of the physical realization of sound sequences (cf., Kent and Minifie, 1977 for a review). Most theories pertain to speech production, and posit articulatory gestures as the primitives of "segmental organization". Gestures are organized into syllables, but the composition is such that exact synchrony is not required between articulatory dimensions (Fujimura 1981). The work of Browman and Goldstein (1986), to date, is the most clearly worked out theory of how this temporal organization may be represented in a linguistic framework. Browman and Goldstein, for example, point to a number of cases involving other forms of variability. By the transformational rule account, their examples appear to be less constrained than by the account given in terms of these elementary articulatory units (Browman and Goldstein, 1986).

One of the questions addressed in the current investigation is, how could some of these articulatory principles be transformed into workable constraints on the acoustic representation. The proposed approach consists of two steps. First, an inventory of dimensions of an utterance's acoustic representation is defined. Secondly, realization rules specify how these primitive acoustic objects are combined and related to parts of the syllable. Specifically, realization rules describe the relation between acoustic autosegments and the initial syllabic sketch.

3.2.1.1 Inventory of Acoustic Autosegments

At present, six types of acoustic autosegments have been incorporated into the proposed framework. They are enumerated in (R 3.15). Like the inventory of features given earlier, the acoustic properties listed are tentative, and are intended to provide a basis for stating the proposed syllable-based theory of acoustic representation.

R 3.15

1. <*sonorant*>
2. <*silence*>
3. <*syllabic*>
4. <*murmur*>
5. <*turbulence*>
6. <*nasalization*>

The list consists of autosegments that were chosen because they may be defined in relatively simple terms. Further, if one were to build a description of an utterance consisting of these acoustic autosegments, it is envisioned that testing for these properties is carried out continuously through the signal in some cases, while inside prespecified regions in others. For example, monitoring for the autosegment <*sonorant*> is continuous. A region of the waveform is considered *sonorant* if there is no acoustic evidence of pressure being built up in the vocal tract and if the source of excitation appears to be at the glottis. In addition, at present it is assumed that sonorants are always voiced (thereby excluding /h/). In this case, one would expect the signal to be be quasi-periodic, with energy concentrated in the lower regions (e.g., below 4kHz) of the signal spectrum. With the exception of sonorants produced with the vocal tract relatively or fully constricted (e.g., sounds produced as a "murmur", i.e., nasals and /l/), all vocal tract resonances are strongly excited.

The remainder of the properties in (R 3.15) require the presence of other properties in the sound stream. For example, the property <*syllabic*> requires that the property <*sonorant*> be present, as do the autosegments <*murmur*> and <*nasalization*>. Although still rather poorly defined in acoustic terms, the autosegment <*syllabic*> is defined to correspond to the temporal region surrounding the maximum of vocal tract aperture. The autosegment <*murmur*> corresponds to regions of the waveform where the vocal tract is relatively constricted, as in the case of a nasal or an /l/. The autosegment <*nasalization*> is implemented simultaneously with the property <*syllabic*>, and corresponds to a region where the vocal tract is open, while the velum is simultaneously lowered.

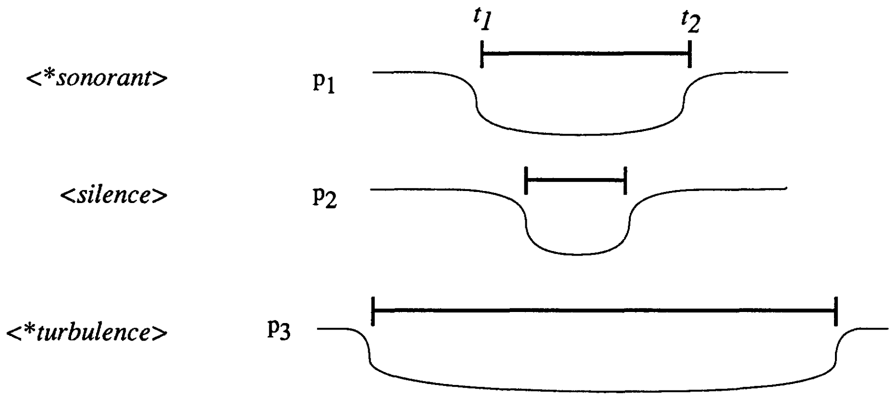


Figure 3.3: Hypothetical parameters extracted from an utterance containing a syllable-initial stop consonant. This figure is intended to illustrate the means by which acoustic autosegments are combined to form elements of the initial syllabic sketch.

The remaining autosegments, <*silence*> and <*turbulence*>, correspond to regions where the vocal tract is constricted to a degree at which a relatively large amount of pressure is built up. In these obstruent regions, a complete closure corresponds to a region where the property <*silence*> is present, whereas a partial closure placed at a location from which airflow is directed towards an obstacle (e.g., the teeth) will produce a region where the property <*turbulence*> is present (Shadle, 1985).

3.2.1.2 Realization Rules

It is assumed that acoustic autosegments are either *present* or *absent* in the sound stream. In both cases, they are defined over regions that are delimited by acoustic landmarks. In addition, they may be assigned a score as a measure of their strength. When taken in varying combinations, autosegments correspond to parts of an utterance's initial syllabic sketch, a grammatical structure that has been already defined at the surface phonemic level of representation.

Figure 3.3 shows three parameters extracted from a hypothetical utterance containing a syllable-initial stop consonant. Parameter p_1 indicates the presence of the autosegment <*sonorant*> in the waveform, that is, the parameter has a "high" value during a sonorant region and a "low" value during an obstruent region. Parameter p_2 has a similar correspondence to the autosegment <*silence*>; it is low during periods of silence and high otherwise. Parameter p_3 corresponds to the autosegment <*turbulence*>; it has a high value during turbulent regions of the waveform and a low value otherwise. Superimposed on these signal streams are symbols denoting the autosegments <* *sonorant*>, <*silence*>, and <* *turbulence*>, where the "*" accompanying the name of an autosegment denotes the absence of the property. In this example, <* *sonorant*> is used to "schematize" the region of parameter p_1 in which this parameter is low. There is a similar schematization of parameters p_2 and p_3 using the autosegments <*silence*> and <* *turbulence*>.

Realization rules specify the spatio-temporal configuration of autosegments corresponding to the assignment of distinctive features to positions within the initial syllabic sketch. For example, for the assignment of the the features

$$\begin{bmatrix} - & \text{SONORANT} \\ - & \text{CONTINUANT} \\ - & \text{STRIDENT} \end{bmatrix}$$

to the OUTER-ONSET position of the syllable, the following constraints on the autosegments <* *sonorant*>, <*silence*>, and <* *turbulence*> must be satisfied:

R 3.16 (Stop in the OUTER-ONSET)

$$\begin{array}{rcl} <\text{silence } t_2> & \leq & <^*\text{sonorant } t_2> \\ <\text{silence } t_2> & > & <^*\text{sonorant } t_1> \\ <^*\text{turbulence } t_2> & \geq & <^*\text{sonorant } t_2> \\ <^*\text{turbulence } t_1> & < & <\text{silence } t_2> \end{array}$$

Rule (R 3.16) consists of a set of inequalities, each making reference to a pair of autosegment endpoints. A set of constraints such as the ones given in (R 3.16) are best understood as stating a set of temporal conditions over a set of acoustic events in the sound stream. For example, in the acoustic realization of an oral stop, the acoustic autosegment $<\textit{silence}>$ denotes the period of occlusion; specifically, $<\textit{silence}\ t_2>$ marks the time of its release. The first two inequalities in (R 3.16) state that the release of a stop must occur in the region delimited by the landmarks $<{}^{*}\textit{sonorant}\ t_1>$ and $<{}^{*}\textit{sonorant}\ t_2>$, i.e., the obstruent region. The latter two statements contained in (R 3.16) specify that during the obstruent region following the release of the stop, turbulence is not to be found.[7] The dependency on syllable structure in the statement of the constraints given in (R 3.16) stems from the fact that the syllable position of a stop encodes information concerning the sonority of the segments that may surround it. For example, for a syllable-initial stop, the endpoint of $<{}^{*}\ \textit{turbulence}>$ may be specified to occur after the onset of the sonorant region. We may make such a stipulation because we have implicit knowledge that a strident segment may not follow a stop in the syllable onset position.

A notation for formally stating realization rules has been developed. The details are provided in Appendix C. An example is presented that illustrates the use of this notation. The realization rule given in (R 3.17) is a more complete statement of the constraints on the timing of autosegments in (R 3.16):

R 3.17

$$\left[\begin{array}{ll}
\textit{constituent:} & \text{Outer-onset} \\[1em]
\textit{output conditions:} & \left[\begin{array}{ll} \text{continuant} & - \\ \text{strident} & - \end{array}\right] \\[1em]
\textit{input realization:} & \left[\begin{array}{rcl}
<\text{silence}\ t_1> & \leq & <^{*}\ \text{sonorant}\ t_1> \\
<\text{silence}\ t_2> & \leq & <^{*}\ \text{sonorant}\ t_2> \\
<\text{silence}\ t_2> & > & <^{*}\ \text{sonorant}\ t_1> \\
<^{*}\ \text{turbulence}\ t_2> & \geq & <^{*}\ \text{sonorant}\ t_2> \\
<^{*}\ \text{turbulence}\ t_1> & \leq & <^{*}\ \text{sonorant}\ t_1>
\end{array}\right]
\end{array}\right]$$

[7] To a first approximation, we may associate the acoustic property $<\textit{turbulence}>$ in one-to-one correspondance with the distinctive feature [*strident*]. As such, we would characterize *aspiration* (associated with a stop's release) as having the property $<{}^{*}\ \textit{turbulence}>$. On the other hand, the frication associated with the release portion of an affricate would be characterized as having the feature $<\textit{turbulence}>$.

This particular realization rule consists of three attributes: the name of a *constituent*, a set of *output conditions*, and an *input realization*. Other realization rules in the grammar will have more attributes depending on the information that is to be represented. The value of the *constituent* attribute specifies the syllable position for which the acoustic and phonological description corresponds. When acoustic syllabification is discussed below, we will describe how this new information is added to an utterance's functional description. The *input conditions* value specifies the values for features that will be assigned to the syllable position specified in the value corresponding to the *constituent* attribute. Finally, the *input realization* value for this realization rule contains the timing constraints stated previously as (R 3.16).

Vowel Nasalization

The rule given in (R 3.17) is but one of a few dozen realization rules that define the initial syllabic sketch at the acoustic-phonetic level. Most rules are similar to the example in (R 3.17), in that they apply to a terminal constituent (e.g., OUTER-ONSET, OUTER-CODA, etc.). However, rules may also make reference to a larger constituent. An example is the rule that describes vowel nasalization in the current framework. The rule is given in (R 3.18).

R 3.18

$$\left[\begin{array}{ll} \textit{constituent:} & \text{Rhyme} \\ & \\ \textit{output conditions:} & \left[\textsc{Coda}\ \left[\textsc{Inner-coda}\ \begin{bmatrix}\text{continuant} & - \\ \text{nasal} & +\end{bmatrix}\right]\right] \\ & \\ \textit{input realization:} & \begin{bmatrix} <\text{nasalization } t_1> & < & <\text{syllabic } t_2> \\ <\text{nasalization } t_1> & \geq & <\text{syllabic } t_1> \\ <\text{nasalization } t_2> & \leq & <\text{syllabic } t_2> \end{bmatrix} \\ & \end{array}\right]$$

In effect, the above rule accepts a nasalized vowel as a well-formed realization of a nasal consonant belonging to the syllable's INNER-CODA. It does so by positing a set of timing constraints as its *input realization* that allow the autosegment <*nasalization*> to overlap with the <*syllabic*> autosegment as shown in Figure 3.4.

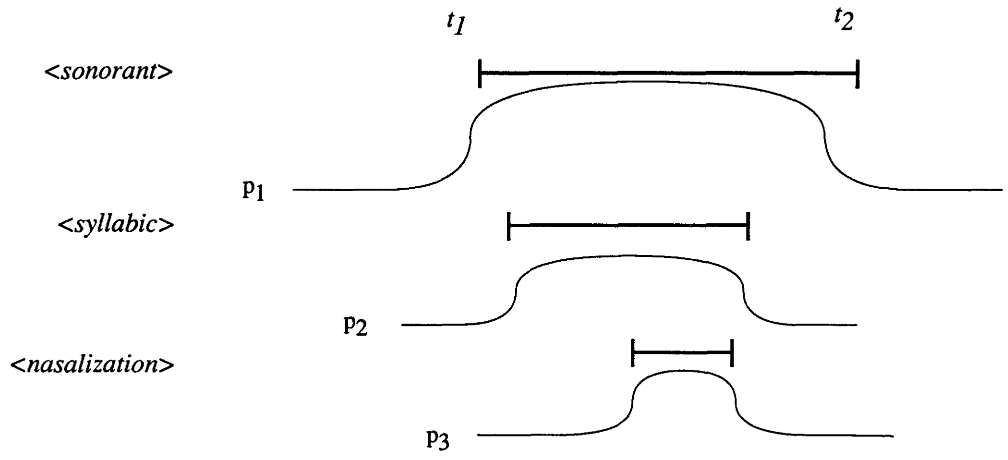


Figure 3.4: Schematic showing the timing of acoustic autosegments for the realization of a vowel-nasal sequence. The realization rule corresponding to this diagram is stated as (R 3.18).

The rule itself applies to the syllable's RHYME, where the RHYME is the smallest constituent of the syllable that dominates both the NUCLEUS and INNER-CODA. The fact that the RHYME does not immediately dominate either of these two constituents requires that a path be specified. Even though the acoustic property corresponding to the feature [+ *nasal*] is to be found in the region corresponding to the syllable NUCLEUS, the feature itself is to be assigned to the INNER-CODA. See Rule (R 3.9) which states the voicing agreement constraint between the OUTER-CODA and the AFFIX-1 positions of the syllable for a similar use of paths.

3.2.2 Transitional Properties

At this point, attention is focused on the second of our three classes of realization rules: namely, rules that specify the implementation of place-of-articulation features. Unlike manner features, which are associated with *regions* of the waveform, place features are associated with *events*. Events are principally located at the transitions between modes of articulation (e.g., at the transition between a vowel and a consonant or at the release of a stop), or at other pre-defined acoustic landmarks. These times identify locations for measuring the shapes of spectra, the transitions of formants, and other properties pertaining to the placement of articulators in the vocal tract. At present, we are not in a position to enumerate a list transitional properties. Instead, the framework by means of which transitional measurements are incorporated into the overall grammatical description is sketched.

Rule (R 3.19) is an example of a realization rule for a transitional measurement. It maps the place-of-articulation features for an alveolar stop onto a spectral shape measurement extracted at the stop's release:

R 3.19

$$\left[\begin{array}{ll}
\textit{constituent:} & \text{Outer-onset} \\
\textit{input conditions:} & \left[\left[\begin{array}{ll}\text{sonorant} & - \\ \text{continuant} & - \\ \text{strident} & -\end{array}\right]\right] \\
\textit{output conditions:} & \left[\left[\begin{array}{ll}\text{high} & - \\ \text{low} & - \\ \text{back} & - \\ \text{labial} & - \\ \text{coronal} & + \\ \text{lateral} & - \\ \text{anterior} & -\end{array}\right]\right] \\
\textit{input realization:} & \left[\ [\text{silence } t_2] \ < \ [\text{*sonorant } t_2]\ \right] \\
\textit{output realization:} & \text{'spectral-shape} \\
\textit{measurement:} & \begin{array}{l}\text{(generic-sample-parameter-at-time} \\ \quad \text{``Wide-band Spectral Slice''} \\ \quad (\text{silence } t_2) \\ \quad \text{rising)}\end{array}
\end{array}\right]$$

As indicated by the value of its *constituent* attribute, the realization rule given in (R 3.19) applies to the syllable's OUTER-ONSET. Unlike the rule stated in (R 3.17), this rule applies to the initial syllabic sketch. Rule (R 3.19) has included an attribute denoted *input conditions*, whose value is a feature bundle which indicates that the rule is to apply if the initial syllabic sketch's OUTER-ONSET position has the specified features assigned to it. The *output conditions* for this rule is a feature bundle containing the complete place-of-articulation specification for the stop. These features will be assigned to OUTER-ONSET of the syllabic sketch if acoustic evidence for the denoted place of articulation is present in the waveform. The *input realization* value for this rule is a logical expression. It ensures that the rule applies to a stop that is released into a sonorant. That is, if the time of the stop's release (denoted as <*silence* t_2>) precedes the time that the obstruent region ends (<**sonorant* t_2>), the rule's *output realization* specifies the name of the acoustic property that is to be extracted, in this case, *spectral-shape.* Finally, the rule has a *measurement* attribute whose value is a *function* that specifies how the acoustic property is to be extracted.

The function *generic-sample-parameter-at-time* is a hypothetical procedure whose purpose is to examine the "Spectral Slice" of the utterance at the time of the stop's release. The term "wide-band" denotes a spectrum computed with a short analysis window. The function is also given the task of determining whether the spectrum is *rising*, which is the prototypical cue for an alveolar consonant (Stevens and Blumstein, 1978).

It should be noted that the rules that map transitional measurements onto the surface phonemic representation of an utterance are less restricted than rules that constrain the temporal orientation of autosegments. Specifically, realization rules of the current kind have a *measurement* attribute which, at present, may contain an arbitrary expression. While on the one hand, this may seem to be an advantage (for example, one may use realization rules of the type presented in (R 3.19) to specify acoustic properties of widely varying types), from a theoretical perspective it is preferable to restrict the properties that rules such as this may characterize. For example, Stevens (1987) proposes a class of *relational properties* for characterizing place-of-articulation features at the acoustic level. These properties involve comparing the values of acoustic parameters, for example, the frequency of one formant in relation to another. The proposed rule formalism is not incompatible with this philosophy.

3.2.3 Constraints on Extrinsic Timing

In the proposed theory, time is an *intrinsic* aspect of the organization of acoustic autosegments within the syllable. That is, autosegments are allowed to overlap. However, no attempt is made to explicitly define the *extent* of overlap in the temporal dimension for the purpose of stating acoustic-phonetic constraints on the syllabic sketch. Rather, the temporal relation among the transitions of autosegments represented on separate tiers is specified in terms of *inequalities* which simply ensure the proper sequencing of acoustic events within the syllabic unit.

Time also has an extrinsic aspect in the proposed framework. For example, the VOT of a stop consonant was defined earlier as the duration between the release of the stop's occlusion and the onset of periodicity in the waveform corresponding to a following sonorant. Voice onset time, in particular, is useful for a variety of purposes. In the current chapter, attention is focused on its use as an acoustic cue to the voicing of a stop. Chapter 4 discusses how VOT may be used to aid in marking the prosodic structure of an utterance.

Voice onset time is one of several critical durations useful for determining voicing for consonants. For example, in a preliminary analysis of a database of fricatives, we found voiced fricatives to be significantly shorter than voiceless fricatives. A similar contrast may be found for affricates as well (see, for example, Klatt (1975) for similar data). The voicing of post-vocalic consonants is often conveyed in the duration of the preceding vowel. Vowels tend to be longer when preceding voiced consonants within the same syllable and shorter before voiceless ones (Denes, 1955). All of these data suggest the importance of using an extrinsic measure of the timing of acoustic autosegments in some way.

A thorough consideration of timing is beyond the scope of the current investigation. Accordingly, the present discussion of this matter will be brief. Nonetheless, it is important to incorporate into the proposed descriptive framework a means by which timing information may be represented explicitly. This is done by defining a class of realization rules similar to ones seen thus far.

The realization rule given in (R 3.20) extracts the VOT for a stop in the OUTER-ONSET of the syllable:

R 3.20

$$\left[\begin{array}{ll}
\textit{constituent:} & \text{OUTER-ONSET} \\
\textit{input conditions:} & \left[\text{Outer-Onset} \left[\begin{array}{ll}\text{SONORANT} & - \\ \text{CONTINUANT} & - \\ \text{STRIDENT} & -\end{array}\right]\right] \\
\textit{output conditions:} & \left[\text{Outer-Onset} \left[\text{VOICED} \; -\right]\right] \\
\textit{output realization:} & \text{'vot} \\
\textit{measurement:} & \begin{array}{l}(- \\ \;\text{(outer-onset *sonorant } t_2) \\ \;\text{(outer-onset silence } t_2))\end{array} \\
\textit{assignments:} & \begin{array}{lll}\text{score} & := & \text{score} \\ & & +l\,[(\text{realization vot}), \text{conditions}]\end{array}
\end{array}\right]$$

One may note that once again the *measurement* attribute is used. In this case, the function specified subtracts the time the sonorant begins <_*sonorant_ t_2> from the time the stop is released <*silence* t_2>. The resulting calculation is the stop's VOT – a symbol that appears as the value of the rule's *output realization* attribute.

The realization rule given in (R 3.20) contains the attribute *assignments*, which has not been seen in any of the realization rules stated in this chapter thus far. In this case, the *assignments* value contains an expression for calculating a score. In general, it may be used for other purposes, most notably to constrain parsing (see

Appendix D). Scores are used to determine the strength of acoustic properties as they relate to underlying phonological categories. In the case of VOT and voicing, longer VOT's correspond to voiceless stop consonants, whereas shorter VOT's correspond to stops that are voiced. In the current framework, the voicing determination, as well as a number of other phonetic decisions, is made on the basis of statistical training (see Chapter 4).

3.2.4 Syllabification

In an earlier discussion of syllabification (in Section 3.1), we described how syllable structure is assigned and represented for an utterance's surface-phonemic level of representation. For the remainder of the present section, the ideas presented earlier are extended to describe acoustic syllabification.

Similar to the case for phonological syllables, the syllabification of an utterance is a process of building an utterance's syllabic sketch by satisfying constraints. Constraints are now derived from a pool of constraints on acoustic realization and on well-formedness of phonological syllables. In the first step of the syllabification process, realization rules that combine acoustic autosegments are applied to specify the syllabic sketch in its initial form. Subsequently, place and voicing features are hypothesized by applying realization rules that correspond to these two categories of features respectively. Throughout this process, a surface phonemic representation of the utterance is being constructed and syllabic constraints on the underlying representation are applied.

In order to accommodate the application of syllabic constraints at both the acoustic and surface-phonemic levels, an utterance's functional description must be expanded to represent both acoustic-phonetic and phonological information. This is done by adding to the *conditions* attribute, already proposed to represent an utterance's phonological information, and a *realization* attribute for representing information pertaining to an utterance's acoustic description. The information represented

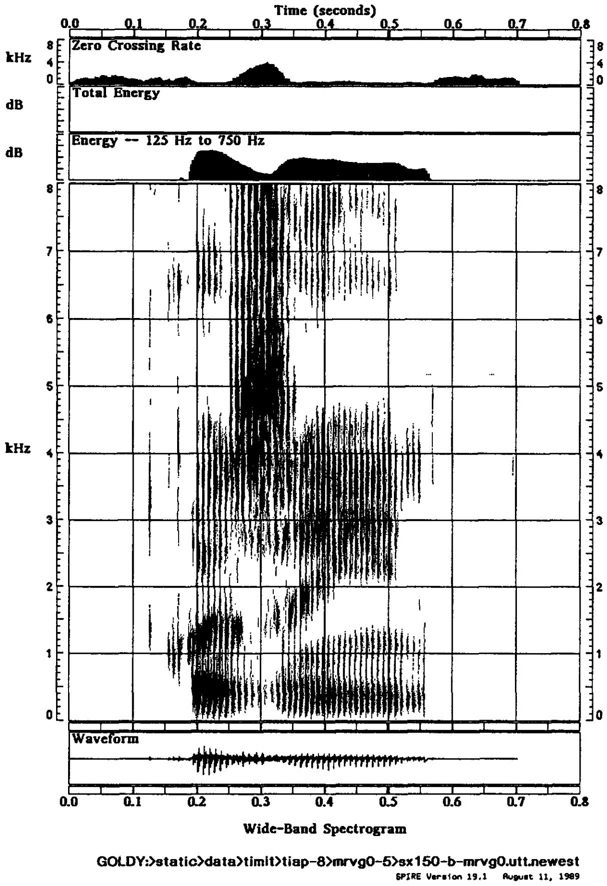


Figure 3.5: Spectrogram of the utterance *cuisine* spoken by a male talker.

as the value for *realization* attribute information is also hierarchically structured.

In the remainder of this discussion, we present two examples. Both illustrate the assignment and representation of an utterance's initial syllabic sketch.

Example I:

In the first example, the initial syllabic sketch for the second syllable of the utterance shown in Figure 3.5 will be presented. The utterance is the word *cuisine* spoken by a male talker. Note that this word contains a word-medial /z/ whose "boundaries" overlap with the surrounding vowels in this particular realization. If one associates the turbulence that is generated by the speaker forming a constriction to produce a /z/ as "belonging" to this segment, then phonetic segmentation is complicated due to the fact that attributes of a vowel-like spectra appear both *after* the beginning of turbulence and *before* its end. In addition, *cuisine* contains a word-final nasal consonant which contributes to nasalization of the second vowel – another problem that complicates phonetic segmentation and labelling.

Acoustic properties that overlap, however, do not present any intrinsic difficulty for a multidimensional acoustic-phonological representation, as long as there are principles to determine how these properties are to be associated to elements of an utterance's underlying surface-phonemic description (i.e., via realization rules). For example, one could formulate a rule for the OUTER-ONSET which did not require the onset of turbulence (i.e., <*turbulence* t_1>) to coincide with the offset of the property (<*sonorant* t_2>); nor would it require the offset of turbulence coincide with the onset of sonorancy.

The syllabic sketch for the second syllable of this word are given in Table 3.1. The functional description given in Table 3.1 consists of the two principal attributes (i.e., *conditions* and *realization*) described above. Both values of these attributes are hierarchically structured. The *conditions* value is a feature specification, whereas the *realization* is a similarly structured collection of acoustic properties.

The functional description shown above is for the initial syllabic sketch of this utterance. Therefore, the features that comprise the *conditions* value are manner-of-articulation features. Also, the acoustic properties that comprise the *realization* value are acoustic autosegments. The notation used to display acoustic autosegments consists of three types of information: 1) the name of the autosegment, 2) the endpoints (shown in parenthesis) and 3) a score or strength (shown as a *log-probability*). The time points correspond to the locations of these autosegments and the score indicates *a posteriori* likelihood the given autosegment exists between the two endpoints.[8]

[8] The details of how one may arrive at an autosegment's score are beyond the scope of the current discussion. The purpose of the score is to assign a measure confidence (or strength) to the presence

Table 3.1: Functional description for the second syllable of the utterance shown in Figure 3.5.

$$\left[\begin{array}{ll}
\textit{constituent:} & \textsc{Syllable} \\
\textit{conditions:} & \left[\begin{array}{ll} \textsc{Core} & \left[\begin{array}{ll} \textsc{Onset} & \left[\begin{array}{ll} \textsc{Outer-onset} & \left[\begin{array}{ll} \textsc{Continuant} & + \\ \textsc{Strident} & + \end{array}\right] \end{array}\right] \\ \textsc{Rhyme} & \left[\begin{array}{ll} \textsc{Nucleus} & [T] \\ \textsc{Coda} & \left[\begin{array}{ll} \textsc{Inner-coda} & \left[\begin{array}{ll} \textsc{Continuant} & - \\ \textsc{Nasal} & + \end{array}\right] \end{array}\right] \end{array}\right] \end{array}\right] \end{array}\right] \\
\textit{realization:} & \left[\begin{array}{ll} \textsc{Core} & \left[\begin{array}{ll} \textsc{Onset} & \left[\begin{array}{ll} \textsc{Outer-onset} & \left[\begin{array}{c} \langle\, \text{*sonorant } (.27\text{s}, .34\text{s})\ {-.05} \,\rangle \\ \langle\, \text{*silence } (.12\text{s}, .57\text{s})\ {-.05} \,\rangle \\ \langle\, \text{turbulence } (.25\text{s}, .36\text{s})\ {-.1} \,\rangle \end{array}\right] \end{array}\right] \\ \textsc{Rhyme} & \left[\begin{array}{ll} \textsc{Nucleus} & \begin{array}{l} \langle\, \text{nasalization } (.41\text{s}, .5\text{s})\ {-.1} \,\rangle \\ \left[\begin{array}{c} \langle\, \text{syllabic } (.34\text{s}, .5\text{s})\ {-.1} \,\rangle \\ \langle\, \text{sonorant } (.34\text{s}, .57\text{s})\ {-.1} \,\rangle \end{array}\right] \end{array} \\ \textsc{Coda} & \left[\begin{array}{ll} \textsc{Inner-coda} & \left[\begin{array}{c} \langle\, \text{sonorant } (.34\text{s}, .57\text{s})\ {-.1} \,\rangle \\ \langle\, \text{*syllabic } (.5\text{s}, .57\text{s})\ {-.1} \,\rangle \\ \langle\, \text{murmur } (.5\text{s}, .57\text{s})\ {-.1} \,\rangle \end{array}\right] \end{array}\right] \end{array}\right] \end{array}\right] \end{array}\right]
\end{array}\right]$$

In the *realization* value of the functional description shown in Table 3.1, one may observe that the <*turbulence*> autosegment has been assigned to the OUTER-ONSET. Its endpoints, as reflected by the spectrogram shown in Figure 3.5, do not coincide with the acoustic autosegment <*sonorant*>. Although, theoretically the property *sonorant* is defined so that no pressure is built up in the vocal tract to cause turbulence. In reality, complete synchrony of boundaries is not found. Nonetheless, a properly structured grammatical description does not preclude the situation observed in this case.

As may be observed in the spectrogram shown in Figure 3.5, both the autosegments <*nasalization*> and <*murmur*> are to be associated with the word-final nasal consonant. As such, both of these acoustic properties are represented on the final syllable's functional description. In particular, in the *realization* value, there is a position for the OUTER-CODA which contains three acoustic autosegments, one of which is <*murmur*>. The autosegment <*nasalization*> is assigned to the <*Rhyme*>. This is denoted in the functional description by having this autosegment placed as a value of the RHYME attribute as opposed to the INNER-CODA.

Example II:

The second example involves the utterance *gently* (used in an earlier discussion motivating a multidimensional view of an utterance's acoustic representation). The functional description for the first syllable of this utterance is given in Table 3.2.

This utterance contains examples of nasal assimilation and nasal murmur deletion (by the transformational rule account). Although there is not an explicit nasal deletion rule in the proposed framework, the fact that the murmur is deleted may be discerned in the utterance's functional description by the absence of the INNER-CODA constituent from its *realization* value and the presence of this terminal constituent on its *conditions* value. As in the previous example, the autosegment <*nasalization*> is assigned to the syllable's RHYME.

of an autosegment.

Table 3.2: Functional description for the first syllable of the word *gently* shown in Figure 3.2.

constituent:	SYLLABLE				
conditions:	CORE	ONSET	OUTER-ONSET		CONTINUANT − STRIDENT +
		RHYME	NUCLEUS	[*T*]	
			CODA	INNER-CODA	CONTINUANT − NASAL +
				OUTER-CODA	CONTINUANT − STRIDENT −
realization:	CORE	ONSET	OUTER-ONSET		< *sonorant (0.0s, 0.24s) −.05 > < silence (0.0s, 0.14s) −.05 > < turbulence (0.14s, 0.24s) −.1 >
		RHYME	NUCLEUS		< nasalization (0.27s, 0.35s) −.1 > < syllabic (0.24s, .35s) −.1 > < sonorant (0.24s, .35s) −.1 >
			CODA	OUTER-CODA	< *sonorant (.35s, 0.38s) −.1 > < *turbulence (0.24s, .66s) −.05 > < silence (0.35s, 0.38s) −.1 >

3.2.5 Concatenating Syllables

In the two examples just presented, the syllabification of monosyllables has been considered. The purpose of these examples was to provide a simple illustration of the means by which syllable structure is assigned to an utterance starting with a multidimensional acoustic representation. Additional problems arise, however, in the syllabification of polysyllables that the theory presented thus far has not explicitly considered. For some polysyllables, it is expected that syllabification does not represent any particular problem. In the case of *gently*, for example, the second syllable would be syllabified by satisfying constraints that are similar to those that enable syllable structure to be assigned to the first syllable.

There are two important cases involving phonetic variability that complicate matters. The theory needs to be extended to address these problems. The first involves allophonic variation in the realization of segments at the boundaries of syllables that are correlated with the placement of the syllable boundary. Examples of this phenomenon include the acoustic realization of the minimal pairs *grey train* vs. *great rain* or *known ocean* vs. *no notion*, etc. In these cases the underlying phoneme sequences are identical, but the surface acoustic structures of these words differs. The second class of problems involve coarticulation that transgresses syllable boundaries. For example, *palatalization* (e.g., *did you* → /dɪǰu/) is a process that often crosses syllable boundaries as does *rounding*, *retroflexion*, among others. Allophonic variability of the first kind may be treated within a theory of the syllable. An acoustic study of these phenomena will be presented in Chapter 4, where a mechanism for incorporating constraints on these kinds of phonetic alternations will be proposed. The second class of processes will require that the theory be extended in a more fundamental way. Perhaps larger constituents might be considered. These larger constituents would serve as domains over which a wider range of processes could occur and be described.

