Working Papers of the Cornell Phonetics Laboratory
1996, vol. 11, pp. 43-116

An Integrated Model of Phonetic Representation in Grammar

G. N. Clements
CNRS (Paris) and University of Paris-3

Susan R. Hertz
Eloquent Technology, Inc. and Cornell University

This study argues that linguistically-determined aspects of phonetics form part of grammatical theory in much the same sense as phonology or syntax do, and can be modeled in terms of similar principles. It proposes that the phonetic component of grammar contains sets of phonetic representations similar to the partially specified, multi-tiered representations of the phonological component. At the acoustic level, these representations include a set of autosegmental tiers specifying values for acoustic parameters such as voicing, nasality, vowel formants, F_0, etc. as well as a duration tier which organizes these values into a succession of discrete acoustic events in the time domain. This integrated representational system, or IRS, defines a complete interpretation of surface phonological representations at the acoustic level, and can be specified with sufficient detail to provide input to an acoustic speech synthesizer.

This framework is applied to a study of the temporal properties of long vowels and diphthongs in General American English. We consider the well-known question whether the long vocalic nuclei of words like *bait*, *boat*, *bite*, and *bout* should be analyzed as one phonological segment or two. Durational evidence involving asymmetries in the distribution of contextually-determined lengthening across the syllable suggests that the nucleus of *bait* is best modeled as a single melodic unit (root node) occupying two skeletal positions, while that of *bite* is best modeled as two melodic units. The phonetically diphthongized quality of the nuclei of words like *bait* and *boat* is determined at the phonetic level by a constraint requiring them to have separate formant target positions at their left and right edges; as a result, no general long vowel diphthongization rule is required in the phonology.

1 Introduction

A major goal of linguistic theory is to account for the ability of the members of a speech community to produce and perceive speech in terms of regular phonetic patterns whose characteristics are in part specific to each language and dialect. However, it is striking fact that current linguistic theory, with some exceptions, does not directly address this problem. In spite of a recent revival of interest in the nature of the phonology/phonetics interface, most phonologists continue to disregard the question of how their descriptions might be interpreted in the physical domain, while most phoneticians and speech scientists show a similar disregard for how phonetic data can be related to the abstract structures posited by linguists. Many research paradigms still treat phonetics as unrelated to phonology, in practice if not necessarily in principle. Yet speech is the physical and behavioral manifestation of cognitively-represented linguistic systems, and cannot be fully understood without reference to the linguistic structure which underlies it. While this point has never been a matter of controversy or serious debate, relatively few researchers in recent times have directly addressed the goal of determining the nature of the relations between phonological representations as studied by linguists and phonetic realizations as examined by phoneticians (some important exceptions will be discussed below). As a result, our knowledge of how phonology and phonetics map into each other is still in a relatively primitive state.

There is very good reason to believe that important aspects of phonetics belong to grammar in much the same sense that phonology does. It is well known that phonetic patterns can differ systematically from one language or dialect to another just as phonological patterns do. Part of what any individual must master in acquiring a language is its particular pattern of phonetic realization, including quantitative, subphonemic detail. Such patterns are clearly internalized by native speakers, since they are extended productively and exceptionlessly to forms never heard before (nonce formations, loanwords, nonsense words, products of speech errors) and are typically generalized to foreign words in second language acquisition. Since phonetic differences among languages are obviously not determined by physiological differences among their speakers, they must be attributed to differences in the phonetic systems that speakers have acquired, and more specifically, to differences in the principles (rules or constraints) that determine the form and content of phonetic representations. Examples of such differences include the fact that [t] normally involves dental contact in French but alveolar contact in English, or the fact that English vowels are lengthened by 50% or more before voiced consonants in certain contexts where Korean vowels are lengthened by about 30%. Many similar examples of language- and dialect-particular differences in the phonetic realization of speech sounds are documented in the literature (see, for example, Ladefoged 1967, Wood 1979, Lindau 1984, Keating 1985, Fourakis and Port 1986, and Labov 1986, among many others); phonetic differences are also found among individuals and among different age and gender groups (see e.g. Zue and Laferriere 1979). Thus language-particular phonetic principles, like those of other grammatical components, form part of each speaker's tacit knowledge of his or her language.