3.3 Summary

In this chapter we have proposed a representational framework for describing constraints on an utterance at both the surface phonemic and acoustic levels. In the current proposal, an utterance's surface phonemic representation is comprised of features; features are binary-valued and align into columns to form phonological segments. Further, features are assigned to the terminal slots of a syllable template. At the surface-phonemic level of representation, the syllable template facilitates the statement of constraints on well-formed sequences of phonemes. At the acoustic level, the template facilitates the statement of constraints on an utterance's pattern of acoustic properties.

The crucial aspect of our framework is the multidimensional nature ascribed to the representation of speech at both the abstract and physical levels. At the physical level, a multidimensional description is deemed necessary to account for the nature of speech production. That is, the speech signal is the result of a coordination of activities of a vocal apparatus in which individual articulators move with differing degrees of sluggishness. As a result, the acoustic properties associated with features are not implemented simultaneously in the sound stream. Our descriptive framework provides the necessary degrees of freedom in order describe constraints on the temporal organization of patterns of acoustic properties.

Our theory should be considered tentative. We have defined a representative inventory of acoustic properties and features. Our purpose in doing so was to provide a concrete basis for illustrating the issues for which the proposed representational framework was developed. Further research will be required to in order to develop a more definitive set of properties and features, and to provide justification for the theory of syllabic organization over these primatives.

Chapter 4

Syllabic Constraints on Extrinsic Variation — A Case Study of the Stop Consonants of American English

This chapter describes the treatment of extrinsic allophones within the syllable theory outlined in Chapter 3. The proposal is to formulate rules for describing extrinsic allophones within a statistical framework. Specifically, a class of regression models based on *binary decision trees* is suggested for predicting the acoustic realization of a phoneme as a function of its distinctive feature specification, local phonetic environment, stress environment, and syllable position. Experimental justification in support of this approach is offered in the form of a case study involving stop consonants in American English. The chapter consists of three main sections. Section 4.1 is an introduction in which the definition of extrinsic allophones is reiterated. In Section 4.2 the details of the proposed rule formalism and regression methods are provided. Finally, Section 4.3 is the case study.

4.1 Introduction

In this thesis, the term "extrinsic allophones" has been used to refer to a class of phonetic alternations that are attributed to the structural context of an underlying phoneme (i.e., its position within some larger phonological unit). To a first approximation, the term is used to distinguish this linguistic behavior from variability in the realizations of a phoneme that may be attributed to the inherent mechanical constraints imposed on the articulatory mechanism (i.e., its intrinsic allophones). We suggest that the nature of the constraints imposed on these two forms of variability differ, and as a consequence, separate types of rules are required to describe these phenomena. In Chapter 3, for instance, it was argued that intrinsic variability requires that speech be treated, at the physical level, as multi-dimensional. The dimensions pertain to individual vocal tract articulators (e.g., the velum, the lips, the glottis, the tongue and the jaw). Although the underlying physiological/physical principles that enable one to predict the exact nature of intrinsic variability have not been specifically addressed, a framework of realization rules was proposed in Chapter 3, whereby constraints on the temporal organization of physical dimensions may be stated.

Extrinsic variation obeys a different set of principles. Several investigators have suggested, for example, that the placements of extrinsic allophones in the sound stream serve as the primary markers of an utterance's phonological structure (see review below). However, there are additional factors, such as speaking rate and dialect, that may also play a role in a speaker's choice of extrinsic allophones. Thus, the proposed use of statistical methods are an attempt to compensate for this lack of understanding.

The position that is adopted in this thesis is that phonetic variability of both forms is highly constrained. Our goal is to propose methods for describing these behaviors in both qualitative and quantitative terms.

4.1.1 Extrinsic Allophones of American English

The theoretical basis proposed for distinguishing extrinsic allophones of a phoneme from its contextual variants, that are dependant on intrinsic factors, is in practice only a first-order approximation. One may cite several cases where there is an interaction among extrinsic and intrinsic factors. For example, a phoneme's acoustic realization may be dependant primarily on its position within the syllable or word, but, the phonemes that surround it (which contribute to its intrinsic variability) will also determine to a certain extent its acoustic realization. What this implies in terms of a rule formalism is that constraints on extrinsic allophones will have a component that describes the phoneme's local phonetic environment in addition to a component that describes its structural context.

Conversely, intrinsic variation may be dependant on structural factors. For example, in the realization of a Vowel-Nasal-Stop sequence, the degree to which the vowel is nasalized varies, as does the duration of the nasal murmur. Variation in both of these acoustic parameters depends on the voicing and syllable position of the stop (Fujimura and Lovins, 1978).

In the discussion that follows, examples will be provided of extrinsic allophones, and the interaction between intrinsic and extrinsic factors will be pointed out.

The Extrinsic Allophones of /t/

The extrinsic allophones of stops represent the most well-studied of such phenomena. It appears that the alveolor voiceless stop /t/ undergoes the most drastic and widely varying of modifications. For example, Figure 4.1 shows a spectrogram of the utterance *Tom Burton tried to steal a butter plate.* In it are seven instances of this phoneme, each having a different phonetic realization. The table given in (R 4.1) summarizes the phonetic alternatives found in this utterance as well as relevant comments regarding their contextual environments.

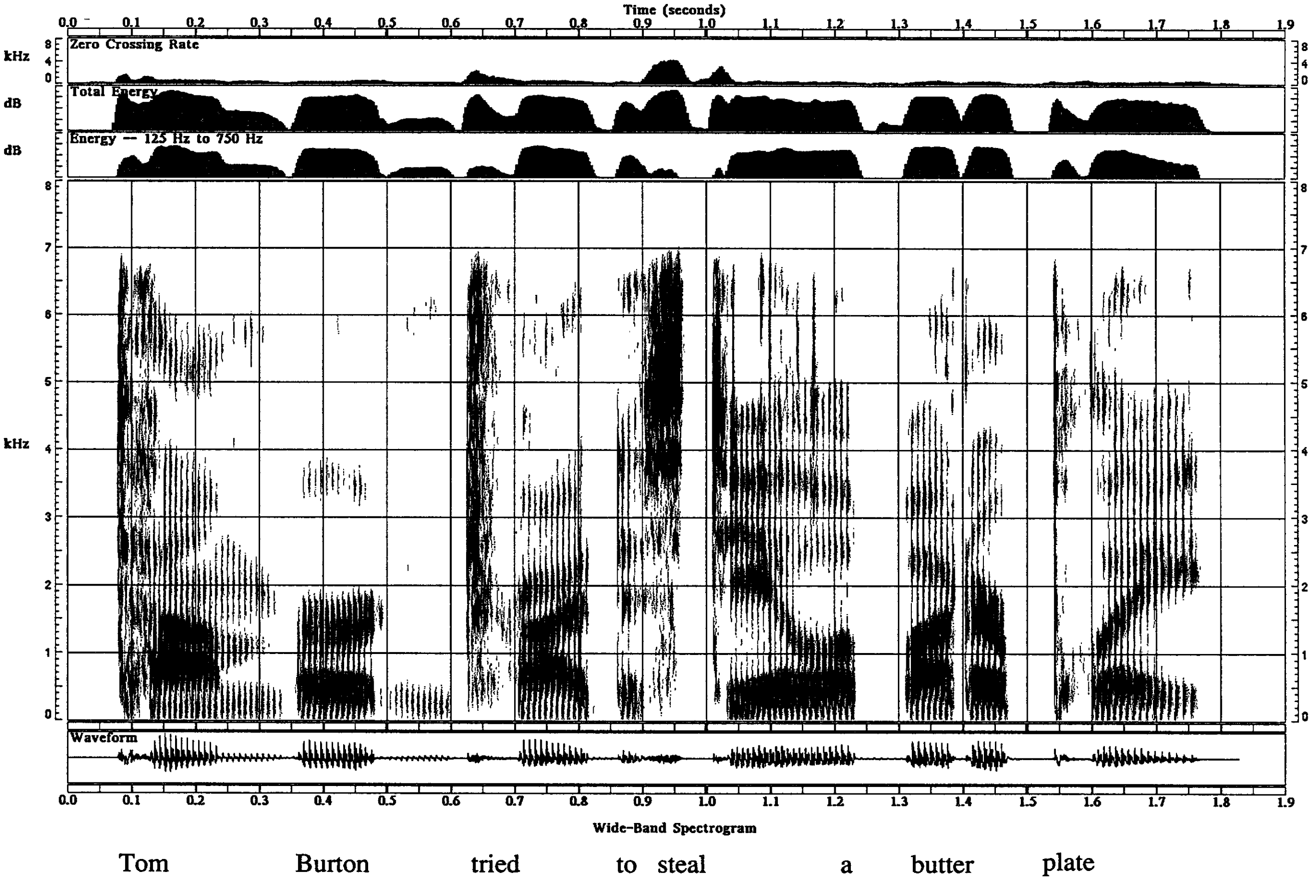


Figure 4.1: Spectrogram of the utterance *Tom Burton tried to steal a butter plate.* spoken by a male talker. This spectrogram illustrates the various acoustic realizations of the phoneme /t/.

		Instance	*Realization*	*Context*
	1)	*Tom*	*aspirated*	*syllable-initial, pre-stressed*
	2)	*Burton*	*glottal stop*	*syllable-final, post-stressed*
R 4.1	*3)*	*tried*	*retroflexed*	*syllable-initial, in cluster with /r/*
	4)	*to*	*released*	*syllable-initial, unstressed*
	5)	*steal*	*unaspirated*	*syllable-initial, in cluster with /s/*
	6)	*butter*	*flap*	*syllable-final, inter-vocalic, post-stressed*
	7)	*plate*	*unreleased*	*syllable-final*

The hypothesis is that the syllable (or word) position of the stop is an important determining factor in its phonetic realization. Also important is the stress environment of the stop, as well as whether the stop is intervocalic, in a cluster with either the phoneme /s/ as in *steal* or with /r/ as in *tried.*

Light vs. Dark /l/

Most phonemes do not possess as wide an array of extrinsic allophones as does the phoneme /t/. For example, two major extrinsic allophones of the phoneme /l/ may be distinguished. Figure 4.2 shows spectrograms of the words *lick* and *kill.* The /l/ in *lick* is the so-called "light" or "clear" /l/ and is the syllable-initial allophone. The "dark" or syllable-final /l/ is the contextual variant found in the word *kill.* The most apparent differences between these two contextual variants are their formant transitions. However, it has been suggested that listeners perceive a more abrupt onset in amplitude accompanying the release of a syllable-initial /l/. On the other hand, the syllable-final /l/ has a more "syllabic" quality (Nakatani and Dukes, 1977).

Extrinsic Allophones of /r/

Similarly, at least two extrinsic allophones of the phoneme /r/ may be found. Figure 4.3 shows spectrograms of the words *rock* and *car* that illustrate these allophones. Once again, the difference between these phonetic variants is manifested in the formant transitions into and out of the vowel. As in the case of /l/, the differences between these two contextual variants is also attributed to syllable-initial /r/ being "more consonantal" and the syllable-final /r/ "more syllabic".

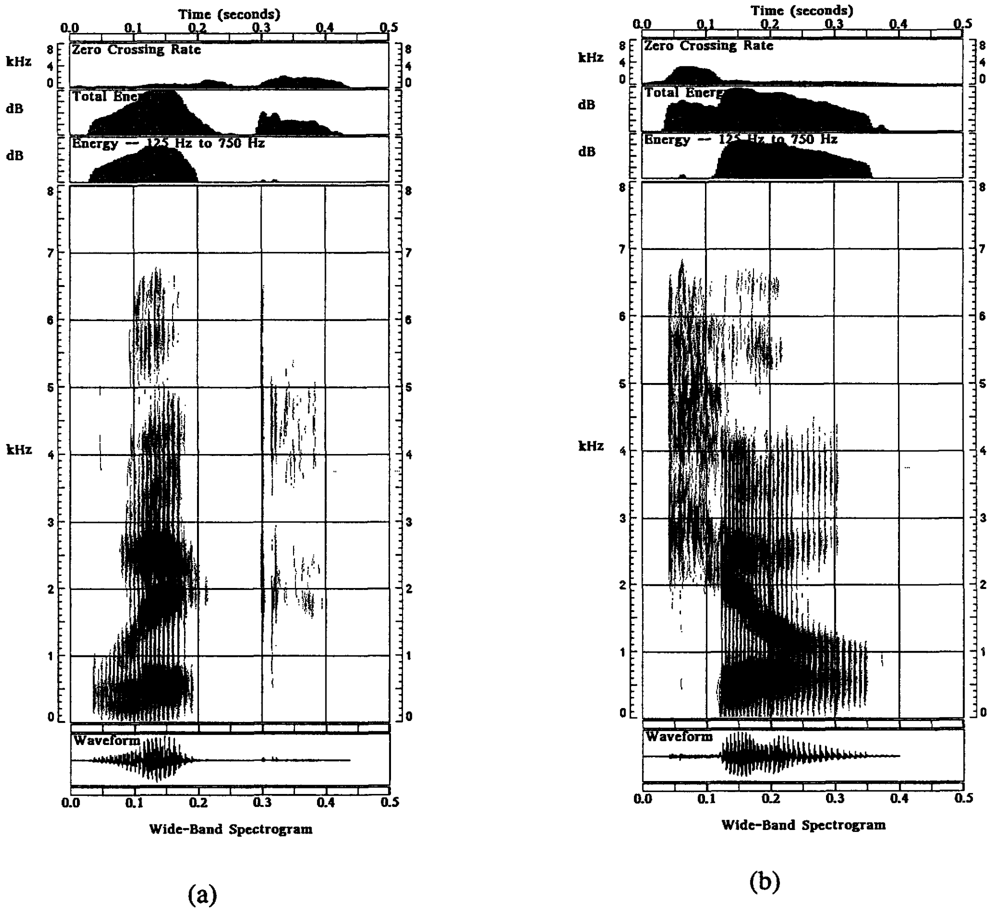
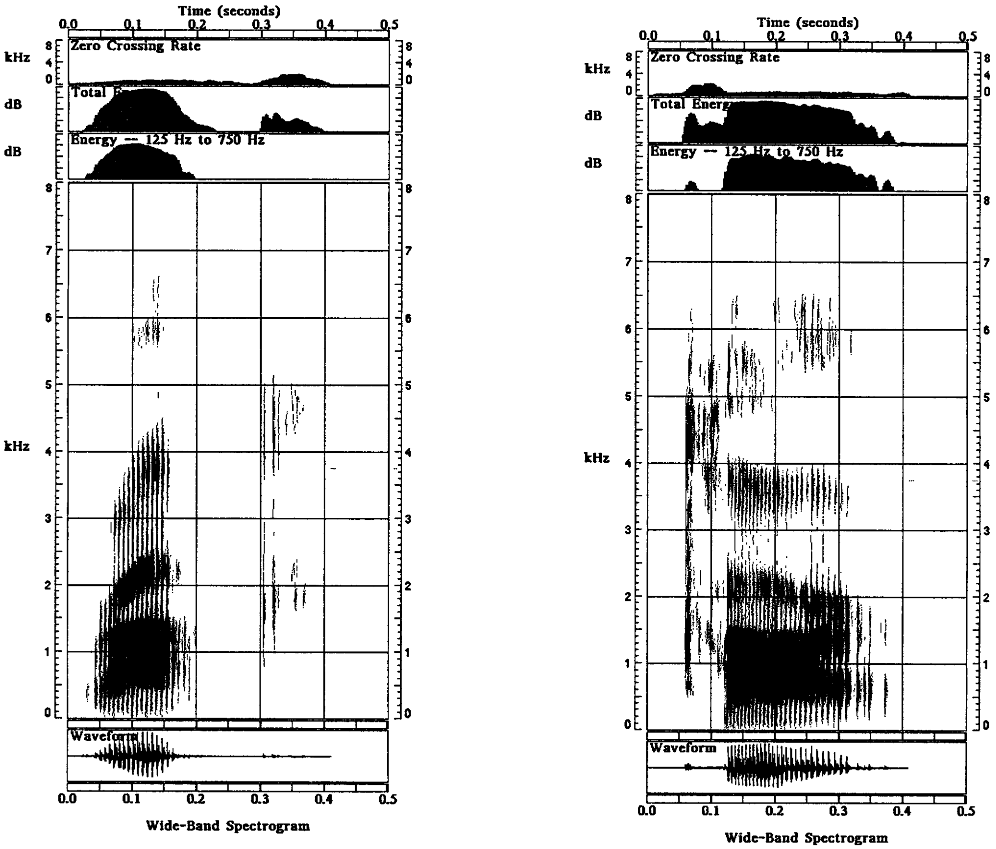


Figure 4.2: Spectrograms of the utterances (a) *lick* and (b) *kill* spoken by a male talker. These examples illustrate the acoustic differences between the "light" /l/ (*lick*) and the dark /l/ (*kill*).



(a)

(b)

Figure 4.3: Spectrograms of the utterances *rock* and *car* spoken by a male talker. These examples illustrate the acoustic differences between the syllable-initial /r/ (*rock*) and the syllable-final /r/ (*car*).

4.1.2 Extrinsic Allophonic Variation: Acoustic Cues to Word and Syllable Juncture

Because of the systematic relationship between a consonant's phonetic realization and its structural context, many past studies of extrinsic allophones have focused on uncovering the strategies used by listeners in determining the phonetic cues that mark syllable and word boundaries in fluent speech (Lehiste, 1960; Christie, 1974; Nakatani and Dukes, 1977). Based on these studies, models of phonetic parsing and word recognition have been proposed that rely on a syllabic structuring of an utterance prior to lexical retrieval (cf., Church, 1983; Frazier, 1987; and the model proposed in Appendix D).

Lehiste (1960) has been a pioneer in these efforts. Using natural speech as stimuli, Lehiste asked listeners to indicate where in phonetic sequences they heard syllable and word juncture. She then correlated responses with visual cues found on spectrograms. Results from these experiments prompted Lehiste to suggest that allophonic variation of this type was more than "excess baggage that the listener must work his way through".

With similar objectives, Christie (1974) and Nakatani and Dukes (1977) have performed studies with synthetic speech. Christie presented subjects with variations of the phoneme sequence /asta/. He tested for the effects of three acoustic cues on listener perception of juncture: 1) the presence or absence of formant transitions from the initial /a/ into the /s/, 2) the presence or absence of aspiration accompanying the /t/, and 3) the length of silence between the end of the /s/ and the /t/ burst. Christie's analysis of the results of his experiment showed no strong statistical effect due to formant transitions. The presence of aspiration seemed to provide the strongest juncture cue. The length of the silence interval had a lesser first order effect on the perception of juncture, but, it did have a strong interaction with presence of aspiration. In terms of *location of juncture*, his results showed that the presence of aspiration signalled a perception of the juncture between the /s/ and the /t/. The absence of aspiration signalled the perception of the boundary between the /a/ and the /s/. Finally, as the the period of silence became longer, there was a greater likelihood for syllable boundary perception between /s/ and the /t/.

The speech material used in the Nakatani and Dukes study was considerably broader. Their method was to splice together *dyads*[1] taken from regions surrounding the word boundaries of minimal pairs such as *grey train* verses *great rain*, or *no notion* vs *known ocean.* Like Lehiste, the authors of this study correlated scores of juncture perception with the visually observable acoustic properties of these regions (as seen on spectrograms). They discovered, for example, glottalization of word-initial vowels to be a strong cue to juncture as well as the modifications in the formant structure of /l/ and /r/ discussed above. Like Christie, Nakatani and Dukes also found the presence of aspiration in stops to indicate the placement of the word (or syllable) boundary directly before it.

In sum, the distributions of extrinsic allophones suggest a set of principles at work that are different in nature from those that explain coarticulatory phenomenon (otherwise known as intrinsic variability). Specifically, extrinsic allophones when placed in the sound stream signal information regarding its phonological structure. In the next section, we turn to the question of how extrinsic allophonic variation is to be given a grammatical account.

4.2 Extrinsic Allophones: Phonological Representation and Rules

4.2.1 Traditional Accounts

The traditional phonological account of extrinsic allophonic processes has been to formulate a set of generative rules for describing this behavior. In (R 4.2) for example, the rule:

[1] Dyads are acoustic segments defined as half of one phonetic segment followed by half of its neighbor where the transition between phones is internal to the unit itself.

R 4.2

$$\left\{ \begin{array}{c} p \\ t \\ k \end{array} \right\} \rightarrow \left\{ \begin{array}{c} p^h \\ t^h \\ k^h \end{array} \right\} / \ \$ \ \_ \ V \cdots$$

states that syllable-initial, pre-stressed voiceless stop consonants are realized as *aspirated* (as denoted by the diacritic ^{h}). The raised dots ("$\cdots$") in this rule indicate that the remainder of the stop's tauto-syllabic context may be ignored.

A number of general criticisms concerning this formalism for stating rules have been raised in previous chapters. We stated, for example, that such rules are associated with linear phonological representations; in order to overcome this limitation, these rules are generally unrestricted in nature. That is, they provide for an arbitrary manipulation of symbols without any inherent restrictions on the kind of information that may be represented. Therefore, these rules provide little, if any, insight into the nature of allophonic variation and its underlying causes. Also, because of their unrestricted nature, these rules have a generative capacity that exceeds the capabilities of a wide range of recognition devices (cf., Church, 1983, for a more complete discussion of this latter issue). A criticism directed more to the specific example given in (R 4.2) is the categorical nature of these rules. The use of allophonic symbols in the rule's output suggests that phonetic variation may be described in qualitative terms. In our view, certain forms of variability are best described by allowing a rule's output to assume a continuum of values.

Context-free Rules:

Church (1983) proposes a more restricted class of rules that are intended to overcome many of the above limitations. Allophonic rules are to be stated in terms of an immediate constituent grammar. For example, the structural context of a segment is represented by its assignment to one of the terminal positions of the syllable's hierarchical structure. According to this arrangement, for instance, the environment for aspirated stops is the syllable onset position. The rule stated in (R 4.3) is a concise statement of this fact.

R 4.3

$$\textsc{Onset} \longrightarrow p^h | t^h | k^h$$

Furthermore, languages described by immediate constituent grammars may be parsed and/or recognized with relative efficiency by a wide class of well-known algorithms (Aho and Ullman, 1972).

Church's framework represents an important improvement over the more standard approach and depicted in (R 4.2) , but it shares one of its principle disadvantages, in that these rules impose constraints on an utterance's systematic phonetic level of representation. In other words, the output of rules are still allophones. Although the systematic phonetic representation is intended to accurately characterize all grammatically-determined contextual variants of phonemes, as we have previously mentioned, phonetic theory has yet to provide an objective basis for deciding the right inventory of allophones.

4.2.2 The Use of Regression Methods for Discovering Phonetic Rules

In the proposed treatment of extrinsic allophones, rules that characterize this behavior are part of our grammatical framework's realization component. At the core of each realization rule is a statistical predictor having the general form given in (R 4.4):

R 4.4

$$\vec{\beta} = r\{\vec{\alpha}, \vec{\gamma_1}, \vec{\gamma_2}, \sigma, S\}.$$

In this equation, the variable $\vec{\beta}$ denotes a given phoneme's acoustic realization (a value that may be qualitative or quantitative) and $\vec{\alpha}$ is the phoneme's feature specification.

The remaining variables characterize the phoneme's local phonetic and prosodic contexts. In particular, $\vec{\gamma_1}$ and $\vec{\gamma_2}$ denote the left and right context of $\vec{\alpha}$, and σ and S denote the phoneme's syllable position and stress environment respectively.

The regression function r given in (R 4.4) may be viewed as an extension of the traditional framework of context-sensitive rules. This similarity becomes evident when the rule description given in (R 4.4) is rewritten in more conventional terms, as is expressed in (R 4.5).

R 4.5

$$\vec{\alpha} \xrightarrow{r} \vec{\beta} \;/\; \vec{\gamma_1} \;\underset{\left\{\begin{matrix}\sigma\\ S\end{matrix}\right\}}{\underline{\qquad}}\; \vec{\gamma_2}$$

In (R 4.5), the symbols σ and S that are beneath the underscore ("___") indicate that the syllable position and stress environment of α are to be considered in predicting its phonetic realization. The symbol "$\xrightarrow{r}$" denotes a special kind of production; to be read "realized with probability r".

We gain two important advantages by formulating allophonic rules within the statistical paradigm of regression analysis. First, acoustic-phonetic modifications attributable to the context of a phoneme may be described with more accuracy. Aside from allowing the output of rules to assume a continuum of values, the improvement in accuracy comes from allowing rules to be stated using probabilities. Thus, rather than describing certain forms of linguistic behaviors as "obligatory" or "optional", we may assign a quantitative measure of certainty. The second important advantage is that rules formulated as regression functions may be inferred automatically from a database of labelled speech and acoustic measurements. Therefore, statistical methods may be used as an aid in the process of testing phonological theories and perhaps "filling in" gaps in existing knowledge.

For the remainder of the current section, the regression procedures that we have implemented are outlined.

4.2.2.1 The Sèlection of a Regression Method

The formula given in (R 4.4) subsumes a variety of regression techniques, most notably a range of linear (or classical) methods (cf., McCullagh and Nelder, 1983; Draper and Smith, 1966) as well as non-linear approaches such as the one that has been adopted based on regression trees (Breiman *et al.*, 1984). In general, a regression model will consist of two components: 1) a *structural model* (i.e., a formula that indicates how model predictions are computed) and 2) a criterion for assessing how well a model fits a given body of data: a so-called "goodness-of-fit" measure. Additional relevant aspects of a regression technique include the algorithm used to obtain the model, as well as some means of assessing the model's complexity.

Linear Methods

Early in our investigation, experiments were performed using linear or "ANOVA-type" models. We used *Log-linear models* (Bishop *et al.*, 1975) in the analysis of data for which the response variable is categorical, and so-called *factorial designs* (Winer, 1962) with continuous responses. In the linear approach, model predictions are computed as a *linear combination* of predictor values. A general expression is given in (4.1),

$$g(\mu_j) = \sum_{i=1}^{p} \beta_i x_{ij}. \tag{4.1}$$

In (4.1), the terms x_{ij}, for $1 \leq i \leq p$ describe an individual token's context. The terms β_j are the model's parameters, which determine the exact nature by which contextual factors are to be weighted and combined. They are estimated from the observed data. Finally, the function $g(\ )$ is called a *link* function, and is a transformation (not necessarily linear) that is applied to the output of the linear combination, μ_j.[2]

For a number of reasons, we abandoned the use of linear methods. We found the principal disadvantage of the linear approach to be its inefficient use of limited

[2] Notation is adopted from McCullagh and Nelder (1983).

amounts of data, although for low-dimensional problems (e.g., when the number of predictor variables is less than four), the parameters of a linear regression model are relatively easy to estimate. However, the following conditions must exist: a) there is a reasonable amount of data, and b) the number of levels for each of the model's dimensions is small to moderate. In addition, under these conditions, the model parameters may be readily interpreted. Efforts to estimate and interpret parameters quickly become intractable as the model's dimensionality becomes larger. Part of the difficulty is the well-known "curse of dimensionality" (Bellman, 1957): as the number of parameters increase, the amount of data required to obtain robust estimates increases at a rate that is exponential in the number of dimensions. Second, the models obtained from these estimates are linear combinations of "first-order effects" and "higher-order interactions" (Winer, 1962). Although we found interpreting first-order terms to be reasonably straightforward, the interaction terms incorporating more than three factors were nearly impossible to understand.

Tree Regression

As an alternative, we explored a tree regression method, which was found to be quite satisfactory. Although regression trees are not completely immune to problems of high dimensionality and sparse data, the interpretation of the hierarchical structure that comprises a regression tree is more straightforward. In addition, there are means of adjusting the complexity of the tree in such a way that any lack of data does not severely retard the robustness of the resulting prediction (although trees that have been simplified fail to lend as much insight into the data).

We have limited our exploration of regression tree techniques to methods involving *binary* trees (i.e., trees with a "branching factor" of 2). In binary tree regression, model predictions are arrived at through a series of yes/no (i.e., binary) decisions. The decisions are represented as the nodes of a tree. The tree's leaves (or terminal nodes) represent the model's predictions. The prediction itself may be qualitative or a real number. For the purposes of illustration, a hypothetical tree that gives qualitative predictions is given in Figure 4.4.

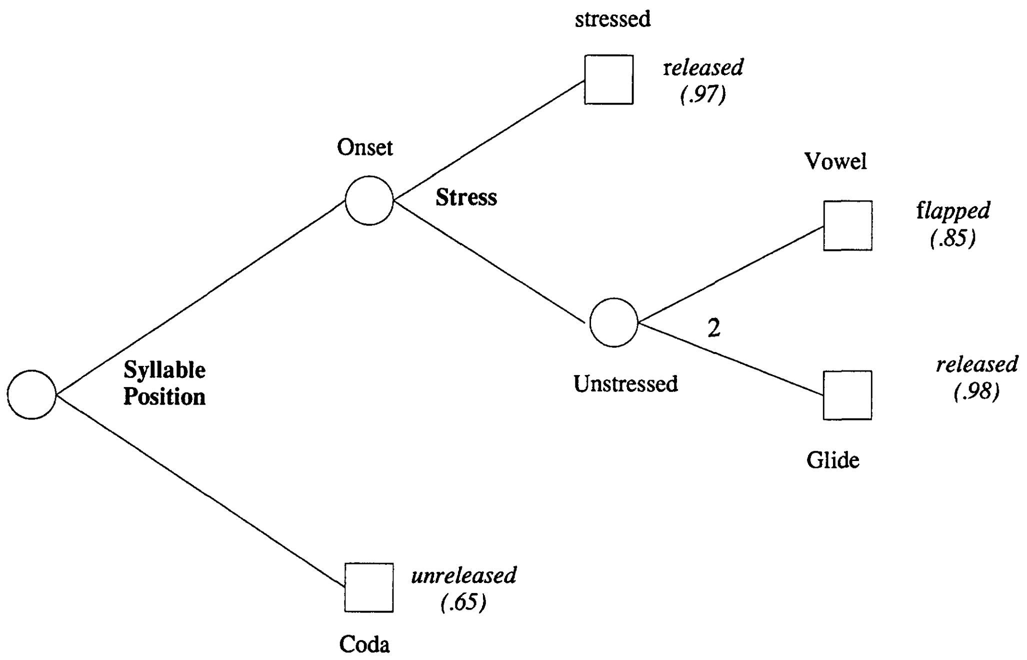


Figure 4.4: A hypothetical tree illustrating the binary tree regression tree approach to formulating allophonic rules.

The hypothetical tree given in Figure 4.4 is an example of what might be used to predict acoustic realizations for the phoneme /t/. The predictor variables are: a stop's *syllable position*, σ; *stress (of the following vowel)*, S; and the stop's *following context*, γ_2. In general, decisions at nodes (also known as *node splits*) are based on the values of variables that represent a particular aspect of a phoneme's feature specification or context. A path that is traversed in reaching a terminal node (i.e., a prediction) specifies a context that the tree deems as being relevant. For example, the tree shown in Figure 4.4 makes the four predictions for the acoustic realizations of the phoneme /t/. The predictions are are listed in tabular form in (R 4.6) (the X entry in the table indicates that a value for the given particular contextual variable is not specified):

		Context		*Predicted Realization (probability)*
	σ :	S :	γ_2 :	
R 4.6	*Onset*	*Stressed*	*X*	*released (.97)*
	”	*Unstressed*	*Vowel*	*flapped (.85)*
	”	”	*Glide*	*released (.98)*
	Coda	*X*	*X*	*unreleased (.65).*

The four predictions enumerated in (R 4.6) correspond to the tree's four leaves. Two leaves represent conditions for a /t/ to be *released* while the remaining two leaves correspond to conditions for a /t/ to be *flapped* and *unreleased* respectively.

4.2.3 Growing Regression Trees

The hypothetical tree used in the above example was constructed by hand for the purposes of illustration. In order to perform this investigation, we have implemented computer software for constructing trees both by hand or using using an automatic procedure based on the CART (Classification and Regression Trees) algorithm (Breiman et al, 1984).

The process of obtaining a regression tree is called "tree growing". The CART algorithm (and others like it) for growing trees automatically is a *step-wise* optimization procedure, where growing a tree consists of computing a set of *node splits.* At each node (or step of the computation), the criterion of maximum mutual information is used to direct individual computations. This measure will be defined once the essentials of tree growing have been provided.[3]

Trees are grown from a *learning* or *training* sample comprised of a set of tokens. Let $X = \{\vec{x_1}, \vec{x_2}, \ldots \vec{x_n})$ denote such a learning sample, where each token,

$$\vec{x_j}^T = [f_{1j}, f_{2j}, \ldots f_{pj}, a_j],$$

is a $(p+1)$-dimensional attribute vector. In the current use of the tree regression method, the first p of the attributes comprising $\vec{x_j}$ (i.e., f_{ij}, for $1 \leq i \leq p$) are the set of predictor variables. The remaining attribute, a_j, is the phoneme's acoustic realization. As mentioned before, a_j is either a real number ($a_j \in \mathcal{R}$) or a qualitative value.

At the start, a tree will contain only its root node. The tree growing procedure is initialized by placing the learning sample in the root of the tree. The next step is to choose one of the attributes for partitioning the root's sample. Once the choice is made, the node is split, and the two subsamples that are obtained are placed in the original node's descendants. At this point, the procedure is repeated. The iteration continues until either one of three conditions are met: 1) the sample resulting from a split is "pure" (i.e., containing acoustic realizations of only one type), 2) the sample is of such a relatively small size that further splitting is no longer warranted, or 3) splitting the node does not significantly improve the tree's performance.[4]

[3] We should point out that the criterion of maximum mutual information has been used in a number of other speech-related contexts for statistical parameter estimation. Most notable is its use for estimating the parameters of a *hidden Markov model* for automatic speech recognition (cf., Bahl *et al.*, 1986).

[4] In the latter case, a simpler tree is preferred. In the CART procedure, Breiman *et al.*, suggest

The details of how a tree's prediction values are obtained differ, depending on whether the response variable is continuous or categorical. Since trees are binary, in both cases, a series of yes/no questions that depend on the predictor values for a given test token are answered. During this question answering process, a path through the tree is traversed. Once a terminal node is reached, the predicted response depends upon the distribution of learning sample tokens that the node contains. For the case where the response variable is categorical, the predicted value is the mode of the distribution, i.e., the response value that corresponds to the majority of the learning sample tokens contained in that node. For a continuous response, the prediction value is the sample mean of the node's distribution.

4.2.3.1 Mutual Information as a Measure of Tree Performance

In practice, a regression tree rarely meets the goal of complete accuracy in making predictions (even for tokens in the learning sample). Therefore, an appropriately defined goodness-of-fit measure is necessary to indicate how well the job has been done. Breiman *et al.*, suggest the use of a tree's *error* or *misclassification rate* to evaluate its performance. Although the use of error rate is completely adequate for our purposes, we have found mutual information to be both satisfactory and most appropriate, given our desire to express and quantify linguistic knowledge within an information-theoretic framework. Specifically, mutual information is used to evaluate the overall performance of trees and to relate what we learn from regression analysis to results from lexical partitioning studies presented in the next chapter. Moreover, maximum mutual information is the criterion used for selecting attributes in computing individual node splits.

It is worthwhile examining some of the details involved in using mutual information in the current context. In the following, we will first develop mutual information

growing trees to a maximum depth. Then, according to a pre-defined criterion that measures the tree's complexity and accuracy (typically classification accuracy), trees are pruned in order to simply them.

as a criterion for evaluating the overall performance of the tree. We will then specialize the result that is obtained as a measure for evaluating an individual node split. Specifically, the maximum mutual information criterion is specified for splitting nodes of a tree grown for categorical response variables. We then define a criterion for splitting nodes of trees grown for continuous responses. This latter criterion is based on the notion of maximizing the amount of node variance that is explained by a split. It is compared and contrasted with criterion of maximum mutual information.

Let the ensemble $A = \{a_j\}$ denote the set of acoustic responses associated with the sample X. For now, we assume a_j to be discrete, i.e.,

$$a_j \in \{b_1, b_2, \ldots, b_K\},$$

where b_k are qualitative values that a_j may take on. Associated with A is a probability assignment $P_A(b_k)$. Further, let T denote the set of nodes in a tree, and $\tilde{T}$ be the subset of T that are leaves. At the completion of the procedure that grows a tree, each token in X will be assigned to at least one element in T, and to exactly one of the elements in $\tilde{T}$. We will let N_T denote the ensemble of values that represent the assignment of tokens to the various elements of T; that is, $N_T = \{t_j\}$, where $t_j \in T$. We may associate with N_T the probability assignment $P_{N_T}(t)$, where each probability value is the proportion of tokens originally contained in X that have been assigned to node t.

Since predictor variables used to split nodes of the tree represent aspects of a phoneme's contextual environment, the objective is to define a measure such that, when optimized, tokens in the learning sample will be partitioned into "meaningful" contexts. *Meaningful,* in the present sense reflects our desire to have the tree provide a set of rules that accurately predict the acoustic realization of a phoneme given a description of its context. In information theoretic terms, the objective in growing a tree is to determine a set of contexts that (on average) convey the most information concerning a token's acoustic realization, or, more concisely, maximize the *average* *mutual information,* $I(A; N_{\tilde{T}})$. In $I(A; N_{\tilde{T}})$, $N_{\tilde{T}}$ denotes the ensemble of values that represent the assignment of tokens to the terminal nodes of the tree.

For $\vec{x_j} \in X$, this average mutual information is given by (4.2),

$$I(A; N_{\tilde{T}}) = \sum_{k=1}^{K} \sum_{t \in \tilde{T}} P_{N_T}(t) P_{A|N_T}(b_k|t) \log \frac{P_{A|N_T}(b_k|t)}{P_A(b_k)}, \tag{4.2}$$

where $P_{N_T}(t)$ and $P_A(b_k)$ are defined as above. The conditional probability, $P_{A|N_T}(b_k|t)$ is the node probability assignment for the elements of A. In other words, it is the distribution A for node t. All logarithms used in (4.2) and subsequent calculations are computed in Base 2 unless specifically noted.

As it is written in (4.2), it is perhaps not apparent that average mutual information is a logical measure to be maximized. A more useful expression is obtained by writing the definition for mutual information in an alternative form (see Appendix B for the details of this manipulation). In (4.3), the average mutual information is more readily seen as the average information concerning the acoustic realization of a phoneme that is provided by having knowledge of its context. This quantity is expressed as the difference in entropies given in (4.3),

$$I(X; N_{\tilde{T}}) = H(A) - H(A|N_{\tilde{T}}) \tag{4.3}$$

In (4.3), $H(A)$ is the average self-entropy for the ensemble A, and is defined in (4.4),

$$H(A) = -\sum_{k=1}^{K} P_A(b_k) \log P_A(b_k). \tag{4.4}$$

The quantity $H(A|N_{\tilde{T}})$ is conditional entropy averaged over the joint ensemble $(A; N_{\tilde{T}})$, i.e.,

$$H(A|N_{\tilde{T}}) = -\sum_{k=1}^{K} \sum_{t \in \tilde{T}} P_{A|N_T}(b_k|t) P_{N_T}(t) \log P_{A|N_T}(b_k|t). \tag{4.5}$$

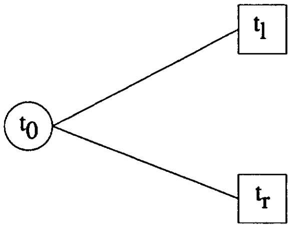


Figure 4.5: Subtree depicting a node that split during the tree growing process. Node t_0 is the subtree's root. Nodes t_l and t_r are its descendants, which are terminal.

The expression given in (4.3) states that the amount of information concerning the acoustic realization of a phoneme provided by its being assigned to one of the terminal nodes of the tree is computed as the *a priori* entropy, computed over the ensemble of acoustic realizations contained in the overall sample, reduced by a weighted sum of entropies corresponding to the individual terminal nodes of the tree. Note also that this information reduction can be stated in relative terms. Specifically, we may define the *Percent Information Extracted* (or PIE) as (4.6),

$$PIE(A; N_{\tilde{T}}) = \frac{H(A) - H(A|N_{\tilde{T}})}{H(A)} \times 100\% \qquad (4.6)$$

In the statement of results of experiments below, we will find the measure defined in (4.6) useful.

4.2.3.2 Maximum Mutual Information as a Criterion for Splitting Nodes

To obtain the expression for mutual information that is to be optimized by splitting individual nodes, we simply recognize that a node and its descendants represent a subtree, and that mutual information computed for the subtree is the corresponding measure to be maximized. In other words, we will specialize the result and interpretation given to (4.3) above for the case of the subtree shown in Figure 4.5.

For the subtree shown in Figure 4.5, $\tilde{T} = \{t_l, t_r\}$ is its set of terminal nodes. Further, we may "shift our frame of reference" and observe that the probability assignment for $\tilde{T}$ is given by

$$P_{N_T}(t) = \left\{ \frac{n_{t_l}}{n_{t_0}}, \frac{n_{t_r}}{n_{t_0}} \right\}.$$

That is, node t_0 is the "root" node in our shifted frame of reference, and the nodes t_l and t_r are leaves. The quantities n_{t_0}, n_{t_l}, and n_{t_r} are the numbers of tokens that comprise the nodes t_0, t_l, and t_r respectively. Finally, mutual information for this subtree is given in (4.7),

$$I(A; t_l, t_r) = h(A|t_0) - \frac{n_{t_l}}{n_{t_0}} h(A|t_l) - \frac{n_{t_l}}{n_{t_0}} h(A|t_l) \tag{4.7}$$

where

$$h(A|t) = \sum_{k=1}^{K} P_{A|N_T}(b_k|t) \log P_{A|N_T}(b_k|t)$$

The expression given in (4.7) is identical to one of the splitting criteria proposed by Breiman *et al.* (1984).

4.2.3.3 Node Splitting Criterion for Continuous Responses

The criterion used to evaluate a regression tree computed for continuous response variables are analogous to those that apply in the case where the response is categorical. What differs are the details of the individual calculations. Specifically, we will use a measure of *variance explained* by the tree as a substitute for *mutual information.*

As mentioned above, the predicted value for a node is the sample mean of A for the tokens of X that the node contains. We may define the *node variance*, $r(t)$, as the corresponding measure of the nodes quality of prediction, i.e., as analogous to $h(A|t)$. The expression for $r(t)$ is given in (4.8),

$$r(t) = \frac{1}{n_t - 1} \sum_j (a_j - \bar{a}(t))^2. \tag{4.8}$$

In (4.8), the values of j included in the summation are those for which tokens in X are assigned to t (i.e., $\{j|\vec{x_j} \in t\}$), n_t denotes the number of these tokens, and $\bar{a}(t)$ is their sample mean.

The variance explained for the tree is defined in (4.9),

$$V(X;\tilde{T}) = r(t_0) - \sum_{t \in \tilde{T}} P_{N_T}(t) r(t). \tag{4.9}$$

The corresponding node splitting criteria is given in (4.10),

$$V(X;t_l,t_r) = r(t) - \frac{n_{t_l}}{n_{t_0}} r(t_l) - \frac{n_{t_r}}{n_{t_0}} r(t_r). \tag{4.10}$$

In (4.10), t_0 denotes the node to be split.

4.2.3.4 Selecting the Best Node Split

The process of selecting the best split for a given node is a procedure in which each predictor variable is tried. A split is computed and evaluated using either (4.7) or (4.10). The split using the predictor variable for which the splitting criterion is maximized is the one chosen for that node.

It should be noted that maximizing mutual information or the variance explained at each of the nodes of a tree does not necessarily produce a tree with maximum mutual information. The procedure that we have defined is an instance of what is known as a "greedy" algorithm. The procedure for optimizing expression (4.7) at each node tends to reduce mutual information quickly, but is not guaranteed to converge to a globally optimal solution.

4.2.3.5 Simplifying Trees

In the use of regression trees, there is generally a compromise between the size or complexity of the tree (typically measured as the number of terminal nodes) and its performance. Trees that are too small fail to be informative. Trees that are grown too large are biased towards the characteristics of the learning sample, and therefore fail to adequately fit data that is unseen. In this way regression methods used to formulate allophonic rules are not much different from the method of introspection used by a human. That is, if a set of rules that a person formulates are too general, they will fail to be linguistically significant. If a person tries to capture too many details in a rule set, the rules formulated will not generalize.

The policy and procedures for obtaining the right-sized tree remains an open question. There are two approaches. The first is to define a *stopping rule.* A condition, that if satisfied during tree growing, will cause the procedure to halt. The alternative is to grow trees to some maximum depth, and then prune nodes until an appropriate tree size is obtained. We have adopted the latter approach, and use a procedure defined in Breiman, *et al.* (1984).

4.2.4 Discussion

The proposed treatment of extrinsic allophones is to formulate rules in the form of a statistical prediction. The predictor is obtained from a sample of acoustic observations. We believe this "data-driven" approach to have two major advantages over an approach whereby rules are formulated by hand. First, the use of explicit rules is based on the assumption that allophonic variation may be adequately predicted. Although, in principle, we have little reason to doubt this assumption, in practice the investigator rarely comes in contact with sufficient data to predict all the contextual variants a given phoneme. On the other hand, a data-driven method has the potential for examining a substantially larger amount of data. Secondly, the use of rules that manipulate symbols tacitly assumes that contextual variation can be described in categorical terms. Once again, in principle, we do not doubt the capability to define an adequate inventory of allophones. In practice, however, we feel it best to describe certain contextual modifications in continuous terms (at least given our current knowledge).

The methods outlined in the present section will be illustrated by way of a case study in the next section.

4.3 A Case Study Involving Stops

In the present section two experiments are described in which the above methods were applied to a sample of American English stop consonants: /p,t,k,b,d,g/. In the first experiment, the acoustic realizations of stops were classified according to five categories: *released*, *unreleased, flapped, deleted,* and *glottalized.* In the second experiment, a stop's *release duration* (as defined below) is used as a determiner of its acoustic realization. In both experiments the same seven predictor variables were used: *syllable position, place of articulation*, *voicing*, *previous context*, *following context*, *Stress*, and *s-stop-cluster?*. The last predictor is a binary variable that indicates whether or not a stop is in a cluster with an /s/.

4.3.1 Design of Experiments

4.3.1.1 Source of Data

Data for both experiments were obtained from three databases, denoted TIMIT, HL, and JP, each containing speech read by subjects. Databases TIMIT and HL contain utterances spoken by multiple speakers. Specifically, the TIMIT database contains approximately 2154 utterances spoken by 376 male and female talkers. Each speaker read 8 sentences. Database HL contains 1000 utterances, spoken by 100 talkers, 50 male and 50 female. Each speaker read 10 sentences. Finally, database JP contains 600 utterances read by a single male talker.

All three databases are general-purpose: they have been designed and collected for a wide community of users. The corpus for the TIMIT database was a subset of sentences developed at MIT. The corpus for databases HL and JP were each subsets of the well-known Harvard list of phonetically-balanced sentences. Both corpora were designed to provide speech material of rich and varied phonetic content. Of the two, the TIMIT sentences are more diverse in terms of phonological sequences and sentence types. In particular, the HL sentences consisted mostly of monosyllable words (the average number of syllables per word for this corpus is 1.1). The TIMIT sentences, on the other hand, contain a larger number of polysyllabic words (avg. syl/word = 1.58). Also, HL sentences are all simple declarative sentences. The TIMIT sentences are more of a variety: statements, questions, sentences with embedded syntactic structures, such as clauses and parentheticals. For these reasons, we will focus on data obtained from the TIMIT database in the reporting of the experiments below. These databases are documented in more detail in Appendix A.

Transcriptions

Associated with each of the utterances used in the current study were three time-aligned transcriptions: 1) a *narrow phonetic* transcription, 2) a syllabified *phonemic* transcription, and 3) an *orthographic* transcription. Utterances were originally transcribed with the phonetic transcription. For example, a stop that is marked with a narrow phonetic transcription would include both its closure and its release. The phonetic transcription also included symbols for glottalized stops as well as flaps.

A semi-automatic procedure was used to align phonetic transcriptions with time waveforms. This procedure consisted of two steps. First, an automatic procedure was used to segment and label the waveform (Leung, 1985). Then, the resulting aligned transcription was verified manually. The syllabified phonemic transcription was obtained from a pronunciation dictionary. Although in this thesis a hierarchical representation for an utterance's syllabification has been advocated, the syllabified phonemic transcription consisted of a more traditional string of phonemic symbols with syllable boundary markers ("$") interspersed.[5] This transcription was aligned with the phonetic transcription also using a semi-automatic procedure. Finally, through the phonemic-phonetic alignment, orthograhic transcriptions were aligned with the time waveform.

At this juncture, we should point out a number of problems that arise when using a phonetic transcription in an acoustic study of this size. These problems are related to errors in the phonetic transcriptions themselves, and the lack of consistency among transcribers. Errors, although infrequent in our study, must be expected to occur. The types of errors that we detected in the transcriptions of stop consonants are of a limited number. For example, stop releases that are weak often go undetected. There is also a bias towards transcribing /t/'s as glottal stops when pitch irregularities are detected in the waveform, but in situations where /p/'s and /k/'s possess similar waveform characteristics, they are transcribed as unreleased. Perhaps a more pervasive problem in transcribed databases is maintaining consistency in the transcriptions among the several phoneticians who participate in the segmentation and labelling processes. Although this problem might be lessened to a certain extent with stop consonants, it is reasonable to assume that individual phonetic impressions of a given speech sound will differ.

4.3.1.2 Predictor Variables

The table given in (R 4.7) lists the values that each of the predictor variables may assume:

[5] The locations of syllable boundaries were determined on the basis of the grammar outlined in Chapter 3. Further details concerning the syllabification of words are given in Chapter 5.

	Variable	*Value*
	Place of Articulation:	Labial, Alveolar, Velar
	Voicing:	Voiced, Unvoiced
R 4.7	*Previous Context:*	Affricate, Fricative, Glide, Nasal, Stop, Vowel
	Following Context:	Affricate, Fricative, Glide, Nasal, Stop, Vowel
	/s/-stop Cluster:	Yes, No
	Syllable Position:	Onset, Coda, Affix-1, Ambisyllabic
	Stress:	High, Low, Rising, Falling

With the exception of *Stress*, these values were determined on the basis of the syllabified phonemic transcription.

A stop's stress environment is determined by the phonetic realizations of its surrounding vowels. In particular, we consider a vowel to be unstressed if it realized as *reduced* (i.e., as a schwa), and otherwise stressed. The manner in which a stop's stress environment is determined is described in the table in (4.8).

	Stress Environment:	*Vowel Stress:*	
		Preceding Vowel	*Following Vowel*
R 4.8	High	Non-reduced	Non-reduced
	Low	reduced	reduced
	Rising	reduced	Non-reduced
	Falling	Non-reduced	reduced

4.3.1.3 Acoustic Realizations

For the first experiment, a stop's acoustic realization was determined on the basis of the phonetic transcription. If, for example, an underlying /t/ surfaced as both a closure ([$\text{t}^{\square}$]) and a release ([t]) or simply as a release, it was marked *released.* If it contained only a closure, it was marked *unreleased.* If a /t/ in the phonemic transcription corresponded to a segment that was missing in the phonetic transcription, it was marked as *deleted.* Finally, glottal stops (ʔ) and flaps (ɾ) were marked directly with no further interpretation. The example utterance shown in Figure 4.1 includes examples of these allophonic variants.

For the second experiment concerned with *release duration*, only released stops were considered. Depending on what follows a stop, its release duration may be determined with varying degrees of accuracy. For example, if a stop is pre-vocalic, or preceding a sonorant, then the release duration for a stop is its voice onset time. This period begins with the location of the stop burst and ends at the first sign of periodicity in the waveform. If the stop precedes an obstruent, then release duration is more difficult to determine. For example, if the following phoneme is a fricative, transcribers would attempt to determine when stop aspiration ends and frication for the following fricative begins.

4.3.1.4 Data Analysis

Data were cataloged and manipulated using a statistical analysis software package that we have written for Symbolics Lisp Machines called *SEARCH.* In addition to serving as an engine for doing the various statistical calculations, *SEARCH* provided the capabilities of a *relational database*, allowing the various transcriptions mentioned above to be related to one another. In addition, this software provided exploratory data analysis capabilities for producing summary statistical plots (e.g., histograms, scatter plots, etc) and for probing the individual tokens in order to investigate anomalies (e.g., outliers).

In the presentation of the data, the intent is to provide the reader with an overall impression of the phonetic variability of stop consonants, and the systematic relation of this variability to the proposed set of explanatory variables. Due to the scope and emphasis of this chapter, an exhaustive study is not possible. The main points will be to demonstrate a) the informational content of syllable structure in describing this variability, and b) the effectiveness of the proposed methodology. A significant analysis of stops could constitute a thesis within itself.

4.3.2 Experiment I: Stop Realization as a Function of Local Phonetic and Prosodic Context

In the first of the two experiments, stop realization is considered a categorical variable, assuming the values *released*, *unreleased, flapped, deleted,* and *glottalized.* All of the predictor variables were included as factors in the experiment. Results will be presented in two parts. First, variables will be examined individually in order to provide an overall impression of how each affects a stop's realization. Then results of the analysis using the tree are presented. The tree analysis provides a means of understanding how the variables interact.

For both parts, results will be given in information-theoretic terms. Specifically, we will use mutual information to evaluate the effects that the contextual variables have on stop realization. Specifically, the *Percent Information Extracted* measure developed in the previous section will be used for quantifying the effects of individual variables and for characterizing the overall performance of the tree.

4.3.2.1 Variables Taken Individually

Figure 4.6 shows *Percent Information Extracted* as a function of predictor variable and database. This summary statistic measures the difference in the amount of unextracted information (i.e., uncertainty) prior to and after the value of a particular explanatory variable is known. As one can see in the figure, knowledge of a stop's syllable position, and following context appear to be most pertinent in this respect. Knowing whether or not it is voiced or in a cluster with an /s/ provides relatively little information. According to the above criteria, knowing the place of articulation of a stop and its previous context seems to provide a marginal amount of information, as does knowledge of whether it precedes a stressed vowel.

From the histograms shown in Figure 4.6 one may observe that, with the exception of *place of articulation*, all predictor variables have a larger effect on *Percent*

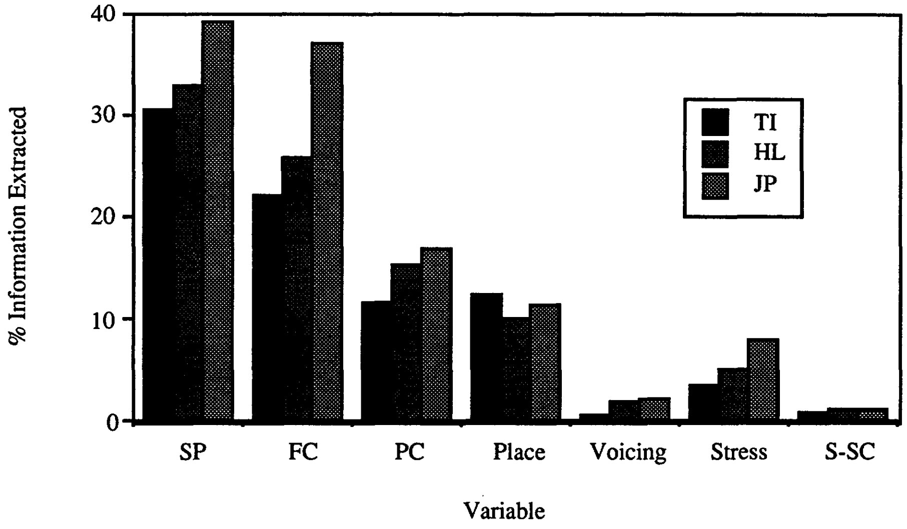


Figure 4.6: Comparison of the effects of each predictor variables as a function of database. Along the vertical axis is Percent Information Extracted, along the horizontal axis is the name of the variable. The abbreviations SP, FC, PC, and S-SC denote Syllable Position, Following Context Previous Context, and /s/-stop Cluster respectively. Values are plotted as a function of database.

Information Extracted for database JP. The effects are the least for tokens in the TIMIT database. This behavior in the data is not totally unexpected. Database JP is a single speaker database. Furthermore, it contains Harvard list sentences read by this speaker, which is simpler speech material. Therefore, the amount of variability one may attribute to differences among speakers, and complexity of speech materials is not present. As a result, a larger percentage of the variability in the acoustic realizations of stops may be explained by their contextual environment.

Having both knowledge of a token's following context and its syllable position is largely redundant, with knowledge of syllable position being slightly more informative. One may observe this in Figure 4.6 where it appears that a stop's *syllable position* and its *following context* are similar in informational content. For example, knowledge of a stop's syllable position provides on average 8.4% more information (30.6% vs. 22.2%) for tokens in database TIMIT, and only 2.1% more information (39.2 vs. 37.1%) for database JP. Figure 4.7 shows histograms for stop realization as a function of (a) *syllable position* and (b) *following context.* Stops in the outer onset position are practically always released (97% of the time), whereas in the coda position, they are mostly unreleased (43% vs. 31% released). One also sees that large percentages of stops are released when they are followed by vowels and glides (a necessary but not sufficient condition for the stop belonging to the onset). Further, when followed by classes of sounds that it cannot precede in the onset of the syllable, a stop tends to be either unreleased or deleted.

We should note that our reporting of this latter set of facts requires an important qualification. The classes of sounds that a stop cannot precede in the onset of the syllable include nasals, affricates, and other stops. These sounds all have the feature [- *continuant*]. At first glance, these data seem to indicate that when speakers are required to produce two stops in succession, they often do not release the initial one. However, it is not always straightforward to determine on the basis of acoustic evidence whether the first stop is actually unreleased in this circumstance.[6]

[6] Although not presented here, in a preliminary study of stops found in *non-continuant – non-*

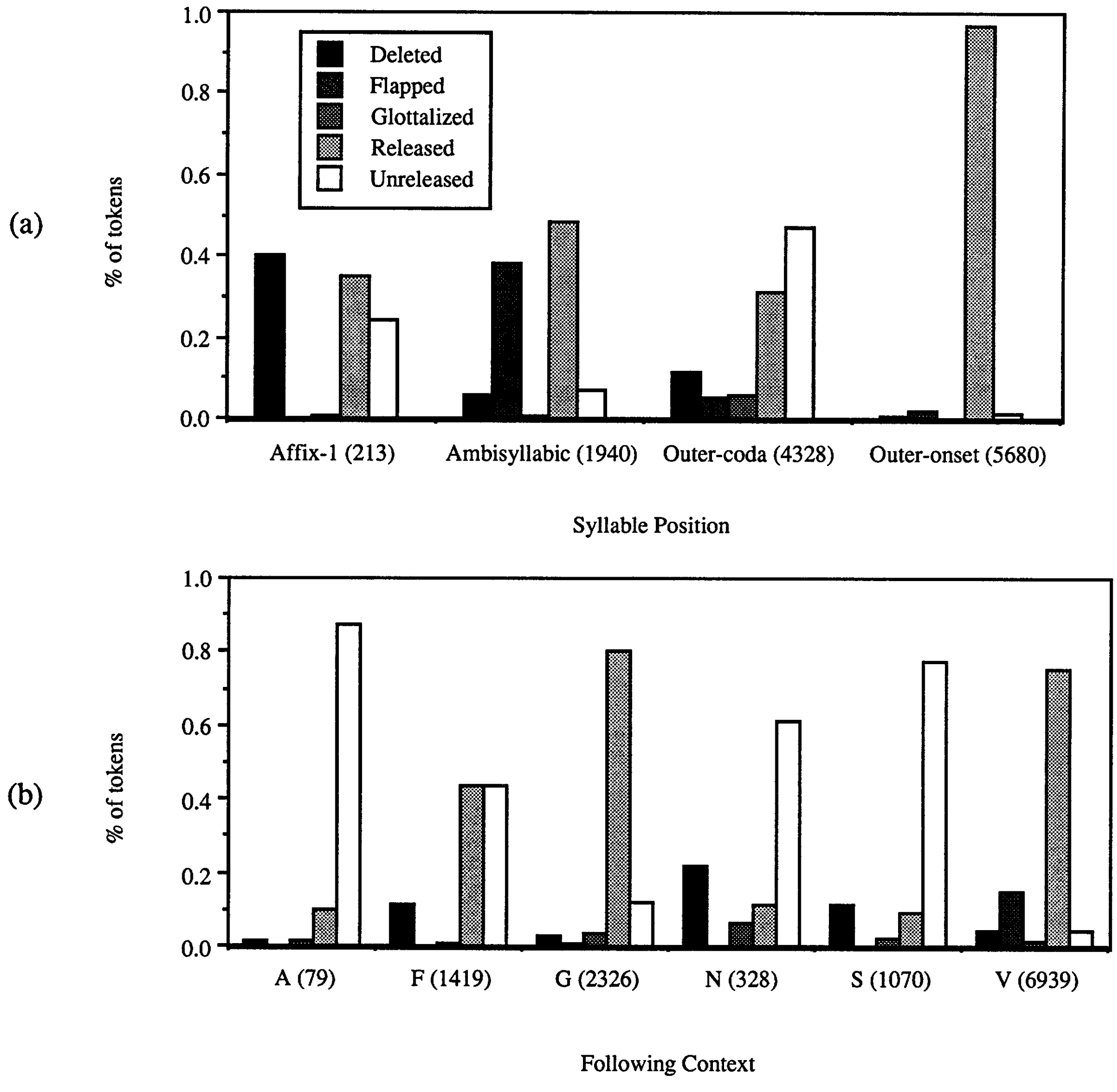


Figure 4.7: Histograms of stop realization as a function of (a) syllable position and (b) following context for tokens in database TIMIT. In (b), the abbreviations A, F, G, N, S, and V denote Affricate, Fricative, Glide, Nasal, Stop, and Vowel respectively. Shown in parenthesis are the number of tokens in each syllable position.

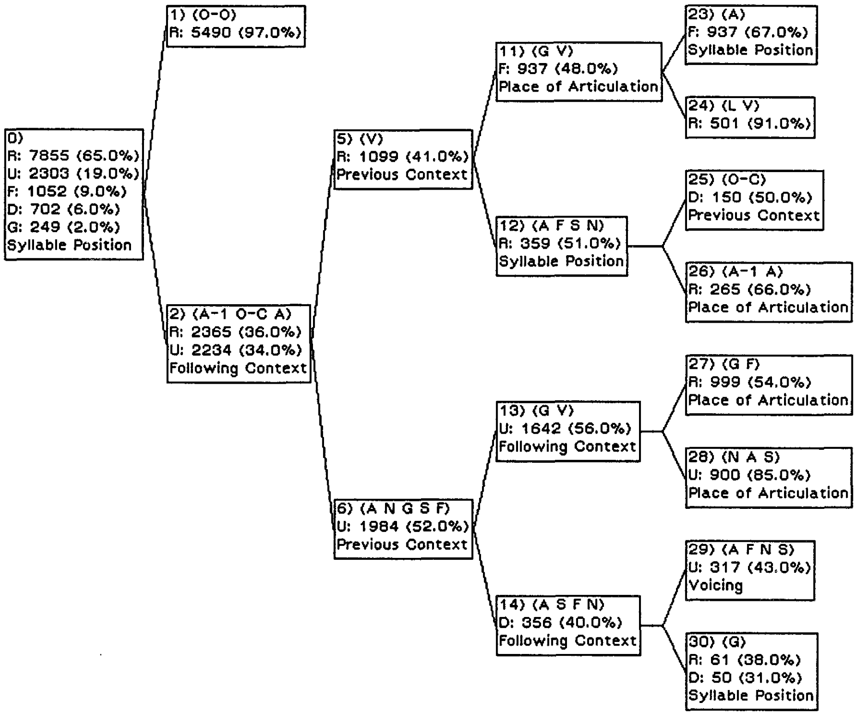


Figure 4.8: Tree obtained using maximal mutual information method for the stop realization data. Nodes are shown as boxes, see text for a description of the information contained in each box.

4.3.2.2 Analysis of Stop Realization Using the Tree

Next, the results of using the tree regression method to analyze the stop realization data are presented. The method of maximal mutual information described in Section 4.2 was used to grow a tree based on the entire data set of 12,000 tokens. All seven predictor variables considered in the previous section were used in obtaining the tree. The tree was grown to maximal length. Nodes were then identified that contributed little to overall tree performance and subsequently pruned. After pruning, the tree contained a total of 62 nodes, 32 of which were terminals.

A major portion of the tree is shown in Figure 4.8. Nodes are shown as boxes, where inside each box is information both about the node itself and the subset of the

continuant sequences, the closure duration of the first stop was substantially increased.

learning sample contained in that node. The information includes: the node label (which is an integer), the value (or values) of the variable used to split the node's parent, and a description of the distribution of tokens contained in the node's sample (i.e., the number of tokens of each stop realization along with its percentage shown in parentheses). For all but the root node, where the proportions of all stop realizations are listed, the proportions that explain the most significant classes are given in the other nodes of the tree. Finally, for all but the tree's terminal nodes, the name of the variable used for splitting the node's sample is displayed last in each box. In order to display all of the information required to describe a node, values of predictor variables have been abbreviated (e.g., o-o denotes OUTER-ONSET), as have the categories of stop realization (e.g., (R denotes *released*)..

Some of the tree's nodes are more informative than others. Therefore, an enumeration of the tree's entire contents would not be the most expedient means of summarizing it. Instead, the approach will be to first describe its overall structure, and then to trace along a couple of its more interesting paths. The walks along the tree's structure will serve as a means of of demonstrating how information is conveyed. Following this exercise, a number of summary statistics will be presented that also provide a reflection of how information is factored among the tree's contents.