To this view it is sometimes objected that grammatical systems deal exclusively with *categories*—units and their relations— and cannot, therefore, involve quantitative information of the sort that is required to assign speech sounds specific values along continuous phonetic parameters. In this view, quantitative regularities are best relegated to such realms as performance or "paralanguage". However, it is often hard to draw a sharp line between the continuous and the noncontinuous in the characterization of grammatical knowledge (see Ladd 1993, Pierrehumbert, Beckman and Ladd (in press) for relevant discussion); moreover, as we will shown below, many aspects of phonetic representation can be regarded as categorical, and can be modeled in terms of constraints on symbolic representations of much the same type as are commonly employed in phonology. It is also sometimes objected that since much subphonemic phonetic detail goes largely undetected by speakers due to the phonemic bias in speech perception, it should not be assigned to grammar, which represents a model of the speaker's internalized (and thus, largely phonemic) linguistic system. However, many studies have shown that distinctive features can be perceptually cued by subphonemic acoustic properties, some of which, like vowel lengthening (an important cue to [+voice]), are at least in part language-dependent (see e.g. Hillenbrand et al. 1984). Furthermore, we must assume that speakers reliably perceive and encode subphonemic regularities in their day-to-day language experience if we are to explain their ability to acquire language-dependent phonetic rules. As Pierrehumbert points out (1994), "even the most inescapably quantitative details of language sound structure are subject to language particular conventions, and hence must be learned and represented in the mind. That is, the cognitive representation of language is not confined to categorical structure and rules, but rather includes arbitrarily fine details of allophony."

We take the position that it is not only possible to integrate phonological and phonetic representation into a single grammatical system, but that such an integration is necessary if we are to provide a unified account of sound structure, subject to a uniform evaluation measure. The idea of such a fusion is not in itself new. It can be found, for example, in Chomsky and Halle's proposal (1968) that phonological and phonetic representations both have the form of fully-specified, two-dimensional feature matrices, differing in that feature specification is binary at the underlying (phonological) level and integer-valued at the surface (phonetic) level. However, subsequent developments in both phonology and phonetics have suggested a somewhat different view of phonological and phonetic form. Phonological representations are no longer assumed to be fully-specified nor to have the form of two-dimensional matrices, but are viewed as partly-specified, hierarchically-organized three-dimensional structures involving many tiers or information channels, each of which corresponds to an independent phonological parameter and whose units may overlap in potentially complex ways (see e.g. Kenstowicz 1994, Goldsmith 1995 for recent overviews). This general conception—which we will term a *partly-specified, multitiered* approach—finds a counterpart in phonetic models such as those proposed by Pierrehumbert 1980, Hertz, Kadin, and Karplus 1985, Browman and Goldstein 1986, Keating 1988, Cohn 1990, and Keyser and Stevens 1994, among others. As most of this work has shown, once such an enriched conception is adopted there is no longer any need to introduce integer-valued phonetic features at the phonological level of description.

The present study builds on these fundamental insights. It proposes that the partially-specified, multi-tiered approach developed in recent phonological theory can be insightfully generalized to phonetics by introducing a level of *acoustic phonetic representation* in which appropriate acoustic and durational values are formally related to the nodes of the surface phonological representation that they interpret. Such representations, or "acoustic scores" to use the felicitous term of Perkell 1980, array phonetic information in much the same way that phonological features are arrayed at the phonological level, thereby eliminating the need for a radical translation between distinct and largely incompatible representational systems.