Overall Structure

Figure 4.8 shows the *skeleton* of the stop realization tree. Its root node contains 12161 tokens, which are distributed as follows:

7855	Released
2303	Unreleased
1052	Flapped
702	Deleted
259	Glottalized

As one may observe from the portion of the tree shown in Figure 4.8, it is highly unbalanced. The split that takes place at its root node (node 0) partitions the sample according to the *syllable position*; those tokens placed in node 1 are in the outer-onset position and the tokens belonging to the remaining syllable positions are placed in node 2. From the above examination of first order statistics, one may recall that syllable position extracts a relatively large amount of information, as is evidenced in the tree by the fact that this variable is used first. There are 5490 tokens in the outer-onset position (node 1), and in this syllable position approximately 97% of the stops are released. Of the remaining stops (there are 190), 99 are flapped, while 69 and 22 tokens respectively were marked as unreleased and deleted.

Tokens that are marked *unreleased* and *deleted* in the outer-onset position of the syllable should be considered anomalous. A closer examination of these tokens revealed most to have weak releases, or to be realized as fricatives (i.e., incomplete closures) of some sort.

The types of stop realizations that are contained in node 2 are considerably more diverse, having an entropy of .97 bits (as compared with an entropy of .12 bits computed for node 1). As a result, the subtree that this node dominates is substantially larger. The distribution of tokens contained in this node is as follows:

2365	Released
2234	Unreleased
953	Flapped
680	Deleted
249	Glottalized

Rather than describing the subtree dominated by node 2 in its entirety, two of its paths will be traced. Specifically, the paths leading to the nodes dominated by tokens *flapped* and nodes that are dominated by tokens that are *unreleased* will be followed. The reader is asked to refer to Figure 4.8.

Example 1: Flapping

The split at Node 2

At node 2, the partition that maximizes mutual information is the one that selects tokens whose following context is a vowel and separates them from the remainder. These tokens are placed in Node 5. Of the original 953 flaps contained in node 2, 940 satisfy this condition.

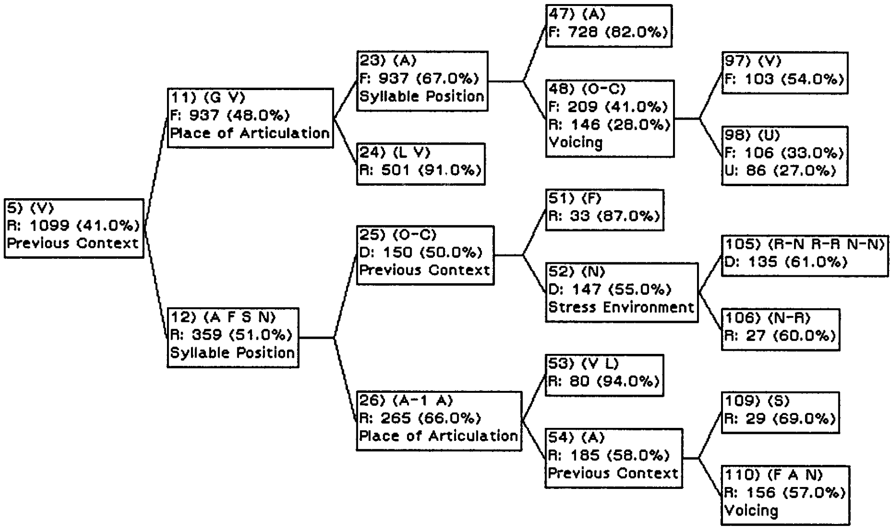


Figure 4.9: Portion of the tree shown in above dominated by node 5. Node 5 contains 940 flaps in addition to 1099 released stops.

The split at Node 5:

The subtree with node 5 as its root is shown in Figure 4.9. Node 5 is dominated by released stops (1099 out of 2657 tokens). It also contains 940 stops that are realized as flaps. Attention will be focused on those branches of the tree that are directed towards distinguishing these two classes of acoustic realizations.

A split of node 5 using the variable *Previous Context* results in 937 (99.7%) of these flaps being placed in node 11. Tokens placed in node 11 are preceded by either glides or vowels.

Node 11:

Node 11 is the first node along the path for flapped stops in which a majority of tokens having this realization emerges. This node contains a total of 1951 tokens, of which 937 are flapped (48%). Of the remaining stops, 740 are released. The split that occurs at this node is to select stops that are alveolar and place then into node 23.

Node 23:

The tokens contained in this node are alveolar stops preceded by a vowel or glide, and that are prevocalic. There are 1398 of these tokens, 937 of which are flapped. Of the remaining tokens, 239 are released, 147 are unreleased, 63 are glottalized, and 12 deleted. Node 23 is split according to syllable position. Placed in Node 47 are tokens that are ambisyllabic.

Node 47:

The end of this path for flaps is node 47. At this point, a contextual variation rule can be formulated corresponding to the path just traced. Such formulation could take a number of forms. For example, a rather direct formulation would be to simply write down the *conditional probability*:

R 4.9

$$P\{a = Flapped|\vec{\alpha}, \gamma_1, \gamma_2, \sigma, S\} = .82$$

where

$$\begin{aligned}\vec{\alpha} &= \{t, d\}\\ \gamma_1 &= \{\textsc{Vowel}, \textsc{Glide}\}\\ \gamma_2 &= \{\textsc{Vowel}\}\\ \sigma &= \textsc{Outer-onset} \wedge \textsc{Outer-coda}\end{aligned}$$

Note that for the "allophonic rule" given in (R 4.9), the conditions are under specified. That is, the stop's voicing is left unspecified, as is its stress. The tree has discovered in the data the context under which these variables are not necessary.

Example 2: Unreleased Stops

As another example, a path will be traced through the tree for stops that are unreleased. Like flaps, the path traced for unreleased stops will begin at node 2 (shown in Figure 4.8). The split at this node partitions tokens according to whether or not their following context is a vowel. The majority of the unreleased stops that are in this node (1984 out of 2234) do not precede vowels. They are placed in node 6.

Node 6:

The distributions of tokens among the various stop realizations in node 6 is as follows:

1266	Released
1984	Unreleased
13	Flapped
420	Deleted
141	Glottalized

One may observe that unreleased stops dominate this node, but there are also a considerable number of released stops. At node 6 these tokens are split according to their previous context. The majority of the unreleased stops (1642/1984) have as their previous context a glide or a vowel. These tokens are placed in node 13 (see Figure 4.10).

Node 13:

The distribution of tokens at node 13 is as follows:

1097	Released
1642	Unreleased
13	Flapped
64	Deleted
109	Glottalized

The majority once again are unreleased stops. The effect of the split at the previous node (node 6) was to remove a number of the deleted stops. Out of the original 420 tokens that were deleted in node 6, only 64 are marked deleted in node 13.

The split at node 13 looks at the variable *Following Context.* It partitions the tokens whose *Following Context* is either a *nasal*, *affricate* or a *stop* into the node 28.

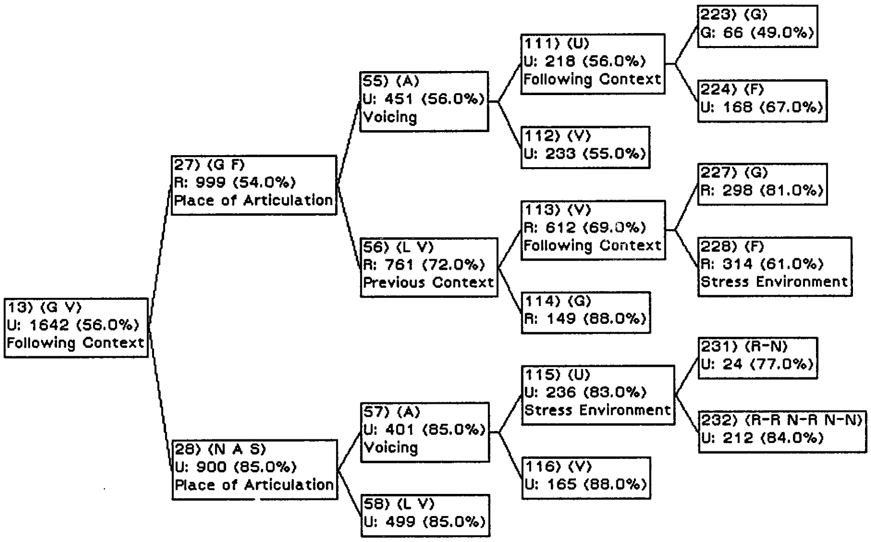


Figure 4.10: Portion of the above tree dominated by node 13.

Node 28

The distribution of tokens at node 28 is as follows:

98	Released
900	Unreleased
0	Flapped
27	Deleted
33	Glottalized

Unreleased stops are in a clear majority. Node 28 itself could have been left as a terminal node (in other words, not split). However, a an improvement in entropy reduction results from its split of velars and labials into node 58 and its placement of alveolars into node 57.

Node 58:

Node 58 is a terminal node, we may now formulate an allophonic a realization rule representing its path. Once again, a *conditional probability* will be written:

R 4.10

$$P\{a = Unreleased|\alpha, \gamma_1, \gamma_2\sigma, S\} = .85$$

where

$$\begin{aligned}\vec{\alpha} &= \{b, p, g, k\}\\ \gamma_1 &= \{\text{Vowel}, \text{Glide}\}\\ \gamma_2 &= \{\text{Stop}, \text{Nasal}, \text{Affricate}\}\\ \Sigma &= \{\text{Outer-coda, Affix-1, Ambisyllabic}\}\end{aligned}$$

This rule states that alveolar and labial stops followed by other non-continuant sounds tend not to be released. This rule too is underspecified (as indicated by both the absence of certain conditional variables and the multiple values for the variables that are present). As the path starting at node 2 of the tree was traced, it was found that the variables stress, voicing, and syllable position do not play a role in the statement of conditions for unreleased stops. It is postulated that the variables *stress* and *voicing* fail to play a role because they do not have a large statistical effect on the data. *Syllable Position* does not play a role because it has already been factored in due to the local phonetic context specification: stops that precede nasals, stops, and affricates in English can only be in either the outer-coda or affix-1 syllable positions.

4.3.2.3 Overall Tree Performance

The examples just presented illustrate how knowledge of context is factored into the structure of the tree. At this point, attention is focused on evaluating the tree's performance. There are two approaches to this problem. The first would be to take path tracings and formulate transformational rules. Then traditional linguistic evaluation metrics could be applied to determine how accurate the tree is, and to what extent is capturing important generalizations. Alternatively, one could rely on the information theoretic measures developed in Section 4.2. The latter approach is taken here.

There were a total of 12,161 tokens used in growing the stop realization tree with a distribution of realizations as follows:

7855	Released
2303	Unreleased
1052	Flapped
702	Deleted
259	Glottalized

The *a priori* entropy, $H(A)$, of this distribution is 1.52 bits. After growing and pruning the tree, the resulting conditional entropy is $H(A|\tilde{T})$ with a reduction of entropy of approximately 56%. For this tree the classification accuracy was also calculated and found it to be approximately 84%.

As mentioned above, there are 32 terminal nodes in this tree. The distributions of tokens assigned to these terminal nodes has an average entropy fo 3.351 bits. The *a posteriori* entropy for this tree, $H(\tilde{T}|A))$, is 2.58 bits giving a reduction of *a posteriori* entropy of approximately 23%. It is interesting to break the a posteriori entropy into its component parts $(H(\tilde{T}|a))$, where

$$a \in \{\ \textit{released, unreleased, flapped, deleted,glottalized}\}.$$

These results are given in (R 4.11):

	Realization	*Entropy*	*Perplexity*
	Flapped	*.128*	*1.09*
R 4.11	*Deleted*	*.189*	*1.14*
	Glottalized	*.0628*	*1.04*
	Unreleased	*.757*	*1.69*
	Released	*1.44*	*2.71*

These numbers indicate, on average, how much information a particular acoustic realization provides about the underlying phonological specification. Also given in this table is the *node perplexity*, the *antilog* of these entropy values. It is observed, for example, from the values given in this table, that, with the exception of glottalized stops, an observance of a flap provides the most information concerning a stop's underlying context. From the path tracings above, it is recalled that these stops are between two sonorants, and in an ambiguous syllable position. Observation of a stop that is released actually provides the least amount of information. For example, it is well established that stops in the onset position are practically always released. However, stops appear to be released in other syllable positions as well, as this high perplexity indicates. It will be seen in the discussion of the results of experiment II that the duration of the stop's release lowers this uncertainty.

Prior to concluding the discussion of this experiment on stop realization, it is interesting to examine briefly what happens when tokens taken from databases HL and JP are passed through the tree grown for the TIMIT data. For reasons given earlier, one might expect the *Percent Information Extracted* computed under these circumstances to improve. Moreover, given that the effects on stop realization due to speech material and inter-speaker differences are, in a way, accounted for in these alternative databases, one might consider the performance observed by passing their tokens though the TIMIT tree to give some indication of how well we could expect the tree to perform given the predictor variables considered in this study.

We found the improvement in *Percent Information Extracted* to be marginal when tokens from database HL were passed through the TIMIT tree. *Percent Information Extracted* for TIMIT tokens in the TIMIT tree was calculated to be 56%. For the tokens from database HL, it improved to 57%. A more substantial increase, from 56% to 74%, was found for the case where tokens from database JP were passed through the TIMIT tree, an improvement of 18%. On the basis of these results, we might postulate that the variation due to speaker variability is substantial. Furthermore, that the additional 25% of information thus far unextracted may be due to one of two factors: a) non-optimal tree performance or b) the omission of crucial predictor variables.

4.3.3 Experiment II: Stop Release Duration

In the previous experiment, a relatively large percentage of stops in the coda position (31%) are released. This observation casts doubt as to whether a purely categorical description of an utterance at the phonetic level is sufficiently accurate. That is, if one were to posit the rules

R 4.12

$$\begin{array}{lcl} \text{ONSET} & \longrightarrow & p^h|t^h|k^h \\ \text{CODA} & \longrightarrow & p^\square|t^\square|k^\square \end{array}$$

as well-formedness constraints on the phonetic representation, the resultant description of the facts would be incorrect at least 31% of the time. Alternatively, as has been suggested in this thesis, rules could impose tighter constraints on a phonetic representation if parameters used to characterize an utterance were allowed to vary along a continuum. In the present experiment, this idea is explored by constructing a regression tree on the basis of stop release duration.

In the current study, the same predictor variables used in the previous experiment are considered once again. A tree was obtained using the method outlined in Section 4.2 for continuous response variables. In the discussion that follows, the tree's structure and contents are described.

4.3.3.1 Overall Structure

A portion of the tree obtained in this experiment is shown in Figure 4.11. Nodes, once again, are shown as boxes, where inside each box is information both about the node and the subset of the learning sample contained in that node. Included

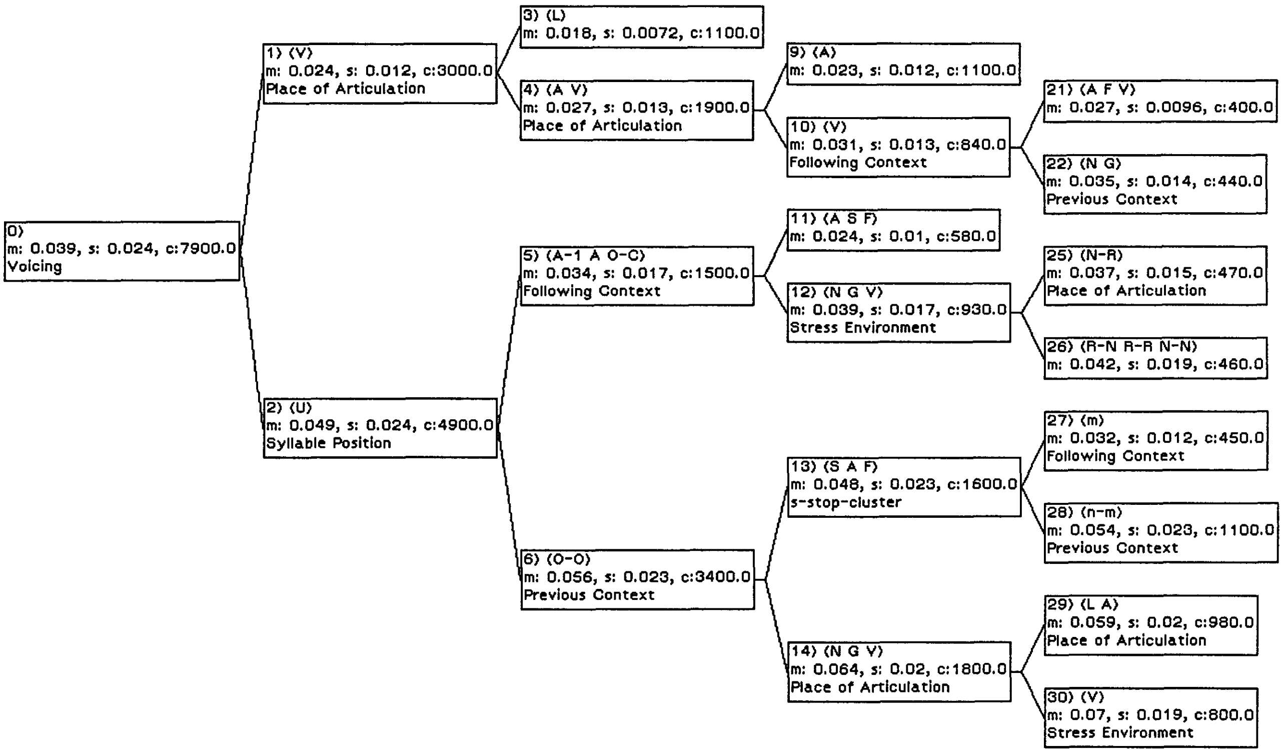

0)
m: 0.039, s: 0.024, c:7900.0
Voicing
1) (V)
m: 0.024, s: 0.012, c:3000.0
Place of Articulation
2) (U)
m: 0.049, s: 0.024, c:4900.0
Syllable Position
3) (L)
m: 0.018, s: 0.0072, c:1100.0
4) (A V)
m: 0.027, s: 0.013, c:1900.0
Place of Articulation
5) (A-1 A O-C)
m: 0.034, s: 0.017, c:1500.0
Following Context
6) (O-O)
m: 0.056, s: 0.023, c:3400.0
Previous Context
9) (A)
m: 0.023, s: 0.012, c:1100.0
10) (V)
m: 0.031, s: 0.013, c:840.0
Following Context
11) (A S F)
m: 0.024, s: 0.01, c:580.0
12) (N G V)
m: 0.039, s: 0.017, c:930.0
Stress Environment
13) (S A F)
m: 0.048, s: 0.023, c:1600.0
s-stop-cluster
14) (N G V)
m: 0.064, s: 0.02, c:1800.0
Place of Articulation
21) (A F V)
m: 0.027, s: 0.0096, c:400.0
22) (N G)
m: 0.035, s: 0.014, c:440.0
Previous Context
25) (N-R)
m: 0.037, s: 0.015, c:470.0
Place of Articulation
26) (R-N R-R N-N)
m: 0.042, s: 0.019, c:460.0
27) (m)
m: 0.032, s: 0.012, c:450.0
Following Context
28) (n-m)
m: 0.054, s: 0.023, c:1100.0
Previous Context
29) (L A)
m: 0.059, s: 0.02, c:980.0
Place of Articulation
30) (V)
m: 0.07, s: 0.019, c:800.0
Stress Environment

Figure 4.11: Tree obtained using maximal variance reduction method for the stop release duration data. Nodes are shown as boxes. See text for a description of the information contained in each box.

in each box are: a node label, which is, an integer, the value (or values) of the variable used to split the node's parent, and information describing the distribution of tokens contained in the node's sample. Specifically, the box includes is the sample mean of the distribution, the estimated standard deviation, and the number of tokens contained in the node's sample (these values are denoted *m, s,* and *c* respectively). Finally, for all but the terminal nodes, the name of the variable used for splitting the node's sample is displayed.

The tree obtained for release duration (shown in Figure 4.11) is considerably more balanced than the tree obtained in Experiment I (Figure 4.8). Voicing is used to split the root node (node 0) with voiced stops being placed in node 1 and voiceless stops in node 2. On the side of the tree containing voiced stops, nodes 1 and 4 are split on the basis of place of articulation, with velar voiced stops (node 10) having, on average, the longest release duration (31 msec.) and labials (node 3) having the shortest (18 msec.). The nodes containing labial and alveolar voiced stops (node 9) are terminals. The node containing velars (node 10) is split further, first according to following context.

The release duration of voiceless stops (node 2) is twice as long as the that of their voiced counterparts (49 msec. vs. 24 msec.). In addition, the standard deviation of the sample of voiceless stops is also twice as large (24 msec. vs. 12 msec.). Voiceless stops are split first on the basis for syllable position. Voiceless stops in the onset position (node 6) have approximately one and half times as long a release duration as stops in the other syllable positions (node 5).

Voiceless stops in the non-initial syllable positions (node 6) are split on the basis of *Following Context.* These stops that precede nasals, glides, and vowels (node 12) have a longer release duration than voiceless stops preceding affricates, other stops, and fricatives. Further, stops in a non-falling stress environment (node 26) have the longest release duration of the stops contained in node 5. This is perhaps an indication of these stops being resyllabified as onset stops.

Stops in the onset position (node 6) are split according to *Previous Context*, with stops preceded by obstruents: other stops, affricates, and fricatives, having a shorter release duration (48 msec.) than stops preceded by nasals, glides and vowels (nodes 14). Of those stops that are preceded by obstruents, those in clusters with /s/ have the shortest release durations (32 msec). Of those onset stops that are preceded by sonorants (node 14), velars have the longest release durations (70 msec) whereas alveolars and labials have release durations on average of approximately 60 msec.

4.3.3.2 Overall Tree Performance

The overall structure of the tree suggests that there is a considerable amount of information not conveyed in an entirely qualitative description of a stop. There are at least two ways in which one may quantify this assertion. The first is in direct information theoretic terms. For example, based on results of information theory for continuous ensembles, mutual information was calculated, $I(A;\tilde{T})$ for the tree given in Figure 4.11 to be .84 bits. Unfortunately, we can not calculate *Percent Information Extracted* in this case since the notion of *a priori* entropy, $H(A)$, is not as meaningful for the continuous case. We also calculated the *Percent Variance Explained* for this tree using the learning sample. We found it to be approximately 60%.

4.4 Summary

In summary, extrinsic allophones have been defined as phonetic alternations attributed to the structural context of a phoneme in addition to its local phonetic environment. A method for treating extrinsic allophones has been proposed. The suggestion is to formulate a set of rules that predict this behavior in the form of a statistical regression. The regression method used is based on binary regression trees.

Binary regression trees have been selected for two reasons. First, they are a statistical optimization procedure. It has been suggested in this chapter that extrinsic allophones represent linguistic behavior that are the result of principles that are less well understood than those that characterize coarticulatory variation. The use of the statistics is a means of compensating for this relatively poor understanding. Unlike a number of other statistical techniques, binary trees are easy to interpret. Moreover, as has been demonstrated in the experiments using this method, contextual variation rules may readily be formulated.

The results of two experiments were reported. The experiments demonstrate the binary tree method as well as suggest a treatment of certain kinds of extrinsic allophonic behavior in direct acoustic terms. For example, stop consonants are practically always released when in the syllable initial position, but they may also be released when syllable final. In order to distinguish these two structural positions, the release duration of the stop (a continuous variable) needs to be considered.

Chapter 5

Syllabic Constraints on the Lexical Representation

This chapter marks a change in our conception of the syllable. In the previous three chapters, the syllable is viewed as a descriptive framework for stating phonological constraints. In the present chapter, we study the potential role of the syllable in facilitating spoken word recognition and lexical access (i.e., the processes of retrieving word entries from the lexicon on the basis of an utterance's acoustic description).

Chapter 2 outlined a number of issues that a theory of syllable structure is to address. Included was the question of how the syllable is to constrain sound patterns described at both the acoustic and the surface-phonemic levels of phonological representation. In Chapter 3, a model of English syllable structure was proposed for this purpose in the form of an augmented context-free grammar. The proposed grammar is essentially an extension of earlier theories of syllable structure, incorporating ideas that may be discerned in the works of Fudge (1969), Selkirk (1982) and Clements and Keyser (1983). In the present chapter, we argue that the grammar should form the basis of procedures for translating the acoustic representation of an utterance into a form that is suitable for matching against lexical entries (which we assume to be stored in the form of distinctive feature matrices). The procedures themselves are outlined in Appendix D.

Our arguments are based on empirical evidence. In particular, we present two experiments that are directed at providing a quantitative understanding of syllabic constraints at the surface-phonemic level. The results of these experiments provide justification for a model of spoken word recognition that is proposed. Both experiments make use of a computer-readable dictionary of syllabified words.

The first experiment, described in Section 5.1, provides partial justification for the syllable theory outlined in Chapter 3. Specifically, this experiment examines collocational restrictions within the syllable and addresses the problem of determining an appropriate syllable structure topology. Mutual information is used as a means of estimating the magnitude of collocational constraints (i.e., the dependance that the occupant of one position within the syllable has on the occupants of others). Mutual information statistics gathered from a sample containing some 5500 syllables form the basis of a cluster analysis in which a hierarchical structure is determined for the syllable's terminal elements.

In the second experiment, described in Section 5.2, we examine the possible advantages that structuring an utterance into syllabic units has during the process of word recognition. The "syllable lexicon" considered in the first experiment, is itself studied. The objective is to ascertain the relative importance of the various parts of the syllable as a means of providing lexical constraint. In this experiment, a series of lexical partitions is examined, where for each, different parts of the syllable are underspecified. Statistics of the resulting equivalence classes are analyzed and compared.

5.1 The Study of Collocational Constraints

5.1.1 Constituents as Phonological Domains

The basic premise underlying the use of immediate constituent grammars in phonology (see Chapter 2) is that constituents will serve as domains over which phonological generalizations may be stated. Similar statements may be made concerning the suprasegmental tiers of autosegmental phonology. The class of generalizations of current interest are restrictions on phoneme sequences: so-called collocational constraints.

Several authors have claimed that the relative "tightness" of collocational constraints are not uniform across the syllable (cf., Pike and Pike, 1947; Fudge, 1969; Selkirk, 1982). Instead, the selection of phonemes that occupy syllable-initial clusters, for instance, depends more on other occupants of the syllable's onset than on the phonemes that occupy more remote places in the syllable's hierarchy. Similar statements may be made concerning syllable-final phonemes. What has emerged from these observations is what are known as the *immediate constituent* principles of phonotactics and the internal structure of the syllable proposed by Fudge (1969) and Selkirk (1982).

In contrast to the position of a highly enriched immediate constituent structure for representing the syllable, Clements and Keyser (1983) envision a much "flatter" syllable structure. These authors state:

> As far as we have been able to determine, there is no linguistic evidence suggesting that phonological rules ever make crucial reference to the categories "onset" and "coda". Thus it appears that the set of syllable structure conditions defining the set of well-formed syllables for each language can be stated with complete adequacy with reference to the categories "syllable" and "peak" (p. 15).

They go on to add that

> ...the distinction between initial consonant clusters and final consonant clusters, which are subject to independent constraints, can be characterized directly with reference to the brackets which delimit the boundaries of the syllable. ...(p. 15)

There is a range of linguistic evidence that one may examine in order to determine whether to adopt a relatively flat syllable structure as proposed by Clements and

Keyser, or a more complex, hierarchical structure (cf., Steriade, 1988). However, we shall limit the scope of our discussion to the phonotactics of English.

In English, for example, there is a tendency towards place *dissimilation* in the syllable onset, but place *assimilation* in the coda (Reilly 1986). For instance, the sequences *bw–, *tl–, *fw–, *θl–, are not allowed syllable initially, but the sequences –mp, –lt, –ŋk, are allowed in the syllable-final positions. These regularities may be stated as *syllable structure conditions* (i.e., as a set of filters that apply over the constituent domains of the syllable). Moreover, in a representation of the syllable that incorporates the categories ONSET and CODA, syllable structure conditions would apply to these constituents directly. On the other hand, if the representation does not posit an internal structure to the syllable, then conditions on the onset and coda apply to the syllable as a whole.

If sufficient empirical evidence can be found, local constraints on sound sequences within the syllable would have importance aside from the problem of providing for a more parsimonious description of the facts of languages such as English. They would help to form the basis of certain predictions regarding the processes of spoken word recognition and lexical access. In Section 5.2, for example, one such set of predictions is made; namely, we discuss a model of spoken word recognition in which lexical decisions are "binded late." Following Nussbaum and Pisoni (1986), we formulate the task as a constraint satisfaction problem in which "structural properties of words interact to specify the identity of an utterance." In adopting this view in the proposed model, we posit constraints on the acoustic properties of words in addition to constraints on phoneme sequences. For both sets of constraints, our suggestion is that the lexicon is the least appropriate for their representation. Alternatively, syllabic constraints are to be represented explicitly, and may be applied early as a means of determining whether a given stretch of the waveform constitutes a *possible* word in English. This is done by first determining whether it consists of a well-formed string of syllables.

5.1.2 Methods

We can assess the need to posit constituents within the syllable for stating phonotactic constraints using the statistical paradigm of *hypothesis testing.* In particular, we may readily formulate a *null* hypothesis of "no privileged groupings [of elements within the syllable]," that may be tested against a substantial amount of data. The test itself is based on a statistic related to mutual information. Prior to developing the mathematical details of this test, it will be helpful to motivate the use of mutual information within the current context.

In general, we have adopted the position that the acoustic properties observed in the sound stream of a language are highly constrained. There are universal constraints on the acoustic realization of an underlying phoneme sequence, as well as language-specific constraints on the phoneme sequence itself. The assimilation/dissimilation facts discussed above are examples of the latter kind of constraint. From a slightly different perspective, we may view such constraints as evidence of the redundancy in a language. Results from information theory, most notably mutual information, provide a natural means of characterizing redundant information, albeit traditionally within the context of statistical communication theory (Gallager, 1968).

The analogy is to view the speaker of a language as an information *source* that encodes a message and the listener as a device that decodes it. Further, the various influences that serve as sources of variability are to be modeled as a *noisy* communication channel. The suggestion, then, is that the phonology of the language provides rules for encoding the message for robust transmission.[1]

[1] We should note that the present information-theoretic analogy forms the basis for the approach to automatic speech recognition adopted by investigators such as Baker (1975), Bahl *et al.* (1983), among others (cf., Levinson (1985) for a review).

5.1.2.1 Mutual Information as a Measure of Collocational Constraint

Let $C_i = \{a_1, a_2, \ldots, a_K\}$ denote the set of phonemes that may occupy constituent i of the syllable. Similarly, let $C_l = \{b_1, b_2, \ldots, b_J\}$ denote the occupants of constituent l, where a_k and b_j are phonemes, and i and l belong to the same syllable. We will use mutual information to estimate the "amount" of information that observing the event $C_i = a_k$ conveys about the event $C_l = b_j$. Put another way, mutual information measures the amount of information that observing the occupant of constituent i of the syllable extracts from the source concerning the occupant of constituent l.

The calculation is defined in (5.1),

$$I(C_i; C_l) = \sum_{a_k \in C_i; b_j \in C_l} P_{C_i;C_l}(a_k; b_j) \log \frac{P_{C_i;C_l}(a_k; b_j)}{P_{C_i}(a_k) P_{C_l}(b_j)}, \tag{5.1}$$

where $P_{C_i;C_l}(a_k; b_j)$ is the probability defined over the space of events corresponding to the pair C_i and C_l. Further, $P_{C_i}(a_k)$ and $P_{C_l}(b_j)$ are the marginal probability functions for the sample spaces corresponding to the constituents C_i and C_l, individually. The logarithm used in this and subsequent calculations are computed in base 2, unless specifically designated.

In our experiment, the above probability functions were determined empirically on the basis of frequency counts. That is, given a sample space of syllables, $L_\sigma = \{\sigma_1, \sigma_2, \ldots \sigma_n\}$, the estimate of the joint probability function, $P_{C_i;C_l}(a_k; b_j)$, was obtained using the expression given in (5.2),

$$K_{C_i;C_l}(a_k; b_j) = \sum_{\sigma_n \in L_\sigma} p_n \mathcal{M}\langle c_i = a_k \wedge c_l = b_j \rangle, \tag{5.2}$$

where p_n is σ_n's frequency of occurrence in some corpus (see below), and $\mathcal{M}$ is a "matching function" defined in (5.3),

$$\mathcal{M}\langle x\rangle = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{otherwise} \end{cases} \tag{5.3}$$

The argument of the matching function is a *boolean variable* (i.e., having values of true or false); the function's output is an integer. In its use in (5.2), $\mathcal{M}\langle x\rangle$ returns the value 1 in the instance of a designated co-occurrence, otherwise it returns 0. These output values are then weighted and summed to produce $K_{C_i;C_l}(a_k;b_j)$. Similarly the counts, $K_{C_i}(a_k)$ and $K_{C_l}(b_j)$ are obtained as estimates for the marginal probability functions $P_{C_i}(a_k)$ and $P_{C_l}(b_j)$ respectively. The count functions are all normalized by scale by the factor,

$$\frac{1}{\sum_n p_n},$$

to ensure that their respective probability estimates sum to 1.0.

The frequency counts, p_n, may be obtained under two conditions: the frequency weighted condition is where the syllable's frequency of occurrence is determined from some corpus of naturally occurring text. For the unweighted condition, the value of p_n is set to 1.

The Likelihood Ratio Statistic

Since our desire is to test the *significance* of groupings of constituents within the syllable, a statistic is required that has some known underlying distribution. Mutual information does not satisfy this criterion. But a closely related statistic, namely the *likelihood ratio* (or G^2) statistic, does. This statistic is often used in the analysis of categorical data (Bishop, Feinberg, and Holland,1975). Its relation to mutual information may be seen from comparing its definition in (5.4) with the expression for mutual information in (5.1):

$$G^2 = 2.0 \times \sum_{a_k \in C_i; b_j \in C_l} P_{C_i;C_l}(a_k;b_j) \log_e \frac{P_{C_i;C_l}(a_k;b_j)}{P_{C_i}(a_k)P_{C_l}(b_j)}. \tag{5.4}$$

Specifically, one may compare these two expressions and observe that the two statistics are related by the scale factor. That is,

$$G^2 = 2.0 \times \log_2 e \times I, \tag{5.5}$$

where in Equation (5.5), the term $\log_2 e$ accounts for the differences in the base used for computing the logarithm.

The statistic G^2 is distributed as a χ^2 random variable having *degrees of freedom* equal to $(K-1)(J-1)$, where $K = |C_i|$, $J = |C_j|$, and $|\cdot|$ denotes the *cardinality* or size of a set.

5.1.2.2 Data Preparation

The syllable lexicon used in the current study was derived from the 20,000-word *Merriam-Webster Pocket Dictionary* stored in computer-readable form (see Appendix A). The following is a description of its preparation. For each word in the dictionary, a phonemic pronunciation (typically the word's most common) is stored along with its frequency of occurrence in the 1 million word *Brown Corpus* (Kucera and Francis, 1967). Words not occurring in the *Brown Corpus* (of which, there were approximately 8500) were arbitrarily assigned a frequency of occurrence value of one. The pronunciation was also marked for lexical stress. Each of the words contained in this dictionary was parsed into syllables. The algorithm used for deciding the locations of syllable boundaries is based on the syllable grammar outlined in Chapter 3. That is, word-internal syllable boundaries were determined largely on the basis of phonotactic constraints on syllable-initial and -final phoneme sequences. The *maximum onset* and *stress resyllabification principles* were used to address cases where placement of the syllable boundary could not be determined unambiguously on the basis of phonotactics alone. Ambisyllabic phonemes in a falling stress environment were assigned to the syllable's coda. The remaining cases of ambisyllabic phonemes were assigned to the syllable onset. Next, the list of syllables from the syllabified dictionary were collected into a "lexicon" of their own, netting a total of 5572 entries. Finally, using the syllable structure assignment principles specified by the grammar given in Chapter 3, we then associated phonemes with terminal slots of the syllable template.

From the standpoint of the current experiment, the syllable constituent labels associated with phonemes (i.e., Inner-onset, Nucleus, Outer-coda) are arbitrary. In other words, since the purpose is to use a statistical procedure to decide whether the Inner-onset and Outer-onset, for example, are to be dominated by the Onset (or perhaps some other constituent), using the above labels is prejudicial. Therefore, the terminal categories of the grammar were assigned arbitrary labels, C_i, where the correspondence between this new set of labels and the former set is given in the table shown in (R 5.1):

	Label	*Terminal Category*
	C_1	Onset-s-slot
	C_2	Outer-onset
R 5.1	C_3	Inner-onset
	C_4	Nucleus
	C_5	Inner-coda
	C_6	Coda-s-slot
	C_7	Outer-coda

The reader will note that the categories Onset-s-slot and Coda-s-slot have been created. In the current experiment, /sp/, /st/, /sk/ clusters were not considered to be single segments as before, but instead, were assigned to their own slots. In addition, although experiments which considered syllable-affix positions were conducted, for the purpose of brevity, their results will not be fully described. Briefly, these constituents showed no special proclivity for associating with any of the other syllable constituents, with the exception of a slight association of the Affix-1 with the second coda position.

5.1.2.3 Limitations of Method

Before proceeding to a discussion of the results of the current study, two comments concerning the use of the above methods and data collection procedures are to be kept in mind. First, a statistical characterization of a language, similar to that which is provided through an information-theoretic approach, entails a number of potential theoretical disadvantages. Collecting statistics from a finite corpus inherently fails to account for the "creative aspects of natural languages": the capability for languages to express indefinitely many thoughts on the basis of a finite inventory of symbols (Chomsky 1965). Moreover, it is unclear as to whether a probabilistic language model determined from a corpus, whatever its size, will account for the linguistic competence or performance of a native speaker/hearer of a language.

The second comment concerns the manner in which phonemes are associated with the terminal categories of the syllable template. Recall from the discussion of Chapter 3 that, at the surface-phonemic level, the terminal elements of the syllable template are analogous to "slots", i.e., placeholders that are filled with phonemes. In the assignment of syllable structure to a phoneme string, not all slots will be filled. For example, for simple syllables consisting of only vowels, all slots but the NUCLEUS are empty. In the monosyllable, *cat*, the OUTER-ONSET, NUCLEUS, and OUTER-CODA are filled, but the other slots remain empty.

There are two consequences of this "slot-and-filler" approach which bear on the results of the current experiment. First, we created the artificial "phoneme", ∅, to fill empty slots. Co-occurrence counts using this "phoneme" have had a definite effect on the mutual information statistics which are calculated. The second consequence of using slots and fillers is that the statistics gathered do not distinguish collocational constraints that are due to the sonority sequencing principle from those that are attributed to constraints on place of articulation. The first of these two consequences has had no adverse effect on our ability to draw conclusions from this experiment. The fact that certain syllable positions tend not to be filled when other syllable positions

	C_1	C_2	C_3	C_4	C_5	C_6	C_7
C_1		.1557 .0375	.0214 .0052	.0039 .0070	.0018 .0024	.0021 .0000	.0053 .0039
C_2			.4863 .2968	.0619 .4360	.0128 .1001	.0053 .0061	.0818 .1474
C_3				.0929 .1879	.0336 .0421	.0012 .0055	.0316 .1052
C_4					.1921 .2893	.0092 .0123	.1929 .3478
C_5						.0168 .0027	.2755 .1163
C_6							.0662 .0380

Table 5.1: Table of mutual information statistics calculated from the syllable lexicon. Each cell contains two numbers. The top number is the mutual information value calculated for the frequency unweighted case. The bottom value has been calculated using frequency of occurrence statistics based on the *Brown Corpus.*

contain various phonemes ought to be reflected in the data. The failure to clearly distinguish the influence on co-occurence statistics that are due to manner versus place represents a greater problem. The statistics will be biased, reflecting to a certain extent, a syllable structure that has been already determined. For instance, the slot C_2 is pre-determined to be obstruent, while C_3 is constrained to be a sonorant. These pre-determined constraints contribute to the magnitude of co-occurence statistics that are calculated, although no attempt has been made to determine their extent.

5.1.3 Results

The numerical results of the experiment are presented in Table 5.1. The data in the table are displayed in the form of an *upper-diagonal* matrix, where each cell contains a mutual information value computed for the constituent pair corresponding the cell's row and column. The lower diagonal terms are not shown since they may be ascertained on the basis of the symmetry properties of mutual information (see Appendix B). Each cell in the table contains two numbers: the top number is mutual information calculated from the frequency unweighted lexicon and the bottom number takes frequency of occurrence statistics from the *Brown Corpus* into consideration.

If mutual information is used as a measure of the "strength" of collocational constraints, then it appears that the collocational constraints among the constituent pair (C_2, C_3) is strongest; the mutual information value calculated for this pair is .486. This pair corresponds to the INNER-ONSET and OUTER-ONSET positions. Next, in the ranking of constituent pairs according to mutual information, is the pair (C_5, C_7) having a value of .276. These positions correspond to the INNER-CODA and OUTER-CODA constituents. Following the pair corresponding to the two coda positions are the pairs (C_4, C_7) (.193) and (C_4, C_5) (.192). These two pairs correspond to the NUCLEUS and OUTER-CODA, and the NUCLEUS and INNER-CODA respectively.

For the frequency weighted data, collocational restrictions are the strongest for the constituent pair (C_2, C_4), having a mutual information value of .436. This pair corresponds to the OUTER-ONSET and the NUCLEUS. The next pair in order of mutual information values is (C_4, C_7) with a value of .348. This pair corresponds to the NUCLEUS and OUTER-CODA.

Hierarchical Cluster Analysis

The data presented in Table 5.1 were also subjected to hierarchical cluster analysis (cf., Hand, 1981; Gnanadesikan, 1977). By applying an agglomerative or "bottom-up" clustering procedure to the data given in Table 5.1, our intent was to determine whether the data given in the table indicated a syllable structure over and above pairs of terminal constituents. The method used the *single-linkage* (or so-called "nearest-neighbor") principle for computing distances among clusters.[2]

Figure 5.1 displays the results of the cluster analysis in the form of *dendrograms,* where in general, a dendrogram is a graph or "tree-like" structure that displays the

[2] The details of hierarchical clustering procedures will not be given here. Instead, the reader is referred to the references cited. We should point out, however, that several principles may be used for computing distances between clusters containing more than one data point, including the *single-*, the *complete-*, and the *average-*linkage principles. The cited references provide an explanation of each. We tried each, and found the results to be more or less the same.

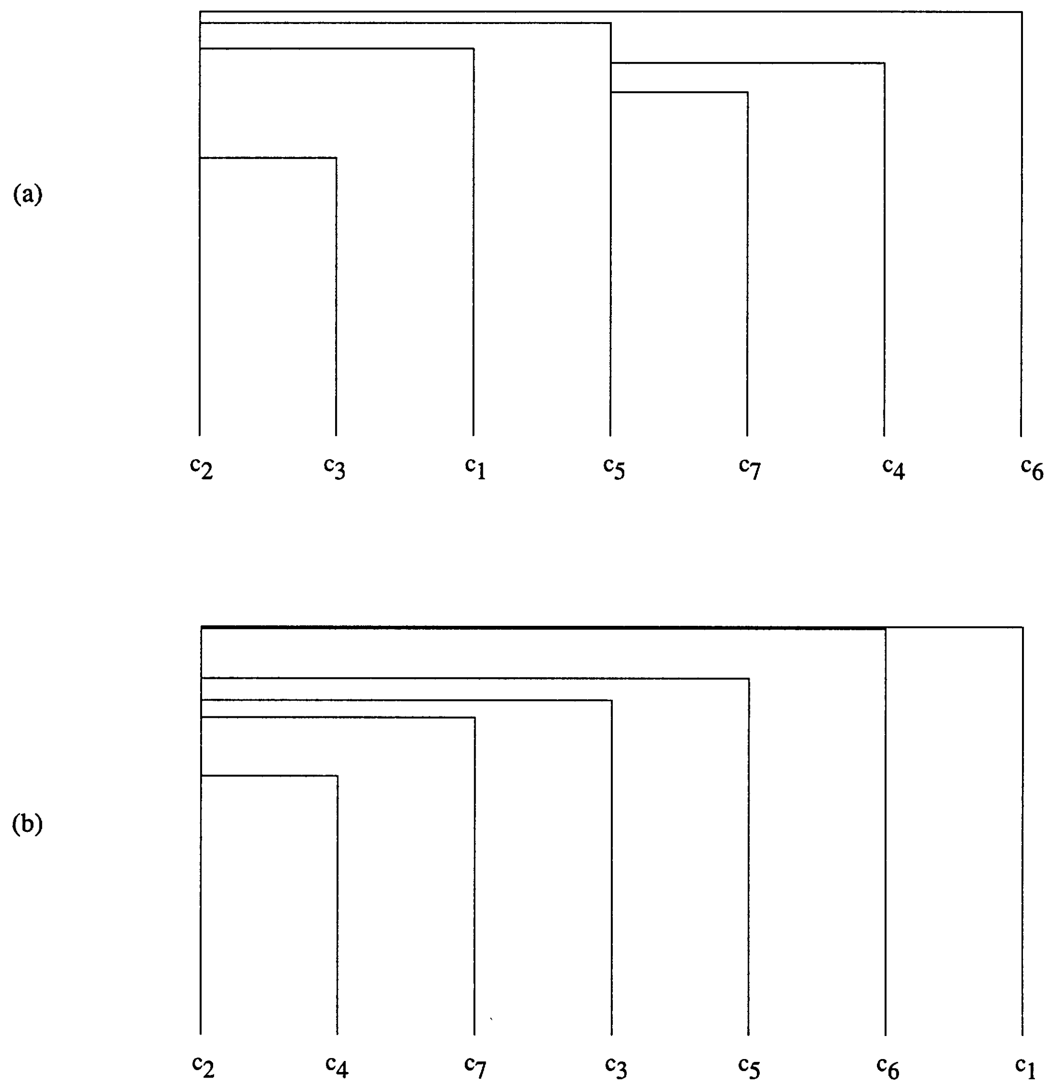


Figure 5.1: Dendrograms computed for the data presented in Table 5.1. The dendrogram shown in (a) is for the frequency-unweighted case. The dendrogram shown in (b) is for frequency-weighted data. In both, mutual information is plotted along the vertical axis.

"strength of clustering". That is, every data point appears with itself in a cluster; different data points that are joined to form a cluster are done so at some level of closeness or proximity. In this case mutual information is the proximity measure. The level of proximity is represented by the height at which the lines in the graph corresponding to the individual clusters are connected. In the dendrograms shown in Figure 5.1, we see pairs of constituents having *higher* mutual information values being joined together at *lower* positions in the dendrogram.[3] This reflects the fact that these pairs of constituents are relatively close.

5.1.4 Discussion and Conclusions

None of the mutual information values presented in Table 5.1 are significant according to the their corresponding G^2 statistics ($p < .05$). Therefore, it is difficult to read a great deal into these data. However, some general comments do apply. For example, depending on whether the probability functions are estimated with frequency weighted data, the data presented in Table 5.1 present two different pictures regarding a syllable hierarchy. This may be seen by comparing the dendrograms in Figure 5.1. From the appearance of the dendrogram for the unweighted lexicon (Figure 5.1(a)), the syllable topology adopted in Chapter 3 seems justified. That is, the syllable's CORE is justifiably represented as having the ONSET and RHYME as its immediate constituents. The ONSET immediately dominates three syllable initial positions, and the RHYME dominates a NUCLEUS and CODA. In contrast, the dendrogram computed for the frequency-weighted lexicon provides evidence for a syllable structure in which the first onset position of the syllable and NUCLEUS are to be grouped together as a constituent. The remainder of the constituent structure suggested by this data would be "flatter" due to an apparent lack of clusters which join the remaining terminal elements.

In conclusion, on the basis of these data, the null hypothesis (no privileged group-

[3] Plotted along the vertical axis of each of the dendrograms is the value, e^{-I}, where I denotes a mutual information value.

ings) is not to be rejected. However, the data corresponding to the unweighted syllable lexicon does have a structure that may be interpreted in terms of the syllable template suggested in Chapter 3. Furthermore, we may only speculate what the reason may be for the apparent disparity between syllable structures found for the frequency-weighted versus frequency-unweighted data. One possible reason may be a bias in the *Brown Corpus* towards the use of relatively simple syllables (i.e., structures of the type *CVC* and simpler). For example, we found that a large number of the more common words (having a high frequency of occurrence) to be function words (e.g., articles and prepositions such as *the* and *at*). Most of these words are monosyllables and are comprised of simple phonological structures. These tokens have an undue influence over the mutual information statistics that are calculated.

The question then is, which set of results do we use as a basis for a model of word recognition. In the psychological literature, language processing models have often assumed frequency of occurrence statistics to be important (Pisoni and Luce, 1987). From this perspective, the results obtained from the frequency-weighted lexicon are perhaps more valuable. In addition, either the notion of the syllable having an immediate constituent structure would have to be dismissed, or perhaps a different (maybe simpler) syllable structure proposed for the use in a processing model. One may also add to this strategy some mechanism for treating function words (or other high-frequency words of relatively simple phonological structure) differently (cf., Bradley, 1978). From a linguistic perspective, the use of frequency of occurrence data is not as crucial. Given that the classical position is that a distinction is to be made between *performance* and *competence* (Chomsky, 1965), the unweighted data would be seen as equally important, since frequency of occurrence statistics speak more to performance issues. In this view, a more enriched constituent structure is once again justified. Finally, from a computational point of view, it has already been suggested in Chapter 2 that local constraints (as reflected in an immediate constituent grammar) have the advantage of simplifying the statement of rules, and, as will be suggested below, may save effort during word recognition.

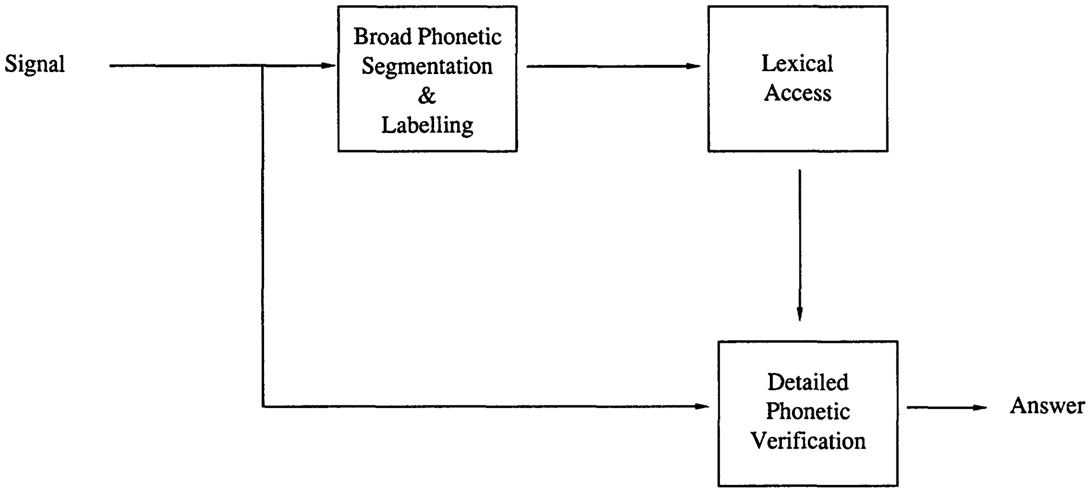

Signal
Broad Phonetic Segmentation & Labelling
Lexical Access
Detailed Phonetic Verification
Answer

Figure 5.2: Model of spoken word recognition proposed by Zue.

5.2 Lexical Constraints

In this section, attention is focused on the interaction between syllabic constraints and constraints imposed by the lexicon. The problem of determining the nature of this interaction is approached from the perspective of specifying a lexical retrieval component of an automatic speech recognition system. As mentioned above, a number of investigators have suggested this task be formulated as a constraint satisfaction problem (see, for instance, Zue, 1983; Shipman and Zue, 1982; Huttenlocher and Zue, 1984; Huttenlocher and Withgott, 1986; and Zue, 1986, for further arguments in support of this view). The following assumptions underlie this methodological approach. First, the constraint satisfaction paradigm acknowledges that an utterance will generally contain only partial information regarding its phonetic identity. However, there is the notion that the relationship between a word's lexical representation and its acoustic properties is highly constrained and consists of redundant information. As a result, producing a full phonological specification of a word prior to accessing the lexicon is not a requirement in all cases.

A well-known recognition framework to evolve from these considerations is the model shown in Figure 5.2. The recognition framework depicted in this figure has been proposed by Zue and his colleagues (Zue, 1986).[4] In this strategy, the speech signal is segmented and labelled in terms of broad phonetic classes (e.g., *Vowel, Semi-vowel, Nasal, Strong Fricative, Weak Fricative,* and *Stop*), and the resulting transcription is then used to probe the lexicon. The final stage of the processing is a detailed phonetic verification step that decides among the cohort of word candidates resulting from the lexical probe.

A number advantages are cited in favor of this strategy over an approach that involves segmenting and labelling the signal in terms of phonemic or allophonic segments. First, rather than having a phonetic classifier decide among the 60 to 100 phonetic classes that would be required in the alternative approach, the initial phonetic analysis would only have to decide among a small number of phonetic categories, each presumably having robust acoustic correlates. Second, an exhaustive lexical search is avoided. By partitioning entries of the lexicon according to their broad phonetic spellings, a relatively small percentage of the lexicon actually needs to be searched during the final stages of recognition. Finally, although narrow phonetic labels must eventually be decided, this classification is made during the process of detailed phonetic verification, a point at which partial phonetic context has already been established.

In the constraint satisfaction approach, the general emphasis is shifted away from developing strategies for coping with variability in the phonetic realizations of phonemes, and towards identifying acoustic patterns that tend to be "stable and reliable" (Huttenlocher and Withgott, 1986). In research based on this paradigm, stable acoustic patterns are given phonological specifications that are systematically evaluated with respect to the degree to which they may be uniquely associated with

[4] We should point out that the model shown in Figure 5.2 differs in key ways from Zue's current thinking. For example, the phonetic segmentation an labelling operation is to be replaced by a strictly acoustic segmentation using on a input signal representation based on characteristics of the human auditory system.

Condition	*Avg.*	Max.	% Unique
Unweighted	21	223	32%
Freq. Weighted	34	223	6%

Table 5.2: Equivalence class sizes obtained by mapping the words in the *Merriam-Webster Pocket Dictionary* into manner-of-articulation class sequences (After Shipman and Zue 1982).

lexical candidates.

The view taken here is that the pursuit of the above constraint-based methodology is to be supported on all accounts. However, studies have revealed a number of limitations associated with basing a speech recognition framework on a strategy that relies entirely on the lexicon as a source of constraint. Experiments involving large lexicons, for example, indicate that lexical constraints are indeed powerful for the problem of isolated word recognition. For example, Table 5.2 summarizes results of a study done by Shipman and Zue (1982), in which the entries of the 20,000 word *Merriam-Webster Pocket Dictionary* were transcribed according to manner-of-articulation features. The figures given in the table are statistics on the resulting lexical partitions, analyzed under both frequency-weighted and unweighted conditions. In the frequency-unweighted case, the average size of an equivalence class (i.e., the set of words having the same manner-of-articulation transcription) numbered around 21. For the weighted case, the average equivalence class size grew to 34. The maximum equivalence class was 223 (approximately 1% of the lexicon). The statistics in the table also show that a large number of words have unique broad phonetic transcriptions.

For continuous speech recognition, however, it is suggested that the lexicon be the place of last resort in the search for constraint, particularly if it contains a large number of entries. Harrington and Johnstone (1987) performed experiments with a lexicon consisting of 4,000 of the most common words in the *American Heritage Dictionary.* Using a *mid-class* transcription (i.e., in terms of phonetic categories slightly more detailed than manner-of-articulation classes), it was found that the typical size of a word lattice resulting from a mid-class phonetic analysis of the signal would easily number into the millions. Even when provided with a completely specified phonemic transcription, this study showed that lattice sizes in the 1,000-word range could be expected.

The numbers that Harrington and Johnstone report are indicative of the number of times that the lexicon would have to be probed in order to find an entry that corresponds to a given stretch of the waveform. What these numbers do not indicate is the number of unsuccessful tries (i.e., the number of times that the lexical matching procedure must backtrack). Therefore, it is suggested that the problem in reality is even worse. For example, in most cases, lexicons are represented as discrimination networks (or trees) (cf., Smith, 1976). These are finite state machines that process an utterance from left-to-right, while matching elements of the input phonetic transcription against phonetic symbols represented on the arcs of the network. This being the case, one could envision times at which a lexical matcher would have to proceed to the end of a phoneme sequence corresponding to a potential lexical candidate before it could determine that the sequence did not correspond to a possible word in English.

This worse case scenario would be predicted in the case of the word candidate *stramk*. That is, a hypothetical lexical matching procedure relying on a fully specified phonetic specification as its input would determine this word to be invalid only after it has detected the failure of the word-final sequence, /mk/, to satisfy the nasal-stop homogamic rule.

The question, then, is are there alternative strategies to probing the word lexicon in such a manner. Nussbaum and Pisoni (1986) suggest that the lexical partitioning results reported by Zue and his colleagues are tacitly taking into consideration two factors that constrain lexical search: the length of phonemic sequences as well as their broad phonotactic shape. The length of a phonological sequence will not be known a priori in continuous speech, and therefore this aspect of lexical constraint is not useful. Constraints on broad phonotactic shape may be stated in terms of well-formedness conditions on phonological syllables. Thus a possible strategy might be to use constraints on well-formed syllables as a "filter". Lexical candidates that are "passed" by the filter are then tested against the lexicon itself. In such a strategy lexical entries may be represented as a strings of phonemes, or alternatively, as strings of syllables. This latter alternative provides the context of the current study.

5.2.1 Delayed Binding of Lexical Decisions

One possible lexical access strategy supported by arguments given thus far may be concisely stated in terms of the constraints represented in (R 5.2):

R 5.2

$$w \longrightarrow \sigma^{+}$$

such that

$$\sigma \in L_{\sigma}$$

and

$$w \in L_{W_{\sigma}}.$$

In (R 5.2), L_{σ} is a "lexicon" of syllables, prepared, perhaps, in the manner outlined in Section 5.1 above. The term $L_{W_{\sigma}}$ denotes a word lexicon in which each word is represented as a concatenation of actually occurring syllables. The constraints stated in (R 5.2) represent a lexical access procedure whereby, it is first determined whether a phoneme string corresponds to a *possible* word (i.e., by satisfying constraints on well-formed syllables). If so, it then is determined whether the string of phonemes comprises a concatenation of actually occurring syllables (by looking up each in L_{σ}). Finally, the word lexicon ($L_{W_{\sigma}}$) is searched.

There have been two investigators to propose lexical access procedures based on this idea. Smith (1976) formulates a strategy in which words are represented in terms of a hierarchy of levels: (1) segments, (2) "sylparts", and (3) syllables. In his framework, the relationship between levels is implemented in the form of a discrimination tree. Church (1983) adopts a similar scheme, but replaces the mapping from "syl-parts" to "syllables" ((2) to (3)) with a context-free parser. The current proposal is to extend these two strategies by first positing the satisfaction of realization constraints, such as those enumerated in Chapter 3, as the means by which phonological syllables are identified in the sound stream. A second aspect of this proposal is to structure the lexical access problem such that a complete phonemic specification of each syllable is not required.

The feasibility of the above strategy was experimentally tested. For the remainder of the present section, the experiment is described and its results presented.

5.2.2 The Syllable Lexicon as a Source of Constraint: Results of Lexical Partitioning Experiments

In the current set of experiments, partitioning experiments have been carried out on the syllable lexicon described in Section 5.1. Specifically, various transcriptions have been applied to individual syllables comprising this lexicon, and statistics of the resulting equivalence classes (i.e., lexical partitions) have been calculated. Underlying this investigation is the assumption that a complete specification of a phonological syllable derived from the waveform is unlikely. In the event that certain parts of the syllable may be completely specified while others only partially so, the study has sought to determine the parts of the syllable that contribute the most towards identifying an eventual lexical candidate.

This determination has been made by transcribing syllables in a manner such that certain parts (e.g., the ONSET, CODA, etc.) were specified according to manner, place, and voicing, while other parts were partially specified (e.g., manner, but not place or voicing) or completely unspecified (i.e., no feature information provided at all). The resulting lexical partitions corresponding to each of possible experimental conditions were then compared.

5.2.2.1 Methods of Comparison

In a review of previous studies, two principal measures of lexical constraint may be discerned. The first is a set of cohort statistics, including: the *expected cohort size*, the *maximum cohort size*, and some measure indicating the number of cohorts containing a single lexical entry (Shipman and Zue, 1982). The second type of measure is based on the notion of mutual information, although this time mutual information is a measure of the amount of information extracted from a speaker concerning the identity of a lexical entry once partial phonetic information concerning an utterance has been supplied (Carter, 1987). In order to define the use of mutual information in the current context, some notation will need to be developed.

Let $\mathcal{T}$ denote a particular *type* of transcription. For example, we may choose to specify manner, place, and voicing features in the ONSET while leaving the rest of the syllable completely unspecified. Following Carter, the lexical partition that results from applying $\mathcal{T}$ will be denoted $\Pi(\mathcal{T})$. As alluded to above, intuitively, applying a transcription to an utterance may be viewed as "extracting information" concerning its lexical identity. As a result, the uncertainty concerning its identity is reduced. To measure this reduction in uncertainty, Carter defines *Percent Information Extracted* (PIE) as (5.6):

$$\text{PIE}(L_\sigma; \Pi(\mathcal{T})) = \frac{H(L_\sigma) - H(L_\sigma|\Pi(\mathcal{T}))}{H(L_\sigma)} \times 100\% \tag{5.6}$$

where

$$H(L_\sigma) = - \sum_{\sigma_i \in L_\sigma} p_i \log p_i. \tag{5.7}$$

and

$$H(L_\sigma|\Pi(\mathcal{T})) = H(L_\alpha) - \sum_{\alpha \in \Pi(\mathcal{T})} P_\alpha \log P_\alpha. \tag{5.8}$$

These quantities are described in detail in Appendix B. Briefly, the quantity $H(L_\sigma)$ in (5.6) is the *a priori* average information extracted from the source by observing syllable, σ_n. The quantity $H(L_\sigma|\Pi(\mathcal{T}))$ is defined as the amount information extracted by applying the transcription $\mathcal{T}$. The reader will note that the numerator may be considered the mutual information between the ensembles L_σ and $\Pi(\mathcal{T})$.

5.2.2.2 Experimental Design and Preparation of Data

Starting with the syllable lexicon collected for the study of collocational restrictions described in Section 5.1, we marked entries using a variety of transcriptions. The choice of transcriptions applied were based on a two-way factorial design, in which one of the factors was the level of specificity in the transcription, while the second was the subset of the syllable's terminal categories that the transcription was applied to. The levels of specificity are defined as follows:

R 5.3

$$\textit{Level of Specificity} = \left\{ \begin{array}{c} \text{Voicing} \\ \text{Manner} \\ \text{Place} \\ \text{Voicing} + \text{Place} \\ \text{Voicing} + \text{Manner} \\ \text{Manner} + \text{Place} \\ \text{Manner+Place+Voicing} \end{array} \right\}$$

The subsets of constituents that the above specification was applied to is denoted by the following expression:

R 5.4

$$\textit{Subset of Constituents} = \left\{ \begin{array}{c} \textsc{Onset} \\ \textsc{Coda} \\ \textsc{Nucleus} \\ \textsc{Onset} + \textsc{Nucleus} \\ \textsc{Onset} + \textsc{Coda} \\ \textsc{Nucleus} + \textsc{Coda} \\ \textsc{Onset}+\ \textsc{Nucleus} + \textsc{Coda} \end{array} \right\}$$

In (R 5.4) the last two subsets of constituents correspond to the Rhyme and Core respectively.

The two sets of factors represented in (R 5.3) and (R 5.4) represent varying experimental conditions whereby information is added to the syllable until it is fully specified.

As a means of reducing the number of experimental conditions, binary features for specifying manner and place were abandoned in favor of marking data with multivalued features. For consonants, the various values of manner, place, and voicing features that may be associated with a terminal position of the syllable template are enumerated in (R 5.5):

R 5.5

Manner	*Place*	*Voicing*
Stop	*Labial*	*Voiced*
Fricative	*Dental*	*Unvoiced*
Affricate	*Alveolar*	∅
Fricative-Stop	*Alveolar-Labial*	
Syllabic	*Alveolar-Alveolar*	
Nasal	*Alveolar-Velar*	
Semi-Vowel	*Palatal*	
∅	*Velar*	
	Lateral	
	Retroflexed	
	Glottis	
	∅	

For vowels (belonging to the nucleus), there is only one manner of articulation. Further, we assume vowels to be voiced. Therefore, the nucleus need only be given a value for place, in which case, the entire set of vowel features is specified.

The following two aspects of the syllabic representation for marking lexical entries should be noted. They are similar to remarks made in Section 5.1.2.3. First, for the most part, manner features are factored into the syllable's hierarchical representation. That is, the well-formedness conditions of the syllabic template prescribe a certain assignment of manner-of-articulation categories to the terminal positions of the template. For example, the INNER-ONSET and INNER-CODA positions will not distinguish manner, and neither will the NUCLEUS. In the results, a consequence of this redundancy will be that specifying manner-of-articulation features does not have a large effect on improving the quality of a lexical partition. Secondly, empty slots are assigned the artificial phonetic category $\emptyset$.

5.2.2.3 Results

Results were obtained using the frequency-weighted lexicon, as well as by assuming uniform frequency of occurrence. For the most part, values will be displayed in the form of a graph in which the PIE value is plotted along the vertical axis and the level of specificity is plotted along the horizontal axis. Curves designating the various subsets of the syllable will be overlaid on top of one another for comparison.