The structure of this paper is as follows. Section 2 discusses some of the theoretical antecedents to our work. Section 3 presents an overview of our integrated representational system (IRS), discussing first its phonological aspects (section 3.1) and then its acoustic aspects (section 3.2). Section 4 addresses the phonological analysis of English vocalic nuclei, focusing on the question of whether long vocalic nuclei should be represented as one phonological segment or two. Section 5 considers the same question from a phonetic point of view, examining durational properties of representative vocalic nuclei, and offering evidence that some long nuclei are diphthongized only in the phonetics. Section 6 considers some further issues raised by our approach, and section 7 summarizes our major proposals and results.[1]

2 The relationship between phonology and phonetics: recent models

This section discusses some recent attempts to define the relationship between phonology and phonetics, emphasizing those that have contributed the most to our own thinking. We review in succession 1) an early autosegmental view, 2) the articulatory phonology of Browman and Goldstein, 3) previous work in speech synthesis by rule, and 4) the target-and-interpolation model of phonetic interpretation.

Phonological theory took a new direction in the 1970s and 1980s with the wide acceptance of nonlinear models of representation. These models reopened the question of how phonological representations can be related to phonetic interpretation. Some first thoughts on this subject were offered by John Goldsmith in his influential 1976 thesis on autoseg-

[1] This paper is aimed at an intended readership of both linguists and phoneticians. Each type of reader will find certain sections of this paper elementary, and we apologize in advance for any belaboring of what may seem to be obvious points.

mental phonology. Although this work was mostly concerned with phonology in the strict sense, Goldsmith also speculated that an adequate phonological model might allow a better integration of phonology and phonetics. He suggested that "autosegmental phonology is a theory of how the various components of the articulatory apparatus—the tongue, the lips, the larynx, the velum—are coordinated" (p. 29). Goldsmith further proposed that while the representations of consonant and vowel segments are mostly linear at the level of underlying phonological representation, i.e. arrayed on very few tiers, their features become "autosegmentalized" in the course of a phonological derivation through rules assigning them to new tiers of their own.

At the surface (or phonetic) level, according to Goldsmith, "the speech signal is broken down into a large number of independent linear parts—autosegmental tiers—with at least as many of these tiers as there are independent articulators. Thus there will be minimally such a tier for the velum, for the laryngeal gesture corresponding to pitch, and so forth" (p. 264). He gives the following schematic example of a phonetic representation, representing the word *pin* spoken with falling intonation:

(1)	Lips	. . . Close up . . . Open . . . .
	Tongue	. . . High and front . . . . touch the palate
	Velum	. . Raise . . . . Lower . . . .
	Larynx	. . . High Pitch . . . . . Low Pitch . . . .

In this diagram, vertical alignment is intended to suggest the pattern of temporal coordination among the articulators. Thus the vertical alignment of lip closure, tongue raising and fronting, velum raising, and high pitch at the beginning of the word indicate that the first segment is an oral labial stop such as [p] coarticulated with a high front tongue position and a toneme of high pitch.

Goldsmith's suggestions were subsequently developed and extended to the articulatory phonetic level in a series of papers by Browman and Goldstein (see e.g. Browman and Goldstein 1986, 1989, 1990). These writers developed a notion of *gestural score* somewhat resembling structures like that in (1), but replacing informal characterizations such as "close up" or "touch the palate" with precise specifications of the amplitude and velocity of vocal tract constrictions, and replacing the (implied) time grid with specific patterns of intergestural temporal coordination. They further differ from Goldsmith in proposing that phonological and phonetic representations are formally identical, thereby eliminating the need for any radical translation between phonology and phonetics. Browman and

Goldstein show that many patterns of apparent assimilation and deletion in casual speech can be described quite naturally in terms of gestural overlap.

However, Browman and Goldstein's approach raises a number of problematical issues. Due to their hypothesis that phonological and phonetic representations are formally identical, their model fails to capture important differences between the nature of linguistic generalizations at these two levels. In particular, it cannot explain the widely-attested observation that segmental units behave in a categorical fashion at the phonological level, and require quantitative specification only at the phonetic level (Clements 1992). One straightforward way of remedying this problem would be to regard the model as a phonetic interpretation model, interfaced it with a standard autosegmental model of the phonology (Zsiga 1993); another would be to suppress all quantitative aspects of gestural organization at the phonological level, while conserving the categorical aspects.

In addition, from a phonetic point of view, their model emphasizes the role of articulatory movements to the virtual exclusion of acoustic and perceptual factors in speech production. Yet much evidence suggests that articulatory organization is oriented toward the goal of achieving relatively stable acoustic outputs with optimal perceptual properties. For example, it is commonly observed that sets of physically similar articulations which have similar acoustic and perceptual effects tend to represent the same phonemic category in any given language, while similar articulations that lie on opposite sides of acoustic or perceptual "quantal" boundaries commonly represent phonemically distinct categories (Stevens 1972, 1989). This fact suggests that languages tend to define their phoneme systems in terms of articulatory configurations that maximize perceptual distinctiveness among phonemes. Furthermore, while Browman and Goldstein offer support for their model primarily from casual speech phenomena which sometimes eliminate acoustic contrasts, more carefully-monitored speech styles tend to coordinate independent articulatory events in such as way as to create salient acoustic "landmarks" at the boundaries between segments (see e.g. Ohala and Kawasaki 1984, Huffman 1990, Halle and Stevens 1991, Stevens 1994); this "boundary-flagging" effect can be understood as a means by which speakers facilitate the segmentation of the signal by listeners. Another point is that many aspects of articulatory coordination appear to be motivated by the requirement that distinctive features should be audible. For example, the feature [spread glottis] in stops is typically coordinated with the stop release, where it can be heard as aspiration overlaying the transition to a following vowel; were it timed instead to coincide with closure it would be inaudible (Kingston 1990, Davis 1994). Further, the existence of redundant articulations, such as the lip-rounding that is typically coordinated with back vowels, can be understood in terms of the principle of acoustic enhancement (Stevens, Keyser, and Kawasaki 1986).

Especially strong evidence for a central role of acoustics in speech production comes from the phenomenon of compensatory articulation, brought to light in recent studies by Maeda (1990, 1991). Maeda has found that speakers of French tend to compensate for deviations in the position of one articulator (such as the jaw) by adjusting the movements of another articulator (such as the tongue body) in such as way as to minimize acoustic variability in the output. These findings confirm and extend earlier results on compensatory articulation from bite-block experiments (e.g. Lubker 1979). Similarly, different speakers may use any of a number of acoustically similar articulatory configurations for producing the same sound. For example, different speakers of English produce / r / by various combinations of lip rounding, tongue blade raising, and pharyngeal constriction, all of which lower the third formant (Pierrehumbert 1994). The fact that these disparate articulations are treated as functionally equivalent can only be explained on acoustic grounds. For these reasons (and others), the phonetic model must provide for the inclusion of appropriate acoustic information.

One way of providing acoustic information, and the one we will adopt, is to provide an explicit acoustic level of representation in the phonetic component. The approach we will make use of draws upon recent work in the area of speech synthesis. Beyond its many practical applications, speech synthesis has proven increasingly valuable as a tool in the explicit modeling of both articulatory organization (e.g. Mermelstein 1973, Coker 1976, Maeda 1990) and acoustic structure (e.g. Hertz 1982, Klatt 1987, Allen et al. 1987). The two main approaches to speech synthesis (using unlimited vocabulary) are concatenative synthesis and rule-based synthesis. In concatenative synthesis (e.g. Peterson, Wang and Sivertson 1958, Dixon and Maxey 1968, Fujimura and Lovins 1978, Olive 1990), acoustic values for certain speech fragments such as diphones and syllables are extracted from natural speech and pieced together to construct utterances. Concatenative synthesis has some practical advantages, but is limited in its usefulness for theoretical modeling since, by its nature, it cannot detect many linguistic generalizations holding internally to (or cross-cutting) the units chosen for concatenation. In rule-based synthesis (e.g. Hertz 1982, Klatt 1987, Allen et al. 1987), in contrast, phonetic parameter values are generated by means of rules which can express generalizations holding all the way down to phoneme-sized segments and their internal constituents. It can therefore be used to formalize predictive hypotheses concerning acoustic regularities and generalizations at both the supra- and sub-segmental levels, and to test these hypotheses through an evaluation of the naturalness of the speech output they predict. Rule-based synthesis can also be used to test detailed hypotheses regarding the relation between phonetic structure and more abstract phonological representations. Our own work in this direction has been carried out in conjunction with a system of rule-based synthesis that makes use of multi-tiered phonological and phonetic representations formulated in terms of the Delta System (Hertz, Kadin, and Karplus 1985, Hertz 1988, 1990a, 1990b, 1991, Hertz and Huffman 1992). Delta-based rule sets generate acoustic parameter values which can serve as input for a formant-based synthesizer (e.g. that described by Klatt 1980 or Klatt and Klatt 1990).