The Onset vs. the Coda

Figure 5.3 shows a comparison of the ONSET vs. the CODA on the basis of PIE for varying levels of specificity. Figure 5.3(a) is for the unweighted case, while Figure 5.3(b) is for the weighted case. Both sets of data show the same trends, although the separation among the curves is larger for the weighted condition than for the unweighted. In general, as one adds more features to the transcription, the resulting lexical partitioning improves. However, there is one exception that presents itself in the data as a "dip" at the point at which place features are replaced in the transcription by voicing and manner features.

The dips are indicative of a situation in which a transcription fails to add new information concerning the identity of syllable as one moves from one level of specificity in the transcription of lexical entries to another. In our data this happens as one moves from a transcription in which place features are specified to one in which only manner and voicing features are specified. This is to be expected, given that man-

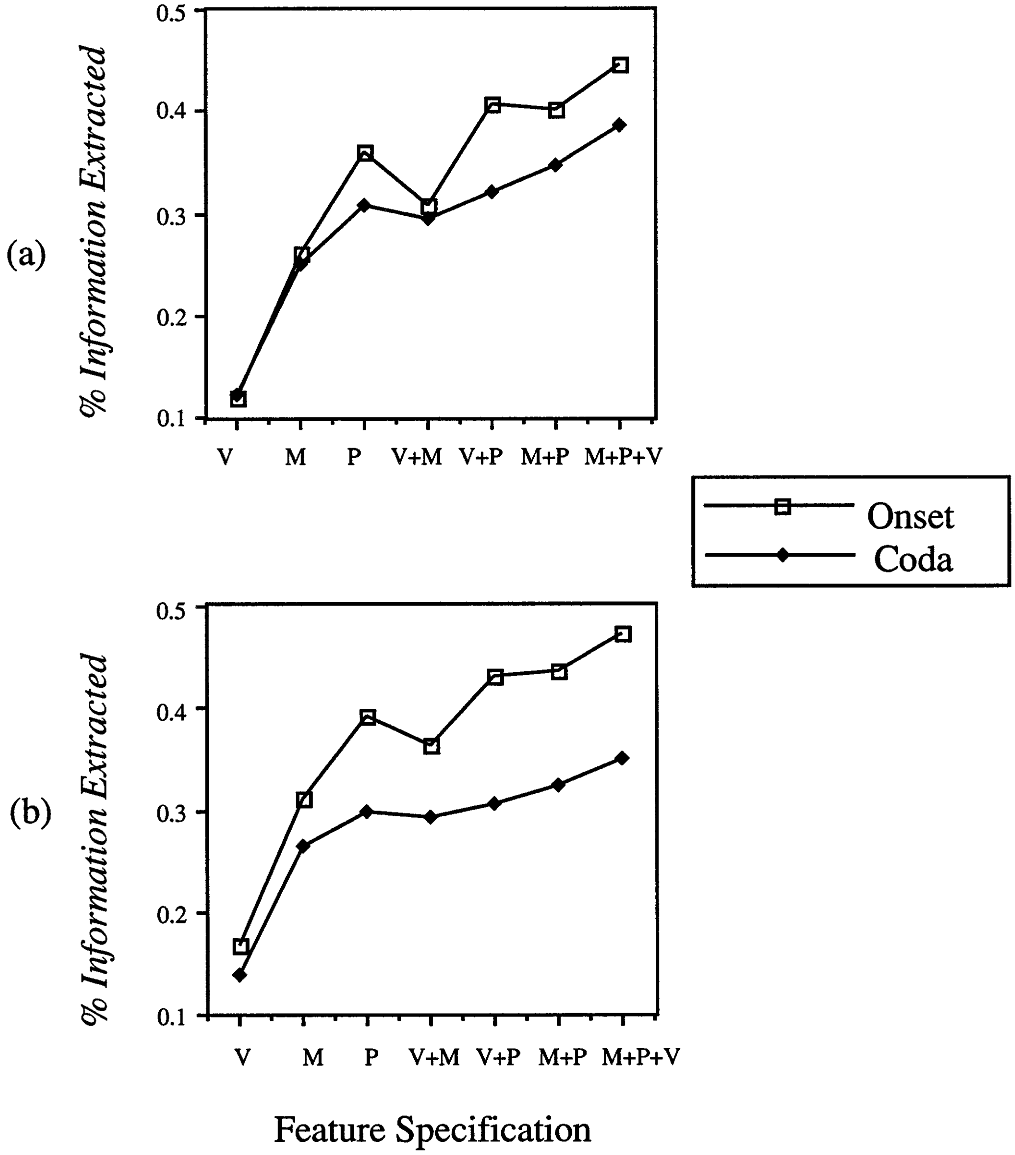


Figure 5.3: Percent information extracted comparing the ONSET versus the CODA for the various levels of specificity in the lexical representation: (a) shows results computed from the frequency-unweighted lexicon, (b) is for the frequency-weighted lexicon. The initials M, P, and V used to label the horizontal scale denote Manner, Place, and Voicing features respectively.

ner and voicing are, for the most part, already factored in the syllable's hierarchical structure (see discussion above).

In both the unweighted (Figure 5.3(a)) and the weighted cases (Figure 5.3(b)), the onset tends to be more informative than the coda. This discrepancy increases as more features are added to the specification.

Constituents taken in Pairs

Figures 5.4 (a) and (b) show comparisons of constituents ONSET, NUCLEUS, and CODA taken two at a time. Trends similar to those found in the data presented in Figure 5.3 once again may be observed. As one increases the level of specificity, the more effective the lexical partition becomes.

One also observes a "dip" in the curves as one moves from the condition in which syllables are transcribed according to place of articulation and to the condition in which only manner and voicing is used. In the current case, the dip observed for NUCLEUS+CODA and ONSET+NUCLEUS conditions appear to be more pronounced than those found in Figure 5.3. Once again, the dip is to be expected, and the explanation for its presence is the same as the one given above. Its magnitude in the current case is due to the large increase in information extracted once place features are specified for the syllable nucleus. Recall that specifying manner and voicing features for the nucleus does not provide any information since this constituent of the syllable is constrained to be voiced and belongs to only one manner-of-articulation class.

As was the case before, specifying the beginning of the syllable has a more positive effect on the lexical partition than specifying its end. In both Figures 5.4 (a) and (b), specifying the combination NUCLEUS+CODA provides the least amount of information, while in the case of statistics computed from the unweighted-lexicon, specifying the combination ONSET+CODA provides the most. In the case of the weighted lexicon, specifying the combination ONSET+PEAK provides the most amount of information.

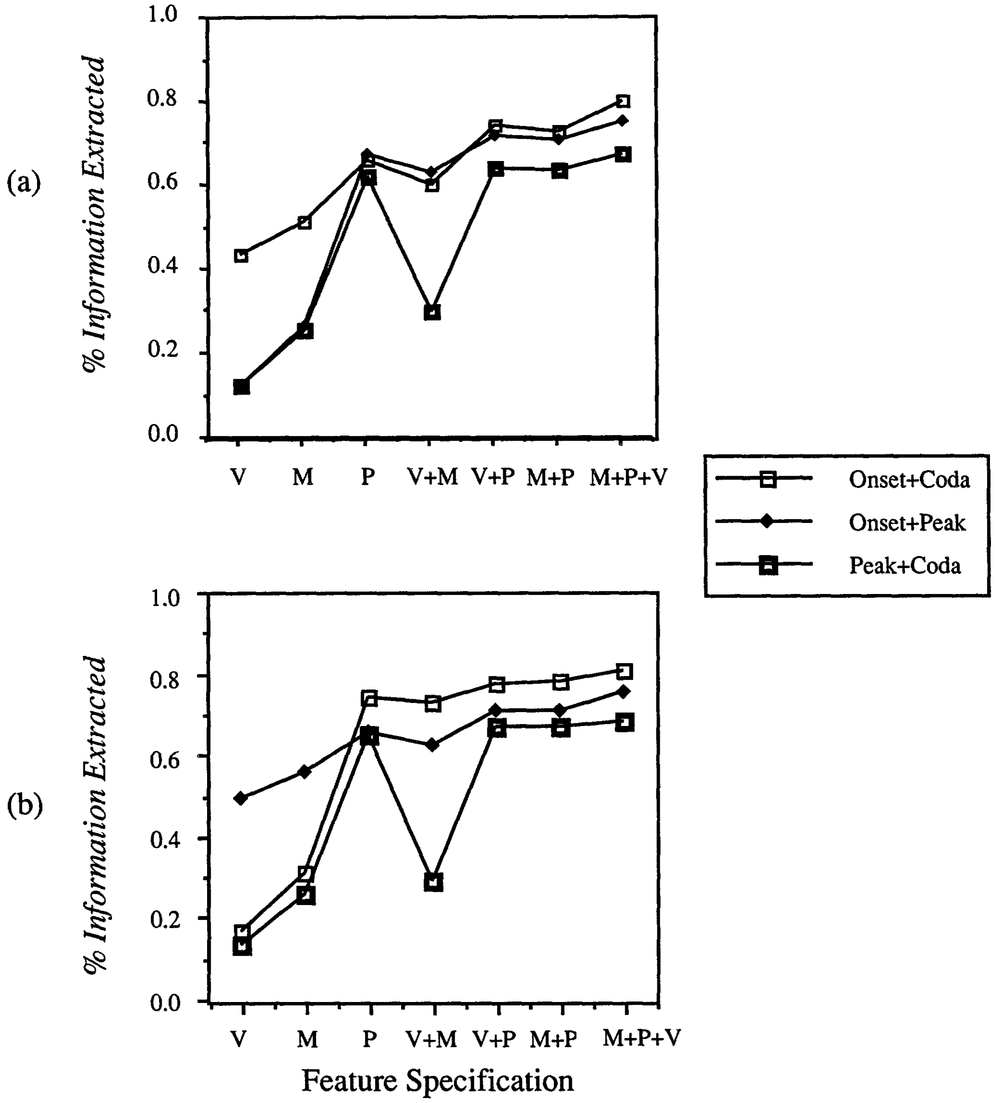


Figure 5.4: Percent information extracted comparing various groupings of the the ONSET, NUCLEUS, and CODA for the various levels of specificity in the lexical representation: (a) shows results computed from the frequency-unweighted lexicon, (b) shows results for the frequency-weighted lexicon. The initials M, P, and V used to label the horizontal scale denote Manner, Place, and Voicing features respectively.

At present, we may only speculate why specifying the combination of the ONSET+NUCLEUS has a more positive effect on the partitions of the frequency-weighted lexicon. Once again, the reasons are similar to those offered in explaining the disparity in collocational constraints measured in the experiment described in Section 5.1. That is, the more frequent words observed in the *Brown Corpus* have simpler syllable structures. As a result, the statistics calculated are slightly biased in their favor.

Calculations Involving the Entire Syllable CORE

Finally, in Figures 5.5 (a) and (b) PIE statistics are plotted for various feature specifications for the syllable CORE (i.e., ONSET+NUCLEUS+CODA). We include these results for the sake of completeness. As has been the case for the previous two sets of data, specifying place, as opposed to manner or voicing, provides the most constraint. Once again, this is a consequence of the fact that the syllable's hierarchical representation encodes constraints on sequences of manner-of-articulation categories implicitly.

5.2.3 Discussion: Implications for a Model of Lexical Access

Earlier in this section, we discussed a model of lexical retrieval in which "lexical decisions are binded late." In the context of this thesis, this meant that constraints on syllables are to be satisfied prior to searching the lexicon for words to match against an utterance's acoustic description. Part of this procedure is to determine whether the input is comprised of a sequence of actually occurring syllables by retrieving entries from a syllable lexicon. The data reported in this section have direct bearing on the computational feasibility of the retrieval task.

The general assumption is that only certain regions of an utterance will receive a full feature specification prior to accessing either the syllable or the word lexicon. Furthermore, these regions that receive full specification may be chosen because they

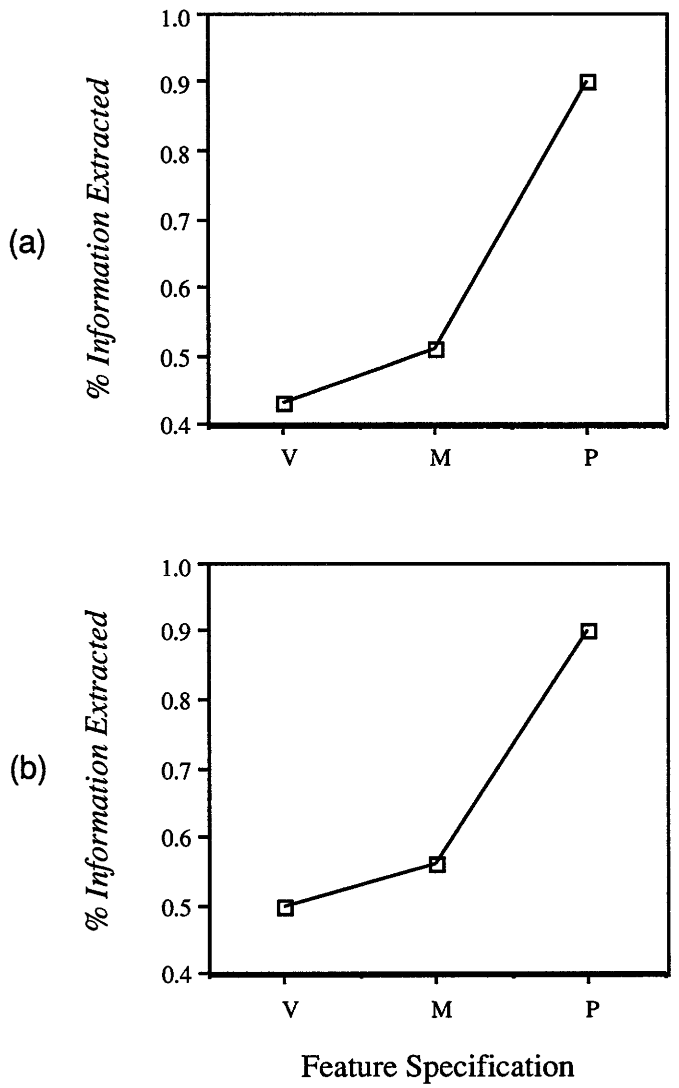


Figure 5.5: Percent information extracted comparing various levels of specificity used to transcribe the syllable CORE: (a) shows results computed from the frequency-unweighted lexicon, (b) shows results for for the frequency-weighted lexicon. The initials M, P, and V used to label the horizontal scale denote Manner, Place, and Voicing features respectively.

represent "islands of phonetic reliability" (i.e., places in the sound stream where phonetic information is most salient).

Where might these regions be, and how much lexical constraint does specifying these regions supply? Zue and his colleagues (cf., Huttenlocher and Zue 1984; Aull, 1984) have posed similar questions, and have conducted lexical partitioning studies directed at assessing the role that stressed syllables play during lexical access. Zue (1985) points out that stressed syllables represent such regions, and the results of lexical studies suggest that stressed syllable indeed are places that provide a relatively large amount of lexical constraint.

We may extend this line of reasoning to other phonological domains, using the methods developed in this thesis in the present and previous chapters. For example, we have strong intuitions that the syllable ONSET represents a place within the syllable where phonetic information is most salient. Our study of stop consonants reported in Chapter 4 confirms this intuition. We found some 5 major allophones of stops. We also observed a rather large variation in stop VOT. However, when stops are placed in the ONSET, they are primarily *released* (having a conditional entropy of .16 bits). In the CODA, their realizations are much less predictable (the entropy increases to 1.7 bits). Stops may be *released* in the syllable-final position, but they also may be *unreleased*, *deleted*, *glottalized*, and so on. Furthermore, when in the ONSET, VOT is a more reliable correlate of voicing for a stop than it is in the CODA. We may also add that, since stops are released in the ONSET we suspect that place-of-articulation features are more reliably represented . For example, the release spectrum is available to measure. Perhaps unsurprisingly, the syllable ONSET is a better location to specify features. In the study reported in this section, we found that, on average, the ONSET extracts the greatest amount of information concerning the lexical identity of an utterance.

5.3 Summary

In this chapter, we have probed the information structure of English syllables. Based on a set of quantitative studies, we concluded that the syllable embodies a set of local constraints that are non-uniform in their informational content. The nonuniformity goes in two directions. First, knowing the occupant of one slot in the syllable allows us to make inferences concerning other slots. It appears that the amount of certainty in these inferences is strongest among the members of the ONSET, and weakest between the ONSET and the CODA. The second dimension of nonuniformity in informational content has to do with the amount of constraint provided by the constituents of the syllable in eventual word identification. That is, if one were to select a location in the syllable in order to focus attention, it appears that the onset is the place to begin. If our studies on VOT and voicing are any indication of the perceptual salience of onsets, then the results presented here would be of considerable interest in any attempt to formulate a word recognition device.

Chapter 6

Concluding Remarks

The research reported in this thesis has had two components. The general problem addressed has been that of developing a framework for research in phonetics and phonology. The more specific problem has been one of defining the syllable at both the surface-phonemic and the systematic-phonetic levels of representation. The latter question has been addressed in Chapters 2 and 3, while Chapters 4, and 5 have dealt primarily with the more general problem.

Chapter 2 provided background on the syllable by pointing to a number of issues that a theory of syllable structure must address. A fundamental issue was how to define the syllable at both the surface-phonemic and phonetic levels of representation. The conclusion drawn was that phonetic syllables (as units that undergo physical concatenation) are ill-defined. Like the phonetic segment, there are no clear-cut principles for dividing an utterance into syllables on the basis of acoustic, articulatory, or any other kind of physical principles. As an alternative, it was suggested that syllables be defined as phonological objects that undergo a systematic process of acoustic-phonetic realization. Two formulations of phonological syllables were discussed in Chapter 2, and subsequently two previous models of phonetic syllable realization were presented. The first model, namely Church's proposal, was to use an immediate constituent grammar (typically used for representing phonological syllable structure) as a means of constraining an utterance's systematic phonetic representation. That is, allophones were to be assigned to the terminal positions of the syllable template. The second proposal, attributed to Fujimura and Lovins, suggests that the syllable be used as a principle of organization over a multi-dimensional phonetic representation comprised of articulatory gestures. Underlyingly, the terminal positions of the syllable template are associated with bundles of features. Features are implemented according to principles of vowel affinity (i.e., sonority sequencing) and syllable phonotactics.

The current model of the syllable was introduced and outlined in Chapter 3. The current proposal has attributes in common with both Church's framework and that of Fujimura and Lovins. Phonological syllables are hierarchically structured. Features are assigned to the terminal positions of the syllable template in bundles, where the assignment of *manner-of-articulation* features obeys the principles of sonority sequencing. Further constraints on patterns of *place* features and *voicing* are imposed by a set of syllable structure conditions stated as filters. Filters apply to the various non-terminal nodes of the syllable template, including the σ node itself in the case of agreement constraints between the syllable's CORE and AFFIX.

The current framework also incorporates a *realization component* which governs the mapping of phonological syllables onto the acoustic domain. The acoustic representation is assumed to consist of two kinds of properties: 1) *acoustic autosegments* that span regions of an utterance, and 2) *transitional measurements* that are to be extracted either at the transitions that delimit autosegments, or at other specified points in an utterance. In Chapter 3, a descriptive framework was introduced that allows constraints to be imposed on the acoustic representation as conditions of phonetic syllable well-formedness.

Chapter 4 was the first of two chapters that provided experimental data. Both chapters provided a set of methods that have general applicability in phonetics and phonology research, but were particularly useful for exploring syllable constraints on sound patterns. Chapter 4, for example, proposed using regression techniques based on *classification and regression trees* in order to discover the quantitative relationship between an utterance's underlying phonological description and certain acoustic measurements.

While Chapter 4 introduced methods for gaining a quantitative understanding of the acoustic-to-phonological mapping, Chapter 5 introduced quantitative methods for studying purely phonological and lexical constraints. Specifically, the notion of mutual information, first used in Chapter 4 as a node splitting criterion in tree regression, was used as a proximity measure for clustering the terminal positions within the syllable into an immediate constituent structure. Mutual information (in the form of *percent information extracted*) was also used as a criterion for evaluating and comparing parts of the syllable as sources of lexical constraint.

From this investigation, the following conclusions may be drawn. From a methodological perspective, the importance of bridging phonetics research and that of phonology cannot be understated. On the one hand, phonetics has traditionally been concerned with the physical properties of speech sounds, while phonology has addressed how sounds are patterned. Both paradigms contribute to one another. Large databases of speech materials are becoming increasingly important resources, containing naturally occurring phenomena which may be systematic in nature. A good phonetic theory is required in order to extract linguistically significant acoustic properties. At the same time, a phonological framework is required in order to ask meaningful questions concerning the relevance of acoustic properties as evidence of how languages are constrained. Computational methods provide yet a third research paradigm. They provide automated procedures for manipulating data and testing hypothesis. In this thesis, a computer algorithm for parsing, described in Appendix D, partially fulfills this latter role.

On a more specific note, we developed a coherent descriptive framework, based on the syllable, for stating constraints on an utterance's acoustic and phonological representations. Further, the acoustic representation need not be comprised of phoneme sized segments. Secondly, incorporating the syllable into a descriptive framework provides a quantifiable simplification in the effort required to specify phonological generalizations. Finally, without actually implementing an algorithm for lexical access, lexical partitioning experiments suggest that the syllable may be helpful in mitigating the process of lexical retrieval.

Suggestions for Extending the Current Work

Few investigations conclude without suggestions for future research. The current project has a number of extensions. Foremost are applications of the current set of proposals in the areas of speech recognition and synthesis. Extensions in these directions would include developing the current framework into a computational model for speech production and lexical access.

A computational model for lexical access, for example, would require that the following steps be completed. First, the "representative" set of acoustic properties proposed in Chapter 3, would have to be turned into into a more definitive set. Along with these efforts, further research is required in order to define signal processing algorithms for the automatic extraction of properties from the waveform, and acoustic studies are needed to gain understanding of the nature of variability during speech production. This understanding would have to be codified in the form of constraints to be included in the grammar. The current grammatical framework permits describing the mapping between the acoustic properties of an utterance and its underlying feature specification. A third step in constructing a model of lexical access would be to develop a procedure for retrieving lexical items on the basis of this feature specification. Finally, the grammar would need to be extended in order to incorporate the role played by larger suprasegmental units, such as the metrical foot, that are necessary for encoding prosodic constraints.

The experimental work begun in this thesis may also be continued and extended. Of particular interest would be to continue the study of interacting constraints. For example, in Chapters 4 and 5, it was noted that the syllable onset provides an environment of relative phonetic robustness for stop consonants, while at the same time providing a relatively large amount of lexical constraint. It will be useful to have future investigations using the same or a similar paradigm consider other classes of sounds, and differing phonological and syntactic contexts.

Bibliography

Aho, A. and Ullman, J. (1972). *The Theory of Parsing, Translation, and Compiling – Volume I: Parsing,* Prentice–Hall, Englewood Cliffs, NJ.

Aho, A.V., Hopcroft, J.E., and Ullman, J.D. (1974). *The Design and Analysis of Computer Algorithms,* Addison–Wesley, Reading, MA.

Allerhand, M. (1986). *A Knowledge-based Approach to Speech Pattern Recognition,* Ph.D. Thesis, Cambridge University, Cambridge, England.

Anderson, S.R. (1985). *Phonology in the Twentieth Century,* The University of Chicago Press, Chicago.

Aull, A.M. (1984). "Lexical Stress and its Application in Large Vocabulary Speech Recognition," S.M. Thesis, Massachusetts Institute of Technology.

Bahl, L.R., Jelinek, F., and Mercer, R.L. (1983). "A Maximum Likelihood Approach to Continuous Speech Recognition," IEEE Trans. Pattern Anal. Machine Intell., vol. PAMI-5, 170–190.

Bahl, L.R., Brown, P.F., de Souza, P.V., and Mercer, R.L. (1986). "Maximum Mutual Information Estimation of Hidden Markov Model Parameters for Speech Recognition," Proceedings Int. Conf. Acoustics, Speech, and Signal Processing, 49–52.

Baker, J.K. (1975). "Stochastic Modeling for Automatic Speech Recognition," in R. Reddy (Ed.), 521–542.

Bell A. and Hooper, J. (1978). *Syllables and Segments,* North-Holland, Amsterdam.

Bellman, R. (1957). *Dynamic Programming,* Princeton University Press, Princeton, NJ.

Bishop, Y.M., Fienberg, S.E., and Holland, P.W. (1975). *Discrete Multivariate Analysis: Theory and Practice,* MIT Press, Cambridge, MA.

Bradley, D. (1978). "Computational Distinctions of Vocabulary Type," Ph.D. Thesis, MIT.

Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984). *Classification and Regression Trees,* Wadsworth International Group, Belmont CA.

Broad, D. and Shoup, J. (1975). "Concepts for Acoustic-phonetic Recognition," in R. Reddy (Ed.), 243–274.

Browman, C. and Goldstein, L.M. (1986). "Towards an Articulatory Phonology," in *Phonology Yearbook 3,* Cambridge University Press, 219–252.

Carter, D.M. (1987). "An Information-theoretic Analysis of Phonetic Dictionary Access," Computer Speech and Language, Vol. 2.

Chomsky, N. (1965). *Aspects of the Theory of Syntax,* MIT Press, Cambridge, MA.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English,* Harper and Row, New York.

Christie, W.M. (1974). "Some Cues for Syllable Juncture in English," J. Acoust. Soc. Am., Vol 55, No. 4, 819-821.

Church, K. W (1983). *Phrase Structure Parsing: A Method for Taking Advantage of Allophonic Constraints,* Ph.D. Thesis, Massachusetts Institute of Technology.

Clements, G.N. (1988). "The Role of the Sonority Cycle in Core Syllabification," Working Papers of the Cornell Phonetics Laboratory, No. 2, April, 1–69.

Clements, G. N. and Keyser, S. J. (1983). *CV Phonology – A Generative Theory of the Syllable*, MIT Press, Cambridge, MA.

Cohen, P.S. and Mercer, R. (1975). "The Phonological Component of an Automatic Speech Recognition System," in Reddy (ed.), 275–320.

Denes, P. (1955). "Effect of Duration on the Perception of Voicing," J. Acoust. Soc. Am., Vol. 27, 761–764.

Draper, N.R. and Smith, H. (1966). *Applied Regression Analysis,* John Wiley and Sons, New York.

Egan, J. (1944). "Articulation Testing Methods II," OSRD Report No. 3802, U.S. Dept. of Commerce Report PB 22848.

Frazier, L (1987). "Structure in Auditory Word Recognition," in *Spoken Word Recognition,* U.H. Frauenfelder and L. Komisarjevsky Tyler (eds.), MIT Press, Cambridge, MA, 157–188.

Fujimura, O. (1975). "Syllable as a Unit of Speech Recognition," in IEEE Trans. Acoustics, Speech, and Signal Processing., Vol ASSP-23, No.1, 82–87.

Fujimura, O. (1981). "Elementary Gestures and Temporal Organization – What Does an Articulatory Constraint Mean?" in *The Cognitive Representation of Speech,* T. Myers, J. Laver, J. Anderson, eds., North-Holland, 101–110.

Fujimura, O. and Lovins J. (1978). "Syllables as Concatenative Phonetic Units," in *Syllables and Segments,* in (Bell and Hooper (eds.).

Fudge, E. C. (1960). "Syllables," Journal of Linguistics, 5, 253–286.

Gallager, R.G. (1968). *Information Theory and Reliable Communication,* John Wiley and Sons, Inc., New York.

Goldsmith, J. (1976). *Autosegmental Phonology,* available from Indiana University Linguistics Club.

Gnanadesikan, R. (1977). *Methods for Statistical Data Analysis of Multivariate Observations,* John Wiley and Sons, New York.

Hand, D. (1981). *Discrimination and Classification,* John Wiley and Sons, Chichester, England.

Harrington, J. and Johnstone, A. (1987). "The effects of word boundary ambiguity in continuous speech recognition," Proceedings of the Eleventh International Congress of Phonetic Sciences, Vol. 4, Tallinn, Estonia, Se 45.5.1–4.

Hogg, R. and McCully, C.B. (1987). *Metrical Phonology – A Coursebook,* Cambridge University Press, Cambridge, England.

Hooper, J. (1972). "The Syllable in Phonological Theory," Language, Vol. 48, No. 3, 525–540.

Huttenlocher, D.P. and Zue, V.W. (1984). "A Model of Lexical Access Based on Partial Phonetic Information," Proc. Int. Conf. on Acoustics, Speech, and Signal Processing, 26.4.1–4.

Huttenlocher, D.P. and Withgott, M. (1986). "On Acoustic Versus Abstract Units of Representation," in Proceedings of Montreal Symposium on Speech Recognition, 61–62.

Kahn, D. (1976). *Syllable-based Generalizations in English Phonology,* Ph.D. Thesis, Massachusetts Institute of Technology.

Kaplan, R.M. and Bresnan, J. (1982). "Lexical-Functional Grammar: A Formal System for Grammatical Representation," in *The Mental Representation of Grammatical Representations*, J. Bresnan (ed.), MIT Press, Cambridge, MA.

Kay, M. (1986). "Parsing in Functional Unification Grammar," in *Readings in Natural Language Processing,* B.J. Grosz, K.S. Jones, and B.L. Webber (eds.), Morgan Kaufman Publishers, Inc., Los Altos, 125–138.

Keating, P. (1985). "The Phonology-Phonetics Interface," UCLA Working Papers 62, 1–20.

Kent, R.D. and Minifie, F.D. (1977). "Coarticulation in Recent Speech Production Models," J. of Phonetics, Vol. 5, 115–133.

Klatt, D.H. (1975). "Voice Onset Time, Frication, and Aspiration in Word-initial Consonant Clusters," J. Speech Res. Hearing, Vol. 18, No. 4, 689–707.

Kohler, K.J. (1966). "Is the Syllable a Phonological Universal," Journal of Linguistics, Vol. 2, 207–208.

Ladefoged, P. (1975). *A Course in Phonetics,* Harcourt Brace Jovanovich, Inc., New York.

Lehiste, I. (1960). "An Acoustic-phonetic Study of Internal Open Juncture," Phonetica Suppl. 5.

Kucera, H. and Francis, W. (1967). *Computational Analysis of Present-Day American English,* Brown University Press, Providence, RI.

Lamel, L.F, Kassel, R.H., Seneff, S. (1986). "Speech Database Development: Design ad Analysis of the Acoustic-phonetic Corpus," Proc. DARPA Speech Recognition Workshop, Report No. SAIC-86/1546, 100–109.

Leung, H. (1985). *A Procedure for Automatic Alignment of Phonetic Transcriptions with Continuous Speech,* S.M. Thesis, Massachusetts Institute of Technology.

Levinson, S.E. (1985). "Structural Methods in Automatic Speech Recognition," Proc. IEEE, Vol 73, No. 11, 1625–1651.

Lewis, H.R. and Papadimitriou, C.H. (1981). *Elements of the Theory of Computation,* Prentice–Hall, Inc., Englewood Cliffs, NJ.

Lisker, L. and Abramson, A.S. (1964). "Some Effects of Context on Voice Onset Time in English Stops," Language and Speech, Vol 10 1–28.

Martin, W.A., Church, K.W., and Patil, R.S. (1987). "Preliminary Analysis of a Breadth-First Parsing Algorithm: Theoretical and Experimental Results," in *Natural Language Parsing Systems*, L. Bolc (ed.), Springer–Verlag, Berlin.

Malecot, A. (1960). "Vowel Nasality as a Distinctive Feature in American English," Language, vol. 36, 222–229.

McCullagh, P. and Nelder, J.A. (1983). *Generalized Linear Models,* Chapman and Hall, New York.

Mermelstein, P. (1975). "Automatic Segmentation of Speech into Syllabic Units," J. Acoust. Soc. Am., Vol 58, No. 4, 880–883.

Mohannan, K.P. (1985). "Syllable Structure and Lexical Strata in English," in *Phonology Yearbook 2,* Cambridge University Press, Cambridge, MA, 139–155.

Nakatani L. and Dukes, K. (1977) "Locus of Segmental Cues for Word Juncture," J. Acoust. Soc. Am., Vol 62, No. 3, 714–719.

Nussbaum, H.C. and Pisoni, D.B. "The Role of Structural Constraints in Auditory Word Recognition," in Proceedings of Montreal Symposium on Speech Recognition, Montreal, 57–58.

Pike, K. and Pike, E. (1947). "Immediate Constituents of Mazatec Syllables," Int. J. of Am. Ling., 13, 78–91.

Price, P.J. (1980). "Sonority and Syllabicity: Acoustic Correlates of Perception," Phonetica, Vol. 37, 327–343.

Reddy, R. (ed.) (1975). *Speech Recognition*, Academic Press, New York.

Reilly, W. (1986). "Asymmetries in Place and Manner in the English Syllable," in Research in Phonetics and Computational Linguistics, No. 5, 123–148.

Selkirk, E. O. (1982). "The Syllable," in *The Structure of Phonological Representations – Part II,* H. van der Hulst and N. Smith, eds. (Foris Publications, Dordrecht), 337–385.

Selkirk, E.O. (1984). "On the Major Class Features and Syllable Theory," in *Language Sound Structure,* M. Arnoff, R. Oehrle, B. Wilker, and F. Kelley, eds., MIT Press, Cambridge, MA.