Speech synthesis rules typically make use of what are sometimes called *target-and-interpolation* models of phonetic interpretation (e.g. Kelly and Gerstman 1961, Holmes, Mattingly, and Shearme 1964, Hertz 1979). These models simulate the continuously-varying time course of given acoustic parameters by extracting certain critical values (typically coinciding with turns or "inflection points" in the observed time course) and interpolating continuous values between them by appropriate algorithms. Target-and-interpolation models have been incorporated in linguistic theories of phonetic realization in such domains as fundamental frequency (Pierrehumbert 1980, Pierrehumbert and Beckman 1988), aspiration (Keating 1988), nasality (Huffman 1990, Cohn 1993), and formant frequencies (Hertz 1991). We can illustrate a target-and-interpolation model with a hypothetical example suggesting the possible F0 interpretation of a High-Low-High tone melody associated with the vowels of a VCVCV sequence:

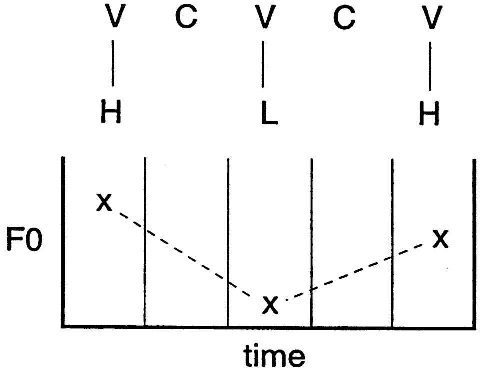


Figure 1. A target-and-interpolation model of the pitch interpretation of a HLH tone sequence.

In this example, critical F0 values (shown by the *x*s), selected on the basis of observed data, are assigned to the middle of each vowel; we may call these "target values". Intervening values, falling along the dashed lines, are computed by interpolation between adjacent target values; we call these "interpolated values". Further smoothing and/or adjustment can be carried out to round off corners and account for local segmental perturbations on the global F0 curve, according to our purposes. If values have been correctly chosen, the resulting F0 track will resemble the original F0 time course.[2]

In a phonetic theory incorporating a target-and-interpolation model, phonetic representations can abstract away from all transitional phenomena that are predictable in terms of abstract targets. This result allows for a considerable simplification of phonetic representations, and, as we shall see, lays a solid basis for extending the multi-tiered representational system of the phonology directly into the phonetics.

3 An integrated system of phonological and acoustic phonetic representation

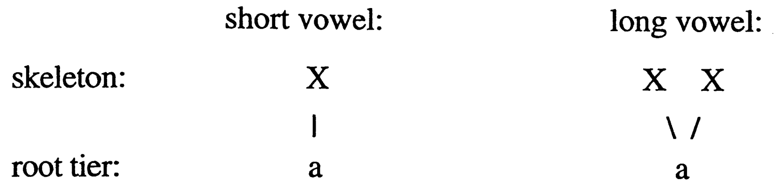
We shall now develop an integrated representational system, or IRS, for phonology and phonetics, which draws much of its inspiration from the various ideas reviewed above. As its name implies, this system integrates phonological and phonetic representation in the form of a single data structure, essentially that of autosegmental phonology. Section 3.1 outlines essential properties of the phonological portion of the system, and section 3.2 presents its phonetic portion in fuller detail. We focus here only on the structure of the representations themselves, postponing until section 5 a discussion of how the phonetic part of the representation is related to the phonological part.

3.1 Phonological representation

For the phonological part of the model, we assume the basic results of nonlinear phonology (see earlier references). In particular, we accept the view that phonological representations provide a *skeleton* or timing tier whose units or "slots" constitute the units of segmental quantity. Among other things, the skeleton functions to distinguish phonologically distinctive length. Thus, simple short segments are represented as units linked to one timing unit and long segments are represented as units linked to two timing units, as shown in (2), illustrating a partial representation of short vs. long /a/:

[2] Pierrehumbert and Beckman (1988) have shown that the use of a target-and-interpolation approach in the phonetics eliminates the need for certain assimilation rules in the phonology.

(2)



(Similar representations can be given in terms of mora theory by substituting "μ" for "X".)

In displays such as (2), the phonetic symbols on the lower tier, usually called the "melodic tier" or the "root tier", stand for nodes annotated for values of the features [sonorant, consonantal]. In a complete representation, these nodes (called "root nodes") dominate phonological features arrayed on other tiers. Such features are grouped into intermediate, hierarchically-organized feature classes designating general categories such as "place of articulation", "oral cavity", and "larynx" on the basis of their patterns of phonological cohesion (see Clements and Hume 1995 for fuller discussion of a model of this type). A highly simplified feature representation of *tide*, showing a selection of major class, laryngeal, and place features, is given below:

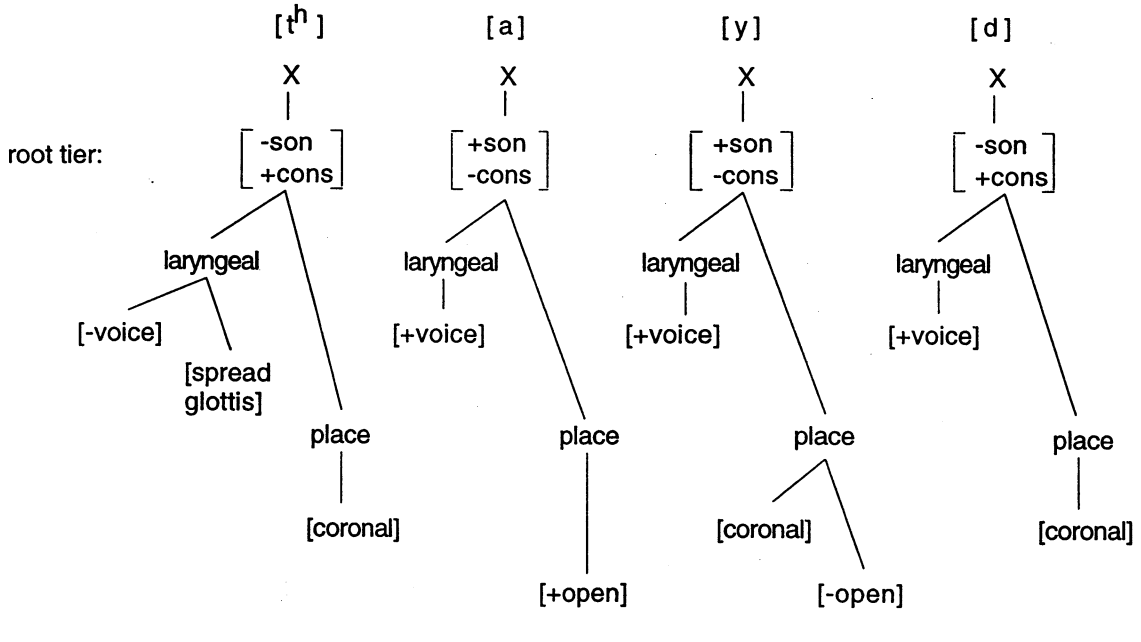


Figure 2. A partial feature representation of *tide* [t^{h}ayd].

A fuller representation of an utterance groups the sequence of timing units into higher-level prosodic constituents. These constituents (syllables, metrical feet, phonological words, and so forth) provide the domains for assigning suprasegmental properties such as tone, stress, and intonation, as well as determining allophonic properties of segmental realization. Following a well-known current of work in syllable theory, we will suppose that English syllables contain a *nucleus* of the form V(G), consisting of the syllable peak V (= the vowel, or syllabic consonant) and an (optional) following glide G.[3] This account is compatible with skeleton-based models of the syllable recognizing a nucleus (e.g. Clements and Keyser 1983, Milliken 1988, Blevins 1995), although not with mora-based frameworks, in which the nucleus is demonstrably superfluous (Steriade 1990). Much of the following discussion could be cast in either a nucleus- or mora-based framework; however, we will see later (in section 5.6.1) that VG sequences are treated as a single constituent by phonetic duration rules.