Shadle, C.H. (1985). *The Acoustics of Fricative Consonants,* Technical Report 506, MIT Research Laboratory of Electronics.

Shipman, D.W. and Zue, V.W. (1982). "Properties of Large Lexicons: Implications for Advanced Isolated Word Recognition Systems," in Proc. Int. Conf. on Acoustics, Speech , and Signal Processing, 546–549.

Smith, A. (1976). *Word Hypothesization in the Hearsay-II Speech System,* Ph.D. Thesis, Carnegie–Mellon University.

Steriade, D. (1988). Review of "Clements and Keyser: CV Phonology," Language, Vol. 64, No. 1, 118–130.

Stevens, K.N. (1986). "Models of Phonetic Recognition II:An Approach to Feature-based Recognition," in Proceedings of Montreal Symposium on Speech Recognition, Montreal, 67–68.

Stevens, K.N. (1987). "Relational Properties as Perceptual Correlates of Phonetic Features," Proceedings of the Eleventh International Congress of Phonetic Sciences, Vol. 4, Tallinn, Estonia, 352–356.

Stevens, K.N. and Blumstein, S.E. (1978). "Invariant Cues for Place of Articulation in Stop Consonants," J. Acoust. Sol. Am., Vol. 64, No. 5, 1358–1368.

Velleman, P.F. and Hoaglin, D.C. (1981). *Applications, Basics, and Computing of Exploratory Data Analysis,* Duxbury Press, Boston, MA.

van der Hulst, H. and Smith, N. (1982). "An Overview of Autosegmental and Metrical Phonology," in *The Structure of Phonological Representations – Part I,* H. van der Hulst and N. Smith, eds., Foris Publications, Dordrecht, 337–385.

Winer, B.J. (1962). *Statistical Principles in Experimental Design,* McGraw-Hill, New York.

Winston, P.H. (1984). *Artificial Intelligence*, Addison–Wesley, Reading, MA.

Woods, W., et. al., (1976). *Speech Understanding Systems: Final Report,* Bolt Beranek and Newman Inc., BBN Report No. 3438.

Zue, V.W. (1983). "The Use of Phonetic Rules in Automatic Speech Recognition," Speech Communication, Vol. 2, 181–186.

Zue, V.W. (1985). "The Use of Speech Knowledge in Automatic Speech Recognition," Proc. IEEE, Vol 73, No. 11, 1602–1615.

Zue, V.W. (1986). "Models of Phonetic Recognition III: The Role of Analysis by Synthesis in Phonetic Recognition," in Proceedings of Montreal Symposium on Speech Recognition, Montreal, 69–70.

Zue, V. and Schwartz, R.M. (1980). "Acoustic Processing and Phonetic Analysis," in *Trends in Speech Recognition,* W. Lea, ed., Prentice–Hall, Inc., Englewood Cliffs, NJ, 101–125.

Zue, V. and Sia, E. (1982). "Nasal Articulation in Homogamic Clusters in American English," Working Papers of the MIT Speech Communication Group, Vol 1, 9–19.

Zue, V.W., et al. (1986). "The Development of the MIT Lisp-machine Based Speech Research Workstation," in Proc. Int. Conf. on Acoustics, Speech , and Signal Processing, 329–332.

Appendix A

Speech Materials, Databases, and Tools

This appendix provides details of the speech materials used in the experiments reported in Chapters. There are two main sections. Section A.1 describes speech databases and Section A.1.1 describes tools. In particular, Section A.1 outlines how data was prepared for the acoustic study summarized in Chapter 4. Section A.1.1 describes computer software that has assisted in the collection and analysis of this data.

A.1 Source of Acoustic Data

Speech material was obtained from three separate databases, all of which were developed at MIT. They are labelled *JP, HL,* and *TIMIT.* Database JP consisted of utterances spoken by a single talker. The corpus was the first 600 of the well-known Harvard list of phonetically-balanced sentences (Egan, 1944). Database HL contained utterances spoken by 100 talkers, 50 male and 50 female. The speech material in this database was also drawn from the Harvard list. Each of the speakers uttered ten sentences, producing a total of 1000 sentences. Database TIMIT was a subset of the TIMIT sentences: a corpus developed at MIT and recorded at Texas

Instruments (Lamel *et al.*, 1986). This databases was also multi-speaker, containing 2154 utterances spoken by 376 speakers. Each speaker read from 5 to 7 sentences. Two of these 9 were so-called *calibration* sentences. They contained phoneme sequences intended to elicit certain kinds of phonological variation. No special use of these sentences was made in our studies. We simply pooled the tokens obtained from these sentences along with the rest of our data.

Associated with each of the utterances used in our study was a variety of time-aligned transcriptions. Utterances were originally transcribed with a *narrow* phonetic transcription. For example, if the phoneme /t/ were realized with both a closure and a release, it would be transcribed as [t$^{\square}$t] under this transcription. The narrow phonetic transcription would also mark flaps and glottalized stops with their own special symbols. Phonemic transcriptions were also provided. These were obtained from a pronunciation dictionary and semi-automatically aligned with the phonetic transcription.[1] Also aligned with the phonetic transcription was a syllabified phonemic transcription of each utterance, along with a phonemic transcription containing word boundaries.

In addition to differences in the numbers of speakers, the two corpora (Harvard list and TIMIT) used in the acoustic study also differed in their distributions of word types, style of sentences, and recording conditions. The Harvard list sentences recorded for databases JP and HL were relatively homogeneous in sentence type and phonetically simple. They consisted of simple declarative sentences, in which the average number of syllables per word was approximately 1.1. On the other hand, the TIMIT sentences contained a variety of sentence types and an average number of syllables per word of 1.58. All three databases were recorded in sound treated rooms. Databases JP and HL were recorded at MIT using a Sony *omni-directional* microphone placed on the speaker's chest. Database TIMIT was recorded at Texas Instruments using a Sennheiser close-talking microphone. It should be pointed out

[1] A two pass procedure was used. In the first pass, a completely automatic phonemic-to-phonetic alignment program was run for each of the utterances. Errors in the alignments were then corrected by hand.

Database	# of Tokens
TIMIT	12161
HL	2615
JP	3374

Table A.1: Total number of tokens collected from each of the databases for the acoustic study reported in Chapter 4.

that the omni-directional microphone was able to pick up sound radiated from tissue vibration, in addition to sound from both the oral and nasal cavities. As a result, weak voiced sounds have more energy at low frequencies in the databases recorded at MIT.

A.1.0.1 Selection of Tokens

From each of the databases, stop consonants were selected. Using the phonemic transcription as a guide, a search was performed whereby stop consonants in a variety of local phonetic environments and syllable positions were selected. Excluded from the study were stop consonants that were located in the sentence-initial and sentence-final positions of utterances. For the case of syllable-final stops in particular, it was felt that the acoustic characteristics of stops in these two sentence positions would vary unsystemically. Table A.1 gives the numbers of tokens collected from each of the three databases.

A.1.0.2 Measurements Obtained

For each token, we obtained a number of acoustic measurements. These included, the the stop's closure duration, its release duration (i.e., its Voice Onset Time or VOT), and durations of segments in the utterance that surrounded the stop. These measurements were obtained on the basis of the phonetic transcription. That is, stops realized as both a closure and a release were transcribed with two symbols.

For example, $\mathbf{t}^{\square}\mathbf{t}$ would be the transcription for a /t/. The closure duration was determined from the duration of the $\mathbf{t}^{\square}$ transcription token, and the release duration was determined from the duration of the /t/ token.

In Chapter 4 we also report results of an experiment in which we examined stop realization. On the basis of VOT and Closure Duration, we were able to ascertain whether a stop was *released* (VOT $>$ 0), *unreleased* (VOT $=$ 0), or *deleted* (VOT $=$ 0 and Closure Duration $=$ 0). In addition, we could determine whether a stop was flapped or realized as a glottal stop from the phonetic transcription directly.

At this juncture, we should point out a number of problems that arise when using a phonetic transcription in an acoustic study of this size. These problems have to do with errors in the phonetic transcriptions themselves, and the lack of consistency among transcribers. Errors, although infrequent in our study, must be expected to occur. The errors that we have found in the transcriptions of stop consonants are typically of a limited number of types. For example, stop releases that are weak often have gone undetected. There is also a bias towards transcribing /t/'s as glottal stops when pitch irregularities are detected in the waveform, but in situations where /p/'s and /k/'s had similar waveform characteristics they were transcribed as being unreleased. Perhaps a more pervasive problem in transcribed databases, is maintaining consistency in the transcriptions among the several phoneticians who participate in the segmentation and labelling processes. Although this problem might be lessened to a certain extent with stop consonants, it is reasonable to assume that individual phonetic impressions of a given speech sound will differ.

A.1.1 Tools

In this section, we will describe software that we and others have written to assist in the collection and preparation of the data outlined in this and the following sections. Specifically, we will describe the *Spire* and *Search* programs developed at MIT.

All experiments were carried out on a Symbolics 3600 series Lisp Machine with at least 4 Mbytes of main memory, and a 474 Mbyte disk. For some of the studies, the machine used was augmented with either internal or external floating point processing hardware. All programs described in this section are implemented in the Lisp programming language under Software Genera 7.

A.1.1.1 Spire

Spire (Speech and Phonetics Interactive Research Environment) was originally developed by David Shipman. More recent versions have been developed by Scott Cyphers and David Kaufman. It is a software package that enables users to digitize, transcribe, process and examine speech signals (Zue *et al.*, 1986). We used this facility primarily for the purpose of displaying utterances. In conjunction with the program *Search* (described below), we used *Spire* to examine statistical outliers.

Spire also provides the capability for programers to write their own interactive programs that include its data displays for the purpose of providing output. An instance of such a program is the one that has been implemented for the purpose of debugging the set of constraints described in Chapter 6.

A.1.1.2 Search

Search (Structured Environment for Assimilating Regularities of speeCH) is a program developed by the author along with Charles Jankowski and David Kaufman (Zue *et al.*, 1986). It is designed principally for the exploration of relatively large quantities of data. It is interactive and includes a number of statistical displays, including: *histograms*, *scatter plots*, *profile displays*, and *boxplots* (Velleman and Hoaglin, 1981). *Search* also interfaces with *Spire*. As such, *Spire* provides the principle source of data that is to be explored in *Search*. Conversely, *Search* incorporates a feature whereby individual tokens selected from one of the statistical displays may be investigated in detail by reading in the token's source utterance into a *Spire* display.

The basis of *Search* is a set of procedures for designing classification trees from a learning sample. The primary method is the CART (Classification and Regression Trees) algorithm described in Chapter 4. CART is a supervised classification algorithm that generates a binary decision tree using a maximum-mutual information criterion. Alternative methods for producing trees include unsupervised clustering procedures, as well mechanisms that *Search* provides for splitting nodes by hand.

Aside from its use for exploring data and for statistical analysis, *search* provides the capabilities of a more general-purpose database management tool. Internally, collections of data points are stored as *relations*: two dimensional tables of tokens and attributes (cite reference on relational databases). Similar structures are maintained for utterance objects. In the utterance relation, tables are indexed by its filename. The attributes are comprised of transcriptions of the varying types discussed above.

Appendix B

Principles of Information Theory and the Evaluation of Phonological Constraints

In this appendix, we will state some fundamental results of information theory. These results form the basis of the node splitting criteria used in growing classification and regression trees in Chapter 4, and the objective measures of phonotactic and lexical constraints used in Chapter 5. For those readers who are familiar with information theory, this appendix will contain little, if any, new information. For those who are interested in details that are not provided, much of the material presented here has been drawn from Gallager (1968).

Information theory is based on the notion that a sample space of events is to be characterized in terms of a probability measure. In so far as such a characterization is appropriate, it provides a set of tools for quantitatively measuring the amount of information conveyed by the observance of any particular event. The sample space itself, may be comprised of discrete events, in which case, the probability measure will be a discrete probability function. The sample space may also be continuous, in this case, the probability function is likewise continuous. In the present section, we will present pertinent results for the discrete case. The reader is referred to Gallager (1968) for results for continuous ensembles.

The remainder of this appendix is organized into two parts. In Section B.1, we define expressions for *entropy*, *conditional entropy*, and *mutual information.* In Section B.2, we will specialize the results of Section B.1 for the problem of evaluating lexical partitions. In particular, we will describe the details of the *percent information extracted* measure proposed by Carter.

B.1 Entropy, Conditional Entropy, and Mutual Information

For developing the material presented in this section, we have adopted a rather general notation. We will let the set $\{a_1, \ldots, a_K\}$ be the sample space for some ensemble X. This sample space is characterized by the probability assignment $P_X(a_k)$. Similarly, $\{b_1, \ldots, b_J\}$ will be a sample space that is associated with some ensemble Y. It is characterized by the probability assignment $P_Y(b_j)$. Finally, XY is the joint ensemble with probability assignment $P_{X;Y}(a_k; b_j)$.

In our use of information-theoretic concepts in the main body of this thesis, X and Y have denoted three kinds of ensembles. In Chapter 4, for example, mutual information was used as the criteria for judging the quality of a node's split in a regression tree. In this context, X was an ensemble of acoustic realizations, the members of the ensemble Y were characterizations of their corresponding contexts through each member's assignment to nodes of the tree. In Chapter 5, mutual information was used in two contexts. Its first use was for cluster analysis, in Section 5.1, the ensembles X and Y were the occupants of two terminal nodes of the syllable template. In Section 5.2, mutual information was used to evaluate the quality of a lexical partitioning. In this context, X was used to denote a lexicon of syllables, for each element of X, elements of Y denoted its membership in some lexical cohort.

Of general interest in all of these contexts are quantitative measures pertaining to the following three questions. 1) On the basis of $P_X(a_k)$, what is the uncertainty of the event $x = a_k$? In other words, how much information is required to resolve the uncertainty of this event? 2) Given that X and Y are jointly distributed according to the probability assignment $P_{X;Y}(a_k; b_j)$, what is the uncertainty of the event $x = a_k$, given that one already has knowledge of the occurrence of the event $y = b_j$? 3) How much uncertainty does knowledge about the event $y = b_j$ resolve about the event $x = a_k$?

Entropy is the quantitative measure pertaining to the first of these questions. Its definition is given as (B.1).

$$I_X(a_k) = -\log P_X(a_k). \tag{B.1}$$

There are three aspects of this expression that are worth noting. First, entropy, $I_X(a_k)$ is a function of the random variable a_k. Therefore, it itself is a random variable. The average entropy of an ensemble is defined as (B.2).

$$H(X) = -\sum_{k=1}^{K} P_X(a_k) \log P_X(a_k). \tag{B.2}$$

The second aspect of the expression given in (B.1) is that the uncertainty or entropy of some event is inversely related to its probability, $P_X(a_k)$. That is, the higher the probability of a particular event, the lower its entropy (uncertainty) and visa versa. Finally, it can be shown that average entropy is maximum when all events are equiprobable, that is, when $P_X(a_k) = 1/K$ (See Gallager (1968, p. 23) for a proof of this latter statement).

The uncertainty measure that pertains to the second of the above questions is conditional entropy. It is defined as

$$I_{X|Y}(a_k|b_j) = -\log P_{X|Y}(a_k|b_j), \tag{B.3}$$

where

$$P_{X|Y}(a_k|b_j) = \frac{P_{X;Y}(a_k;b_j)}{P_Y(a_j)}.$$

The quantity $I_{X|Y}(a_k|b_j)$ is also a random variable. It is defined over the joint ensemble XY, its average is given as

$$H(X|Y) = -\sum_{k=1}^{K}\sum_{j=1}^{J} P_{X;Y}(a_k;b_j) \log P_{X|Y}(a_k|b_j), \tag{B.4}$$

Finally, mutual information is the relevant measure for change in uncertainty (Question 3 above). Its definition is given as (B.5).

$$I_{X;Y}(a_k;b_j) = I_X(a_k) - I_{X|Y}(a_k|b_j) \tag{B.5}$$

The expression given in (B.5) defines mutual information as the amount of uncertainty removed by having conditional information. It is the definition that matches the intuition given above. A completely equivalent definition of mutual information is given as (B.6).

$$I_{X;Y}(a_k;b_j) = \log \frac{P_{X;Y}(a_k;b_j)}{P_X(a_k)P_Y(b_j)}. \tag{B.6}$$

Average mutual information is computed over the joint ensemble XY and is given as (B.7).

$$I(X;Y) = \sum_{k=1}^{K}\sum_{j=1}^{J} P_{X;Y}(a_k;b_j) \log \frac{P_{X;Y}(a_k;b_j)}{P_X(a_k)P_Y(b_j)}. \tag{B.7}$$

It follows from (B.5) that average mutual information may also be written as

$$I(X;Y) = H(X) - H(X|Y) \tag{B.8}$$

B.1.1 Properties of Mutual Information as a Measure of Collocational Constraint

In the main body of this thesis, we have made use of mutual information as a measure of collocational constraint. On the basis of a cluster analysis, using this measure, we were able to "derive" the shape of the syllable's hierarchical structure. In the next few paragraphs, we will discuss the mathematical properties of mutual information in regards to its suitability for this purpose.

For use in cluster analysis, a *distance measure* or *measure of similarity*, $d(x,y)$, will satisfy the following two conditions:

1. $d(x,y) = d(y,x)$.
2. $d(x,y) \geq 0$; $d(x,x) = 0$

Condition 1 states that d is symmetric, Condition 2, states that d is positive-semi-definite. If $d(x,y)$ satisfies a third condition, called the *triangle inequality,*

$$d(x,y) \leq d(x,z) + d(z,y),$$

it is called a *metric.*

From (B.7) it is easy to show that the function for mutual information is symmetric,

$$I(X;Y) = I(Y;X). \tag{B.9}$$

One can also show that the function for mutual information is positive semi-definite (see, Gallager 1968, p. 24),

$$I(X;Y) \geq 0. \tag{B.10}$$

Under general circumstances, however, mutual information does not satisfy the triangle inequality. The special case of when it does occurs when the ensembles X and Y are conditionally independent. That is,

$$I(X;Y) \leq I(X;Z) + I(Z;Y) \tag{B.11}$$

if, and only if,

$$P(x,y|z) = P(x|y)P(y|z). \tag{B.12}$$

In our use of mutual information, for the study of phonotactic constraints, we suspect that the condition given as (B.12) would be rarely satisfied.

B.1.2 Percent Information Extracted

Before ending the discussion on basic concepts of information theory, we will given an additional interpretation of mutual information. The alternative interpretation of this quantity will lead to a definition of the *percent information extracted* measure used in Chapters 4 and 5.

In the above discussion, entropy was given a dual interpretation. It first was defined as a measure of uncertainty concerning the outcome of some event. It is likewise a measure of the amount of information that is required to resolve this uncertainty. If the event in question is the emission of some symbol from a information source (e.g., a grammar) one may also think of the occurrence of an event as the means by which the source puts out information. Intuitively, the more probable the symbol, the less information that the source is providing. Conversely, a symbol that is less probable conveys more information.[1] On the basis of this reasoning, one may establish $H(X)$ as the average information that is contained in a source about the symbol x.

By observing the symbols that are put out by a source, the observer is *extracting* information. The inequality given in (B.13) provides a mathematical justification for this statement.

$$H(X) \geq H(X|Y). \tag{B.13}$$

In words, the expression given in (B.13) says that, on average, a source contains less information about the outcome x once the outcome y has observed. These interpretations of entropy and conditional entropy lead to an interpretation of mutual information as a measure of the *amount of information extracted* from a source. This interpretation is most apparent when average mutual information is written as

$$I(X;Y) = H(X) - H(X|Y). \tag{B.14}$$

Once mutual information is viewed in such terms, then it is natural to think of relative measures of this quantity. If $H(X)$ is the *a priori* amount of information contained in a source, then one can define the *percent information extracted*, as the expression given in (B.15).

$$\text{Percent Information Extracted} = \frac{H(X) - H(X|Y)}{H(X)} \times 100\% \tag{B.15}$$

B.2 Carter's Evaluation Metric

The notion of "information extracted" provides the basis of an evaluation metric proposed by Carter (1987) for use in studying lexical constraints. For the remainder

[1] In the extreme case, the entropy of an event with probability 1 is zero, in which case, the source is providing no information. In such a situation one could guess the outcome of the event with certainty.

of this section, we will derive it.

We are given a lexicon L_σ, in our case, containing the syllables, $\sigma_1, \ldots, \sigma_N$. Each syllable, σ_i, has probability p_i of occurring in some spoken message. We would like to study the consequences of applying the transcription $\mathcal{T}$ to each of L_σ's entries. Since, in general, each lexical entry will no longer be unique under this representation, $\mathcal{T}$ has the effect of partitioning L_σ into a set of equivalence classes, $\Pi(\mathcal{T})$.

There are a number of ways of characterizing the resulting lexical partitioning. Carter suggests that one should view $\mathcal{T}$, when applied to an unknown word, as having the effect of extracting information from L_σ. That is, once one has identified the partition that σ_i belongs to, less information is required in order to determine its complete identity. Carter proposes percent information extracted for the purpose of measuring the change in required information.

Prior to applying any transcription, the lexicon puts out symbols with average entropy

$$H(L_\sigma) = -\sum_{\sigma_i \in L_\sigma} p_i \log p_i. \tag{B.16}$$

$H(L_\sigma)$ is the measure of information that has yet to be extracted.

On the basis of $\mathcal{T}$ a given lexical item, σ_i, will be assigned to a partition α, where $\alpha \in \Pi(\mathcal{T})$. The item, σ_i's conditional probability, once this assignment is made, is given in (B.17),

$$p\{\sigma_i | \sigma_i \in \alpha\} = \frac{p_i}{P_\alpha}, \tag{B.17}$$

where

$$P_\alpha = \sum_{\sigma_k \in \alpha} P_k, \tag{B.18}$$

is α's probability of occurrence.

Once the assignment, $\sigma_i \in \alpha$, is made, the information that remains to be extracted is calculated as the entropy of the ensemble of lexical items contained in α. This quantity, $H(\alpha)$, is defined in (B.19)

$$H(\alpha) = -\sum_{\sigma_i \in \alpha} P\{\sigma_i | \sigma_i \in \alpha\} \log P\{\sigma_i | \sigma_i \in \alpha\}. \tag{B.19}$$

On average, the information that is extracted from a lexicon, once a lexical partition has been identified is given by the expression in (B.20).

$$H(L_\sigma | \Pi(\mathcal{T})) = \sum_{\alpha \in \Pi(\mathcal{T})} P_\alpha H(\alpha). \tag{B.20}$$

Substituting (B.18) and (B.17) into (B.20), one obtains

$$H(L_\sigma | \Pi(\mathcal{T})) = H(L_\sigma) - \sum_{\alpha \in \Pi(\mathcal{T})} P_\alpha \log P_\alpha. \tag{B.21}$$

Upon further manipulation of (B.21), the quantity $H(L_\sigma | \Pi(\mathcal{T}))$, may be rewritten

$$-\sum_{\alpha \in \Pi(\mathcal{T})} P_\alpha \log P_\alpha = H(L_\alpha) - H(L_\alpha | \Pi(\mathcal{T})). \tag{B.22}$$

The expressions on both the left and right-hand sides of the equation given in (B.22) are measures of a lexical partitioning's quality. Intuitively, a good lexical partitioning is one in which, once a partition, α, has been identified, the remaining uncertainty about a given lexical entry is small. This occurs when $H(L_\alpha | \Pi(\mathcal{T}))$ is minimized, or when the quantity on the right-hand side of (B.22) is maximized. This latter quantity is the amount of information that $\mathcal{T}$ extracts from the lexicon L_σ. One may also note that a transcription, $\mathcal{T}$ is optimal, when it maximizes the average entropy of the distributions of lexical partitions (the quantity on the left-hand side of (B.22)).

The appropriate relative measure for evaluating a lexical partition is given as (B.23).

$$\mathrm{PIE}(L; \Pi(\mathcal{T})) = \frac{H(L_\sigma) - H(L_\sigma | \Pi(\mathcal{T}))}{H(\mathcal{T})} \times 100\% \qquad \text{(B.23)}$$

Appendix C

Formal Rules and Representation

In this appendix, the *constraint description language* developed in this thesis for expressing theories of acoustic-phonological representation is presented. A constraint is defined as a means of expressing a relationship between objects. In an utterance's acoustic-phonological representation, objects are *acoustic-properties*, *phonological features*, and *phonological categories*, that, at times, are to be arranged into hierarchical structures. Therefore, the language allows these elements of an utterance's description to be defined and manipulated. In the main body of the thesis, examples have been given of how this language has been used for stating syllabic constraints. In this appendix, emphasis is placed on the language's syntax.

This appendix is to treated like a "programmer's reference manual," analogous to a manual describing any general purpose programming language, such as Fortran, Lisp, or Pascal. Like programming languages in general, the way in which the language is described in this appendix reflects the intention that it be theory-neutral. That is, a theory of syllable structure is but one, of a range of ideas, that can be expressed with this descriptive framework.

This appendix is organized into five principal sections. Section C.1 reviews attribute-value matrices: the general notational device used to assign analyses to utterances, as well as to state grammatical constraints on the acoustic-phonological representation. Section C.2 gives a definition of functional descriptions. In particular, the primitives of an utterance's acoustic-phonological description are defined as well as the means by which primitives are combined to produce an overall analysis. In Sections C.3 through C.5 the elements of the constraint description language itself are presented. Section C.3 describes the means by which hierarchical constraints are stated. Sections C.4 and and C.5 describe filters and realization rules respectively. These constructs extend the language in order to capture constraints on processes such as *agreement*, *assimilation*, *acoustic realization.*

C.1 Attribute-value Matrices

As mentioned above, the attribute-value matrix is a general notational device that serves a dual purpose in the descriptive framework. First it is used for stating grammatical constraints (e.g., realization rules). It is also used to display the structural analysis assigned to an utterance. In serving this latter purpose, an attribute-value matrix is called an utterance's *functional description* (see below).

Intuitively, an attribute-value matrix is an object that bears information describing the *attributes* of some grammatical construct by specifying its *values.* When written, an attribute-value matrix is an indefinite list of attribute-value pairs, that are enclosed in square brackets:

R C.1

$$\begin{bmatrix} a_1 & v_1 \\ a_2 & v_2 \\ \vdots & \vdots \\ a_n & v_n \end{bmatrix}.$$

In (R C.1) the terms a_i denote attributes and v_i denote values.

A useful property of attribute-value matrices is their ability to encode hierarchical relationships. This is done by allowing these structures to be the values that specify other attributes. An example is given in (R C.2).

R C.2

$$\begin{bmatrix} a_1 & \begin{bmatrix} a_{11} & v_{11} \\ a_{12} & v_{12} \end{bmatrix} \\ a_2 & v_2 \\ a_3 & \begin{bmatrix} a_{31} & \begin{bmatrix} a_{311} & v_{311} \\ a_{312} & v_{312} \end{bmatrix} \\ a_{32} & v_{32} \end{bmatrix} \end{bmatrix}$$

In the attribute-value matrix shown in (R C.2), the embedding is two levels deep.

C.1.1 Paths

Because AVM's are sometimes embedded, it is useful to have a notion of a *path*, where paths are sequences of attributes enclosed in parenthesis. Further, associated with paths are values. For example, the value in the AVM shown in (R C.2) associated with the path $(a_1\ a_{11})$ is v_{11}. Likewise, the value associated with the path $(a_3\ a_{31})$ is the attribute-value matrix

$$\begin{bmatrix} a_{311} & v_{311} \\ a_{312} & v_{312} \end{bmatrix}$$

C.1.2 Shared Structures

Another feature of attribute-value matrices is the facility they provide for sharing information (i.e., instances when the same value is used to specify more than one attribute). An example is given in(R C.3)

R C.3

$$\begin{bmatrix} a_1 & \begin{bmatrix} a_{11} & \boxed{1} \begin{bmatrix} a_{111} & v_{111} \\ a_{112} & v_{112} \end{bmatrix} \end{bmatrix} \\ a_2 & v_2 \\ a_3 & \begin{bmatrix} a_{31} & \boxed{1} \end{bmatrix} \end{bmatrix}$$

In this example, the structure that is shared is indicated by the *tag* "$\boxed{1}$".

In using attribute-value matrices as functional descriptions, the facility provided for sharing of structures may be useful to denote categories that are assigned to more than one constituent. For example, if one were to represent ambisyllabic consonants, sharing would provide a convenient notation.

C.2 Functional Descriptions

As indicated above, when used to describe the analysis assigned to an utterance by a grammar, attribute-value matrices are called *functional descriptions* (or fd's). At present an utterances's fd consists of four attributes, 1) the name of a *constituent*, 2) a set of *conditions* describing the utterance's surface-phonemic representation, 3) the utterance's acoustic *realization*, and 4) a *score*. The general format is given in (R C.4):

R C.4 (Functional Description)

$$\begin{bmatrix} \mathit{Constituent} & - \\ \mathit{Conditions} & - \\ \mathit{Realization} & - \\ \mathit{score} & - \end{bmatrix}$$

The score reflects how well the conditions "match" the acoustic realization. For a given set of acoustic properties that comprise the value of the *realization* attribute, there generally will be more than one *conditions* value that may correspond. The score reflects how well a given phonological description fits the acoustic data. In the discussion on realization rules below, the facility provided for computing and assigning scores will be described.

Through the capability for attribute-value matrices to be embedded inside one another, functional descriptions may naturally represent hierarchical structures common in metrical phonology (syllable structure in particular). The primitives of functional descriptions are *specified features* and *acoustic properties.*

C.2.1 Specified Features

Specified features are attribute-value pairs consisting of the name of the feature (as the attribute) and its specification (the value). An example is given in (R C.5):

R C.5

$$\begin{bmatrix} f & v \end{bmatrix},$$

In (R C.5) f denotes the feature name, and v its specification. One should note that the notation given in (R C.5) differs from how specified features are written in running text (i.e., as [v *f*]). In particular, the values for f and v are reversed.

Specified features are phonological objects, therefore, their specifications may only denote symbolic values. In particular, the specifications may be constants, $+, -$, or x, or in the case when specified features are incorporated into rules, values may be variables (e.g., α, β, etc.).

C.2.2 Acoustic Properties

Acoustic properties are primitives pertaining to an utterance's acoustic representation. At present, properties are not understood well enough to defined them as concretely as specified features. That is, any number of attributes may be used to specify an acoustic property. Therefore acoustic properties are represented as general attribute-value matrices, with an indeterminate set of attributes-value pairs may be used to specify a property. For example, the AVM given in (R C.6) is the attribute-value matrix defined in the syllable theory for acoustic autosegments:

R C.6

$$\begin{bmatrix} \textit{name} & \textit{the name of the property} \\ t_1 & \textit{a real number denoting the begin-time} \\ t_2 & \textit{a real number denoting the end-time} \\ \textit{score} & \textit{a real number indicating strength} \end{bmatrix}$$

In Chapter 3 a short hand notation of $< name,\ [t_1,\ t_2]\ score >$ has been used to represent the attribute-value matrix given in (R C.6).

C.2.3 Functional Descriptions and paths

All of the principles that apply to attribute-value matrices in general, apply specifically to functional descriptions. The functional description given in (R C.7) will be used as an example:

R C.7

$$\begin{bmatrix} \textit{constituent} & A \\ \textit{conditions} & \begin{bmatrix} \alpha & \begin{bmatrix} f_1 & v_1 \\ \vdots & \vdots \end{bmatrix} \\ \beta & \cdots \end{bmatrix} \\ \textit{realization} & \begin{bmatrix} \alpha & \begin{bmatrix} a & < a\ [t_b,\ t_e]\ s > \\ \vdots & \vdots \end{bmatrix} \\ \beta & \cdots \end{bmatrix} \\ \textit{score} & \cdots \end{bmatrix}$$

The functional description given in (R C.7) is a hypothetical structure corresponding to constituent A which is composed of subconstituent α and β. The constituent α has a feature specification that includes the value v_1 for the feature f_1. Further, the realization of α contains the acoustic property $<a\ [t_b, t_e],\ s>$, where the property $<a>$ is an autosegment having endpoints enclosed in the square brackets.

For the example functional description given in (R C.7) the value that corresponds to the path

$$(\textit{conditions},\ \alpha,\ f_1),$$

for this fd is v_1. A path of

$$(\mathit{conditions},\ \alpha),$$

is the feature-value pair $[f_1\ v_1]$. Paths refer to values on the acoustic property in a similar way. For instance, the path $(\mathit{conditions},\ \alpha,\ a_1,\ t_1)$ would retrieve the value t_b.