Figure 3 shows a partial surface representation of the English syllable *tide* [t^hayd] following these assumptions:

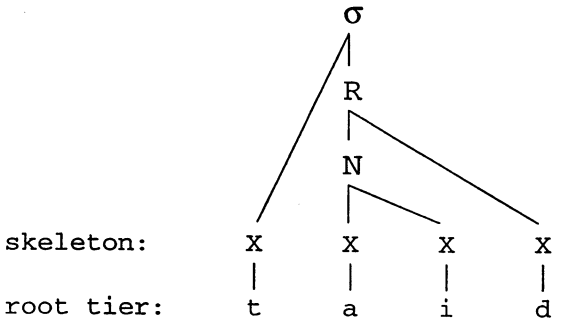


Figure 3. Partial representation of the word *tide* [t^hayd].

Here and elsewhere in the following discussion, we adopt the graphic convention of conflating the root node and all the features characterizing it into a single unit, labelled by an appropriate phonetic symbol. Thus, for example, the symbol "t" in Figure 3 stands for a root node characterized by the features [-sonorant], [-voice], [spread glottis], etc. as

[3] We exclude liquids from the nucleus on grounds that they exhibit looser cooccurrence constraints with the preceding vowel than do semivowels, and that early stress rules single out long vowels and diphthongs as a class of stressable syllable nuclei to the exclusion of vowel + liquid (VL) sequences (see e.g. Selkirk 1982). In fact, we know of no evidence that VL sequences form a constituent in the phonology. However, we will show later (section 5.6.1) that the nucleus must be expanded to include VL at the phonetic level in order to account for important aspects of phonetic duration.

shown by the feature representation in Figure 2. Since the root node designated by "i" in this figure does not function as the syllable peak (which is always the leftmost member of the nucleus in English), it is interpreted as the glide [y] (IPA [j]).

3.2 Acoustic representation

We now consider the acoustic phonetic component of the integrated representational system (IRS). We view this component as forming part of the general theory of representation that underlies speech production at the acoustic phonetic level. Specifically, we hypothesize that in producing utterances, speakers try to produce acoustic patterns that will result in perceptual effects similar to those produced by the speech of other members of the same speech community. To achieve this, speakers construct and execute acoustic *scores* in conformity with the principles of their internalized phonetic grammar. Such scores define the relatively stable properties of the acoustic pattern they wish to produce, and are implemented by an appropriate, continuously-varying sequence of vocal tract shapes.

We propose that acoustic scores are constructed from surface phonological representations through the introduction of a new set of tiers on which acoustic parameter values and duration values are arrayed. Like phonological tiers, these "acoustic tiers" consist of independent sequences of units, or autosegments; instead of phonological features, however, these units consist of *acoustic parameter values*. Phonetic specification at the acoustic level involves in part, therefore, the assignment of appropriate acoustic parameter values to each phonological root node. A root node can be said to be "specified" for a given acoustic parameter if it is formally characterized by a value of that parameter in the acoustic score. As in the case of phonological segments, phonetic segments are not necessarily characterized by values of all acoustic parameters; if a segment is not characterized for a certain parameter in the acoustic representation, it will receive a specification for that parameter at the physical level by interpolation from neighboring values, as discussed above.

Why must acoustic values be specified on separate tiers, rather than on phonological feature nodes, for example? The answer is that there is typically no one-to-one correspondence between phonological features and particular acoustic parameter values: some features are expressed along more than one acoustic parameter (e.g. [labial] involves the lowering of all formants), while some acoustic parameters express several different features (e.g. the value of F1 depends on the specification of [labial], [open], [pharyngeal], and [nasal]). We shall see that the assignment of each acoustic parameter to a separate tier of its own greatly simplifies the formal structure of the representational system, and provides the simplest possible basis for implementation in a target-and-interpolation model.