C.3 Context-free Grammar

Since it is to express hierarchical relationships among grammatical categories, the constraint description language incorporates the capability for stating a context-free grammar. Formally, a context-free grammar is a 4-tuple, $G = (N, \Sigma, P, U)$. The attribute-value matrix given in (R C.8) defines each of the grammar's elements:

R C.8

$$\begin{bmatrix} N & \textit{a set of non-terminal categories} \\ \Sigma & \textit{a set of terminal categories} \\ P & \textit{a the set of productions} \\ U & \textit{the grammar's distinguished symbol} \end{bmatrix}$$

The element P of a context-free grammar's definition is comprised of *productions* of the form $A \longrightarrow \alpha$, where, by the definition of context-free grammars, A is a nonterminal symbol (i.e., $A \in N$), and nominally, the string α, is a concatenation of terminals and non-terminals (i.e., $\alpha \in (N \cup \Sigma)^*$). Further, the distinguished symbol U, where $U \in N$, stands for *utterance.*

For the sake of conciseness, strings on the right hand sides of productions are allowed to to be *regular expressions*, consisting of elements of the set $(N \cup \Sigma)$. In particular, the symbols (), *, and +, are allowed on the right hand sides to indicate *optionality*, *closure* and *positive closure,* respectively.

C.4 Filters

In general, filters are used to state constraints on an utterance's surface phoneme specification. As such, they specify restrictions on an fd's *conditions* value. The general filter description is given as (R C.9):

R C.9 (General Filter Description)

$$\begin{bmatrix} \text{constituent} & \textit{a non-terminal} \\ \text{antecedent} & \textit{a feature specification} \\ \text{consequent} & \textit{a feature specification} \\ \text{polarity} & \textit{+ or –} \end{bmatrix}$$

As indicated by (R C.9), filters are *logical implications* implemented as attribute-value structures. A filter will contain an *antecendant*, a *consequent*, and a *polarity* attribute as well as a *constituent* attribute that is specified by the name of a category in the grammar. Antecedents and consequents are *feature specifications*, where a feature specification is a conjunctive set of conditions that when applied to an individual functional description, the fd must satisfy. It is comprised of a set of path-feature matrix pairs, $([p_1\ \vec{F_1}], [p_2\ \vec{F_2}] \cdots [p_k\ \vec{F_k}])$, where the paths, p_i, have the definition given above. Paths enable a filter to enforce dependencies that are not necessarily adjacent. A path's corresponding feature matrix, $\vec{F_i} = [\vec{f_1}\vec{f_2} \cdots \vec{f_l}]$, is a sequence of feature vectors, where each vector, $\vec{f_i} = [f_1 v_1], [f_2 v_2] \cdots [f_m v_m]$, is composed of a list of features. Features may have constant specifications (i.e., + or -), or they may assume variable values.

Owing to their status as logical implications, filters are to satisfy conditions as specified by *truth tables.* The truth table for filters with positive polarity is given in (R C.10). The truth table for negative filters is given in (R C.11).

R C.10 (Truth Table for a Positive Filter)

Satisfies antecedent?	Satisfies consequent?	Satisfies Filter?
Yes	*Yes*	*Yes*
Yes	*No*	*No*
No	*Yes*	*Yes*
No	*No*	*Yes*

R C.11 (Truth Table for a Negative Filter)

Satisfies antecedent?	Satisfies consequent?	Satisfies Filter?
Yes	*Yes*	*No*
Yes	*No*	*Yes*
No	*Yes*	*Yes*
No	*No*	*Yes*

C.5 Realization Rules

The final grammatical construct is the realization rule; its general definition is given in (R C.12):

R C.12

constituent	*a non-terminal*
input conditions	*(a feature specification)*
output conditions	*(a feature specification)*
input realization	*(an expression)*
output realization	*(name of an acoustic property)*
measurement	*a measurement specification*
assignments	*a list of expressions*

In this general description, the use of parenthesis denotes *values* that are optional. If a realization rule is not specified for the corresponding attribute, the value is null (i.e., $\emptyset$).

In several respects, realization rules are like filters. For example, a realization rule makes reference to an individual category of the grammar. In addition, analogous to the *antecendant* attribute of a filter, a realization rule has an *input conditions* value that is comprised of a feature specification. The feature specification, defined in the identical way as the value for a filter's *antecedent* specifies a set of constraints that apply to a functional description's *conditions* value.

Unlike filters, realization rules make reference to an fd's *realization* value as well as its *conditions.* In particular, the *input realization* value for a realization rule is a logical expression, that specifies a set of conditions that the value of an fd's *realization* must satisfy if the realization rule is to apply. The *input realization* value generally consists of a boolean expression comprising terms that contain relational functions.

Since both a realization rule's *input conditions* and *input realization* values are optional, the grammar writer has the option of leaving both unspecified. If this is the case, the realization rules would apply to all fd's of the designated grammatical type.

A realization rule's *output conditions* and *output realization* are also optional. When present they contain information that is to be added to the input FD's *conditions* and *realization* values respectively. The *output conditions* is a feature specification. During a realization rule's application, the output FD's *conditions* value will contain this information where perhaps the input FD did not. The *output realization* value is either a symbol or a list of symbols. These symbols, along with a set of accompanying values will become part of the output fd's *realization.* The output realization values will be computed by the function that is specified as the realization rule's *measurement* value.

Finally, a realization rule will contain a somewhat arbitrary list of assignment statements. Assignment statements are general expressions that can be used for a variety of purposes. These expressions could be used, for example, for specifying how scores are created and propagated during parsing. They can also be used to pass *special* instructions to the parser in order to direct the analysis of an utterance. Some examples will be included below.

Appendix D

Phonological Parsing

The principal objective of this thesis has been to develop a framework for expressing the relationship between the acoustic representation of an utterance and its surface-phonemic form. In Chapter 3, we developed a descriptive framework for stating this relationship. On the basis of this form of description, we have formulated a theory of syllabic organization over an utterance's acoustic properties and distinctive features.

In the present chapter, we outline one possible scheme whereby the proposed descriptive framework and syllable theory may be implemented in the form of a parser. Parsing, in general, is the problem of assigning a structure to an utterance that satisfies a set of well-formedness conditions specified by a grammar. In the present case, the structure to be assigned to an utterance is its syllable structure; well-formedness conditions describe the syllable's internal organization, state restrictions on phoneme sequences within the syllable, and describe the acoustic realization of distinctive features.

This chapter is organized into two major sections. In the first section, Section D.1, parsing is formulated as a *constraint propagation problem* (Winston, 1984). That is, constraints are viewed as *partial* statements of syllable structure well-formedness. Each applies over a relatively local domain in an utterance and states conditions that this region must satisfy. Through their hierarchical arrangement, local constraints are propagated into global conditions of syllable structure well-formedness. In Section D.1 we present a conceptual model that describes this propagation process. The hierarchical model for clustering constraints readily lends itself to implementation in the form of an augmented context-free parser. Section D.2 provides an overview of a chart parsing algorithm (cf., Martin *et al.*, 1987) that we have adopted and modified.

D.1 Conceptual Framework

The acoustic and phonological descriptions of an utterance each consist of a set of primitives. The acoustic description is comprised of *acoustic properties.* The phonological description consists of *distinctive features.* Grammatical constraints link these two domains of representation. This chapter addresses the problem how to apply constraints while translating the acoustic description of an utterance into its underlying phonological form.

For grammatical constraints to apply, it has been suggested that the primitives of an utterance's acoustic-phonological description be structured in an appropriate way. One such structuring, corresponding to the phonological representation advanced in *SPE*, is to have features "bundle" into column vectors, and for columns to concatenate to form matrices. Further, the literal interpretation of the theory outlined in *SPE* is for acoustic properties to similarly align into columns representing the *time-synchronous* implementation of phonological features (see Chapter 1). Grammatical constraints in the standard theory are expressed as transformational rules, where the rules describe a phonological derivation in which a linear string of phonetic segments are surface images of an underlying phoneme string that has been re-expressed by a process that inserts, deletes, and replaces symbols of one domain of representation with symbols of another.

Although relatively straightforward to apply in the "forward direction" to derive of an utterance's acoustic properties, inverting transformational rules during phonological parsing has proven infeasible (Church, 1983; Woods *et al.*, 1976). As an alternative, a computationally tractable solution to the parsing problem has been *lexical expansion.* Starting with a lexicon of phonemic "base-forms", transformational rules are applied in expanding each entry into all conceivable pronunciations. The resulting expanded lexicon, typically represented as a finite-state grammar or network, describes a language that is considerably easier to parse.

Zue (1983) points to three disadvantages to the lexical expansion approach:

> ... dictionary expansion does not capture the nature of phonetic variability, namely that certain segments of a word are highly variant while others are relatively invariant. It is also difficult to assign a likelihood measure to each of the pronunciations. Finally, storing all alternate pronunciations is computationally expensive, since the size of the lexicon can increase substantially (p. 185).

The limitations of lexical expansion cited by Zue add to those already suggested in Chapter 5. Probing the lexicon is a computationally expensive process. Applying rules to expand it only exacerbates the problem.

As an alternative to lexicon expansion, we may return to the approach of applying phonological constraints during the "on-line" decoding of an utterance. However, constraints must be formulated in a more restricted form. Specifically, the approach advocated here is to augment a context-free grammar. During parsing acoustic properties and distinctive features may be *parsed* or assigned to hierarchical structures such as the syllable and its internal constituents. Grammatical constraints then apply to these hierarchically structured objects. This latter proposal has been adopted by investigators such as Church (1983) and more recently Allerhand (1986). As will be seen below, hierarchical structures may be constructed at relatively small cost without increasing the size of the lexicon substantially.

Formalism

The current approach may be understood in terms of the scheme shown in (R D.1).

R D.1

$$\vec{A} \text{—}\boxed{\phi_C}\text{—} \vec{F}$$

The block diagram given in (R D.1) denotes a relation or "pairing" of an utterance's acoustic representation (denoted $\vec{A}$) with its surface-phonemic specification ($\vec{F}$). The pairing is through the relation $\phi_C(\vec{A};\vec{F})$. The relation ϕ_c formally denotes an utterance's functional description: a means of representation introduced in Chapter 3. The subscript C represents a set of constraints that $\phi_C(\vec{A};\vec{F})$ must satisfy if it is to be considered well-formed. For the remainder of the current discussion, the picture given in (R D.1) will serve as the conceptual basis for parsing.

D.1.1 The Hierarchical Clustering of Constraints

It is useful to consider the set of constraints as "clustering" around the nodes of the syllable template as suggested in Chapter 5. The context of that discussion was an study of collocational restrictions within the syllable. In particular, a hypothesis concerning the nature of the syllable's internal structure was tested using mutual information statistics gathered from a lexicon consisting of some 5500 syllables. Although statistical tests (based on a χ^2 statistic related to mutual information) did not entirely support hierarchical structure, cluster analysis of the data for the frequency unweighted lexicon did suggest the template shape proposed by Fudge (1969) and Selkirk (1982), and adopted in this thesis. Furthermore, hierarchically structured constraints have been motivated by other considerations advanced in Chapter 5. In particular, "principles of locality" and the notion of a "delayed binding of lexical decisions" are two considerations that would warrant a hierarchical structuring of syllabic constraints. These two principles served as the basis of heuristic arguments used to support the notion of structuring the acoustic input into syllabic and sub-syllabic constituents prior to accessing the lexicon. This argument was supported with experimental evidence that suggests that *parts* of the syllable serve as "islands" of phonetic reliability and lexical constraint. Specifically, in certain instances, only parts of the syllable need be specified in order to identify lexical entries. Furthermore, those parts tend to be more reliability represented in the sound stream.

The Syllabic Sketch

Prior to developing an explicit parsing algorithm based on the notion of hierarchically structured constraints, it is useful to review the steps involved in building the syllabic sketch of an utterance first presented in Chapter 3.

Constraints on overall syllable structure well-formedness include those that are to be satisfied while building the initial syllabic sketch and those that apply to the syllabic sketch after its initial form has been constructed. Both types of constraints apply to either terminal or non-terminal categories of the syllable's grammar. Building the initial syllabic sketch, for example, requires that acoustic autosegments be detected in the sound stream and related to manner-of-articulation features. Subsequently, manner features are assigned to the various positions of the syllable template. For the most apart, the positions in the syllable's hierarchical structure that are associated with manner-of-articulation features are terminal (e.g., OUTER-ONSET, INNER-ONSET, etc.). During the process of building the initial syllabic sketch, constraints on the temporal arrangement of these properties must be satisfied prior to making the initial syllable structure assignment. Once the initial syllabic sketch has been obtained, transitional measurements are extracted at places specified by realization rules in order for place-of-articulation and voicing features to be proposed. Finally, constraints on patterns of place and voicing features, represented as filters, are imposed.

D.1.2 A Constraint Propagation Model for Taking Advantage of Hierarchical Constraints

With a hierarchical structure being associated with a set of acoustic-phonological constraints, the parsing problem may be given numerous formulations. From the current perspective, parsing is viewed as a *constraint propagation* problem, where

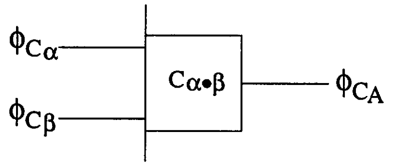


Figure D.1: This figure provides a schematic of the process of combining two functional descriptions while satisfying constraints on the constituent that they form.

constraint propagation is defined to be a procedure that "achieves *global consistency* [in an analysis] through [a series of] *local computations* (Winston, 1984; p. 45)."

A constraint propagation model is naturally suited for implementing conditions of well-formedness that are related to syllable structure. Consider, for example, the phrase structure rule given in (R D.2), taken from some immediate constituent grammar similar to that which describes the syllable:

R D.2

$$A \rightarrow \alpha\beta.$$

In (R D.2), the categories α and β are immediate constituents of A.

Let C_α and C_β denote the set of constraints that are imposed by the grammar on the constituents α and β. That is, the constraints comprising these two sets are filters and realization rules that apply to the *conditions* and *realization* values of the functional descriptions ϕ_{C_α} and ϕ_{C_β} (see Chapter 3). The process of building structure during parsing requires that the constituents α and β be combined to produce the constituent A. Put another way,

R D.3

$$A = \alpha \circ \beta.$$

In typical parsing implementations, α and β denote phrases (or strings), and the operator "o" indicates a procedure whereby the phrases are concatenated. However, in the present case, the constituents α and β are represented by their functional descriptions, and the procedure for combining constituents is one in which the two functional descriptions are *unified.*[1] The process of unification which is denoted by the expression given in (R D.4):

R D.4

$$\phi_{C_A} = \phi_{C_\alpha} \sqcup \phi_{C_\beta}.$$

Unification is a process in which the *conditions* and *realization* values corresponding to the two constituent functional descriptions, ϕ_{C_α} and ϕ_{C_β}, are combined to produce the *conditions* and *realization* values of the resultant functional description, ϕ_{C_A}. During the process of unifying ϕ_{C_α} and ϕ_{C_β}, the set of constraints on the resultant functional description, denoted $C_{\alpha o \beta}$ are to be satisfied. As a result, the set of constraints that the constituent A has satisfied during the implementation of the production rule given in (R D.2) is given by the expression in (R D.5):

R D.5

$$C_A = C_\alpha \cup C_\beta \cup C_{\alpha o \beta},$$

where the operator "$\cup$" denotes *set union.* In other words, a functional description corresponding to some constituent in the grammar is considered well-formed only if a) its immediate constituents are well-formed and b) the their combination is well-formed. In this way, well-formedness conditions on small constituents propagate into constraints on larger constituents. A block diagram is given in Figure D.1 that illustrates the satisfaction of a local set of constraints.

[1] Our use of the term *unification* is derived from a similar use in the theory of *Functional Unification Grammars* (Kay, 1986). One may also find this term used in describing a number of other grammatical formalisms (cf., Kaplan and Bresnan, 1982).

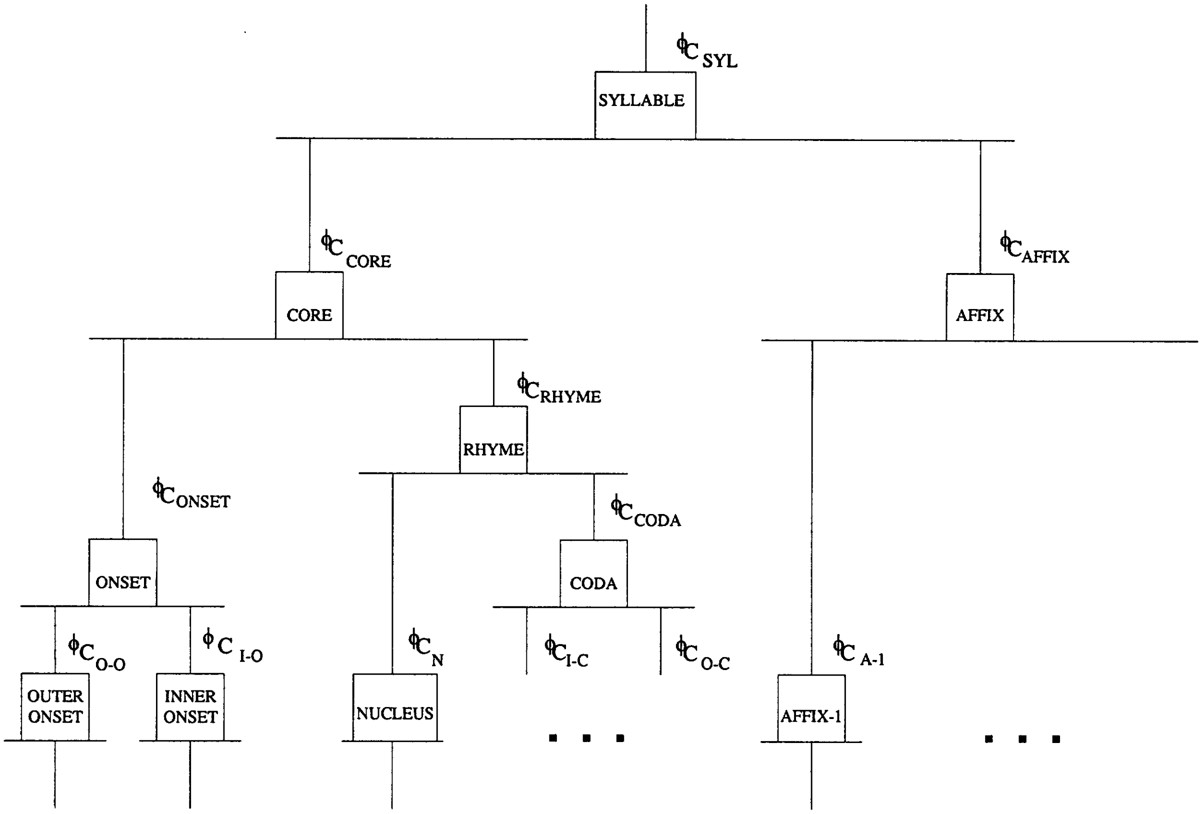


Figure D.2: Constraint propagation network schematically depicting the satisfaction of constraints while constructing the syllabic sketch for an utterance. The input to the network consists of the utterance's acoustic properties.

In general, the symbols A, α, and β denote any three symbols in the grammar, as long as A is a non-terminal. For example, A may denote the constituent Onset and α and β may denote the categories Outer-onset and Inner-onset respectively. In this case, local constraints on the Onset are satisfied in producing the functional description $\phi_{C_{ONSET}}$. Similarly, A may denote Core, and α and β, Onset and Rhyme respectively. In this situation, $\phi_{C_{CORE}}$ will be a functional description satisfying constraints on the Onset and the Rhyme individually, and the Core (i.e., Onset $\circ$ Rhyme). By induction, ϕ_{C_σ} is a functional description satisfying all of the constraints in the grammar.

A *constraint propagation network* corresponding to the above conceptual framework is given in Figure D.2. In the network, each of the nodes, represented as boxes, denotes a "packet" of constraints, each corresponding to the category for which the box is labelled.

D.2 Implementation

In the previous section, parsing is viewed as an iterative procedure, consisting of two basic steps, each applied during the process of building the internal structure of syllables (i.e., the syllabic sketch). In the first step, a functional description corresponding to a constituent in the syllable's grammar is formed by combining functional descriptions corresponding to its immediate constituents. Once this functional description is obtained, the second step is to apply constraints. Constraints consist of filters and realization rules, and are applied disjunctively. If a functional description is deemed well-formed according to each, it may then be combined into some larger structure, at which point the iteration is repeated. We referred to these two steps as unification.

In the present section, we sketch the algorithm we have used for implementing this parsing scheme. We extend the basic steps executed in the above iteration for constructing the internal structure of a single syllable, to a set of steps, executed in parallel, for constructing the internal structures of a *lattice* of syllables. A lattice is a general data structure used to record all possible syllabic analysis for an utterance. In the scheme we have implemented, lattices are represented as *charts.* The algorithm we will describe is a modified version of a very general context-free parsing procedure proposed by Martin *et al.* (1987), based on a set of procedures for manipulating charts.

The set of modifications we propose to Martin's chart parsing procedure extend a "core" of parsing operations for constructing the immediate constituent structure corresponding to a context-free grammar. Our modifications are two-fold. First, we must embed into the overall procedure the unification operation discussed above. Second, the input to parsing is not a simple string of symbols, as is more typically the case for parsing languages, rather, the proposed input is a multi-dimensional acoustic representation of an utterance in which acoustic properties (e.g., acoustic autosegments) are represented on individual tiers.

D.2.1 Charts and Chart Parsing

In general, a parser takes as input a grammar and a symbolic representation of an utterance. In most situations, the input is in the form of a concatenation of symbols (i.e., a string). As output, a parser produces a structural analysis of the input that may assume a variety of forms. For parsing context-free languages, for example, a useful structural analysis is in the form of a *parse tree* (i.e., a representation of an utterance's immediate constituent structure). Alternatively, the structural analysis may be displayed in terms of a *chart.*

Formally, a chart is a *directed graph*, consisting of a network of *vertices* and *edges.* Each vertex represents a juncture between input symbols, while edges are labelled with either symbols, phrases (i.e., strings of symbols) or in the case of the current implementation, functional descriptions. As means of illustration, we will consider assigning syllable structure to the phoneme string /ǰɛntli/. The analysis performed is based on the syllable grammar given in Chapter 3. Its relevant production rules are repeated in (R D.6).

R D.6

$$\begin{array}{rcl}
\mathrm{U} & \longrightarrow & \sigma^{+} \\
\sigma & \longrightarrow & \textsc{Core}\ (\textsc{Affix}) \\
\textsc{Core} & \longrightarrow & (\textsc{Onset})\ \textsc{Rhyme} \\
\textsc{Onset} & \longrightarrow & \textsc{Outer-Onset}\ |\ \textsc{Inner-Onset}\ |
\end{array}$$

Outer-Onset Inner-Onset

Rhyme ⟶ Nucleus (Coda)

Coda ⟶ Inner-Coda | Outer-Coda |
Inner-Coda Outer-Coda

Outer-onset ⟶ ǰ | t

Inner-onset ⟶ l |n

Inner-coda ⟶ n | l

Outer-coda ⟶ t |ǰ

Nucleus ⟶ ε | i

Initially, the parser constructs a chart containing only the phoneme string. In (R D.7) the chart is displayed in a *pseudo*-directed graph form.

R D.7

$${}_0\check{j}_1\varepsilon_2 n_3 t_4 l_5 i_6$$

A more convenient means of representing the chart (both visually, and from the standpoint of computation) is in the form of a two-dimensional table (or matrix). The cells of the matrix contain constituents that span a region in the utterance, where this region is indicated by the cell's position in the matrix. Specifically, the cell's *row* represents the constituent's starting position and its *column* denotes the position where it ends.

The initial chart configuration for the example phoneme string /ǰεentli/ displayed in (R D.7) above, is displayed as a matrix in (R D.8).

R D.8

$$\begin{array}{ccccccc}
0 & \{j\} & & & & & \\
1 & & \{\varepsilon\} & & & & \\
2 & & & \{n\} & & & \\
3 & & & & \{t\} & & \\
4 & & & & & \{l\} & \\
5 & & & & & & \{i\} \\
 & 1 & 2 & 3 & 4 & 5 & 6
\end{array}$$

During parsing, edges are entered into the chart as the parser finds productions in the grammar whose right-hand sides *match* edges already present. The matching procedure is specified as follows. For some production, $A \longrightarrow \alpha\beta$, in the grammar, where A is a nonterminal category, and α and β are either terminals or non-terminals, a new edge with label A is entered into the chart starting at vertex i and ending at vertex j, if for some k, there are edges already present that are labelled α and β spanning vertices $< i, k >$ and $< k, j >$ respectively. The following expression describes this basic chart operation.

R D.9

$$chart(i,j) := \bigcup_{i<k<j} \{A \mid \alpha = chart(i,k) \wedge \beta = chart(k,j)\}.$$

Parsing proceeds by applying the operation given in (R D.9) iteratively until the symbol, *U,* (i.e., the grammar's distinguished symbol) is contained in the chart's upper-right hand corner. The following is a summary of the complete algorithm presented in a "pascal-like" programming language.

R D.10 (Basic Chart Parsing Algorithm)

```
for j := 1 to n do
    $chart(j-1, j) := \{A \mid A \rightarrow phoneme_j\}$
    for i := $j-2$ downto 0 do
        $chart(i,j) := \bigcup_{i<k<j}\{A \mid \alpha = chart(i,k) \wedge \beta = chart(k,j)\}$
if S is in $chart(0, n)$ then accept
else reject
```

The chart that results from applying algorithm (R D.10) to the example phoneme string using the grammar listed in (R D.6) is given in Table D.1. The reader is invited to work his way through this example.

We should note that the grammar given in (R D.6) does not impose phonotactic restrictions. As a consequence, there is more than one syllable structure that may be assigned to this phoneme sequence. Specifically, the /t/ in this word is assigned to both the OUTER-CODA of the first syllable and the OUTER-ONSET of the second. This is indicated in Table D.1, where a ONSET constituent has been placed in cell (3,5), a CORE in cell (3,6), and a σ in cell (3,6).

One may extend the grammar listed in (R D.6) to impose phonotactic restrictions, either by adding additional production rules, or augmenting it with filters. The latter alternative is the approach that we have adopted. When applied to this chart, the filters would eliminate extraneous analysis.[2]

D.2.1.1 Matrix Multiplication

The procedure summarized in (R D.10) is essentially the core of the parsing procedure; its task is to build hierarchical structure. Martin points to important similarities between it and the algorithm for matrix multiplication. In particular, in a comparison between the expression for the general chart entry in chart parsing,

$$\mathit{chart}(i,j) := \bigcup_{i<k<j} \{A \mid \alpha = \mathit{chart}(i,k) \wedge \beta = \mathit{chart}(k,j)\}$$

and the expression for computing the general entry in the matrix C, where $C = A \times B$,

$$c_{ij} = \sum_k a_{ik} * b_{kj},$$

one finds that the two are structurally similar.

These results have led Martin to propose a parsing algorithm based on *factoring* the original chart into individual charts, each corresponding to a particular category in

[2] We should point out that all analysis have not been included in this chart. For the sake of brevity, we have included only those pertinent to the current discussion.

the grammar. For the grammar fragment given in (R D.6), for example, the factoring would be as follows:

R D.11

$$M = M_\sigma + M_{\text{CORE}} + M_{\text{ONSET}} + \cdots .$$

The expression given in (R D.11) represents an operation whereby the original chart M is obtained from each of the individual charts through a "cell by cell" union (denoted "+") of their elements. Each individual term is itself obtained by the parsing analog of matrix multiplication; for instance,

R D.12

$$M_{\text{CORE}} = M_{\text{ONSET}} \circ M_{\text{RHYME}},$$

where the operator "o" in (R D.12) denotes a process whereby the elements of the cells of the individual matrices are concatenated.

D.2.2 Modifications to the Basic Chart Parsing Algorithm

The algorithm outlined thus far is not entirely appropriate for the parsing problem that needed to be solved in the current investigation. For example, the input that is envisioned to the required parser is not a simple string of symbols, but rather acoustic properties represented on multiple tiers. Further, the edges of the chart are to be labelled with functional descriptions rather than categories denoting phrases. The latter problem has a straightforward solution, and therefore is addressed first. Afterwards, the first problem is considered along with other modifications to the basic chart parsing procedure.

D.2.2.1 Filling Charts with Functional Descriptions

Given that the general chart entry in the current implementation is a list of functional descriptions as opposed to a one-bit integer as is the case in Church's implementation, the parsing analog of matrix multiplication needs to incorporate an operation for unification. The expression given in (R D.9) for the production $A \longrightarrow \alpha\beta$ now becomes

R D.13

$$\phi_{C_A} < i,j >= \bigcup_{i<k<j} \phi_{C_\alpha} < i,k > \sqcup \phi_{C_\beta} < k,j >,$$

where, once again, while unifying two functional descriptions, filters (to impose phonotactic restrictions) and realization rules belonging to the set of constraints C_A are applied.

D.2.2.2 Combining Acoustic Autosegments

In order to combine acoustic properties represented on separate tiers, we have added an additional parsing operation which is to combine the elements of charts representing these properties. The proposed operation is similar to matrix multiplication, but with the modifications defined below.

Unlike what happens with two phrases in the string parsing problem, the combination of two autosegments does not necessarily imply they concatenate in time. Instead, their combination to form a constituent depends on their satisfying temporal constraints imposed by the grammar (see Section 3.2). Let $\phi_{a_i} < m,n >$ denote the functional description for some acoustic autosegment a_i, where a_i spans the positions $[m,n]$ in the waveform. The values m and n are integers denoting timepoints which have been digitized. Furthermore, let $\phi_{a_j} < m',n' >$ denote a functional description defined similarly for autosegment a_j, spanning $[m',n']$. In order for a_i and a_j to combine, forming some constituent B in the grammar, there will be timing constraints, stated as a set of inequalities, involving the begin- and end-positions of these acoustic autosegments. In addition, these realization constraints will designate where in a chart computed for constituent B the functional description ϕ_{C_B} is to be placed. We will denote this position $[m'', n'']$. Therefore, we may write an expression for combining ϕ_{a_i} and ϕ_{a_j} as

R D.14

$$\phi_{C_B} < m'', n'' >= \phi_{a_i} < m, n > \sqcup \phi_{a_j} < m', n' > .$$

Finally, we may write the expression for computing the general chart entry for constituent B by combining charts corresponding to a_i and a_j as

R D.15

$$\phi_{C_B} < m'', n'' >= \bigcup_{m,n} \bigcup_{m',n'} \phi_{a_i} < m, n > \sqcup \phi_{a_j} < m', n' > .$$

In other words, in the charts corresponding to a_i and a_j there will be several pairs of autosegments that, when combined, satisfy the timing constraints imposed for constituent B. The operation defined in (R D.15) computes all such combinations.

D.2.2.3 Transitive Closure vs. Optimal Search

The final parsing operation we discuss is the one for implementing the production rule,

R D.16

$$\mathrm{U} \longrightarrow \sigma^{+}.$$

Rule (R D.16) may be read "an utterance is comprised of one or more syllables". In order to implement this rule, a procedure is required for computing the *transitive closure* of the σ chart (or graph). There are a number of general purpose algorithms for implementing the transitive closure of a directed graph. A somewhat inefficient, but nonetheless useful technique is based on Floyd's algorithm which accepts a graph stored in the form of a two-dimensional table edges (see, for example, Aho *et al.* (1974) for its details).

Optimal Search

An alternative way of viewing the rule stated in (R D.16) is to think of the constituent U as comprised of the *most likely* syllable structure to be assigned to an utterance, where the likelihood criteria is to defined in a way that it is meaningful in the phonological sense, and computationally tractable to implement. Consider, for example, the phoneme sequences belonging to the members of the minimal pair *grey train* and *great rain.* Although they are identical, the /t/ in *grey train* is heavily aspirated, while the /t/ in *great rain*, may be released with a short VOT, unreleased, or glottalized (see Chapter 4). From the standpoint of parsing, an utterance consisting of either of these two word sequences could be assigned more than one syllable structure, although on the basis of the stop VOT, for example, one syllabic analysis would be preferred.

An idea that we began to explore towards the end of our investigation was the notion of using probabilities to assign to individual syllabic analysis. These probabilities were derived from a number of sources, including regression trees. Then, as an alternative to computing the full transitive closure of a chart, an optimal search was used. Any number of search procedures are appropriate for this purpose; in our implementation, the *branch and bound* algorithm (Winston 1984) is used.

D.3 Summary

In summary a set of procedures for implementing the proposed syllable theory in the form of a parser have been proposed. Conceptually, parsing has been formulated as constraint propagation problem, although a very general chart parsing algorithm has been modified for the actual implementation. The algorithm itself is based on an algorithm proposed by Martin. The modifications proposed include adding the capability for unifying functional descriptions and manipulating an input that consists of acoustic properties represented on separate tiers.

Our purpose in developing this implementation has been two-fold. First, we wanted a computational framework for debugging the descriptive framework and grammar proposed in Chapter 3. Secondly, we wanted to demonstrate the feasibility of a acoustic-phonetic representation for speech recognition that did not require the signal to be segmented and labelled in terms of allophonic symbols.

Table D.1: Final contents of the chart for the word *gently*.

	1	2	3	4	5	6
0	{j, Outer-Onset, Onset}			{Core, σ }		{U }
1		{E, Nucleus }		{Rhyme }		
2			{n, Inner-Coda }	{ Coda }		
3				{ t, Outer-Coda }	{Onset}	
4					{ l, Inner-Onset, Onset }	{Core, σ }
5						{i, Nucleus, Rhyme }