We now consider in more detail how acoustic scores can be abstracted from the information present in the speech signal. Section 3.2.1 first outlines the phone-and-transition strategy that underlies our approach to acoustic segmentation. Section 3.2.2 then introduces a model of acoustic structure, and shows how acoustic representations can be integrated with feature-based phonological representations in a single, formally unified representational system. The next two subsections motivate the use of multi-tiered representations in the phonetics by discussing two types of non-bijectivity at the acoustic level: overlap (section 3.2.3) and contour phones (section 3.2.4).

3.2.1 A phone-and-transition strategy for segmentation

An essential basis for any linguistically-motivated phonetic analysis is a consistent strategy of segmentation determining how pieces of the continuous acoustic signal can be related to the units and categories of the surface phonological representation in such a way as to permit the straightforward expression of regularities at the phonetic level. The *phone-and-transition* approach to segmentation (Hertz 1991, 1992) has been devised with this requirement in mind. Unlike conventional segmentation strategies, which parse the speech signal into segments of a single type (often termed "phones"), the phone-and-transition model analyzes speech into separate *phone* and *transition* segments.

We will exemplify this segmentation strategy with a simple illustration. Consider the spectrogram of [t^hayd], as uttered by a female speaker (SRH) of General American English in the frame *Say __for me*, shown in Figure 4.

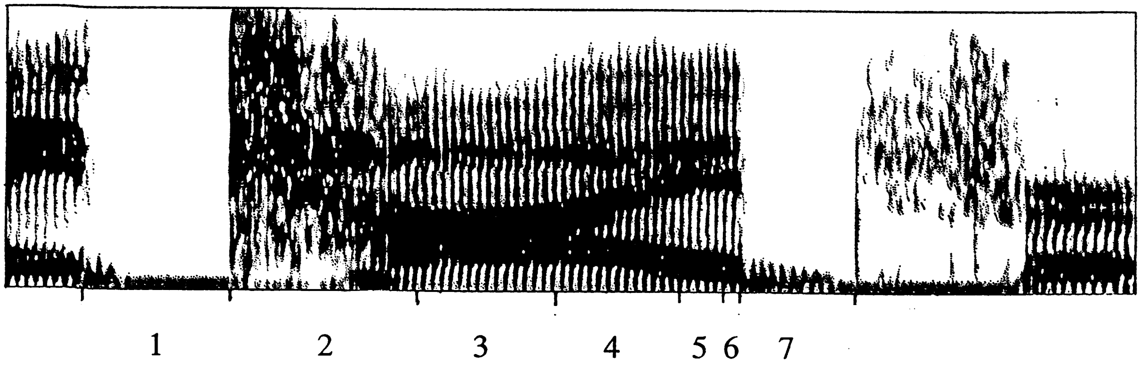


Figure 4. Spectrogram of the word *tide* [t^hayd].

We have segmented this spectrogram, which is representative of many that we have examined for this speaker, into two types of acoustic segments: *phones* and *transitions*.

This segmentation is based on the view that speaking involves the production of a succession of sounds corresponding to the sequence of root nodes (the so-called "melodic" segments) of the surface phonological representation. The sounds that correspond to each root node are not always, or even typically, adjacent to each other in time, but may be separated from each other by intervals during which the lips and tongue move from the articulatory positions appropriate for one to those appropriate for the next. Intuitively speaking, *phones* are the portions of the signal corresponding to the time intervals in which the lips and tongue have achieved their target positions, and *transitions* are the portions that separate them. Phones often appear as relatively steady state portions of a spectrogram (we discuss the less typical case of internally dynamic or "contour" phones in section 3.2.4), while transitions between phones usually involve changes in formant patterns, which are often quite rapid. Not all phones are separated by formant transitions, however; when two phones have the same place of articulation, as in the cluster [mb] in *umbrella*, or when their target positions overlap, as in the cluster [fr] in *free*, there are usually no observable transitions between them.

We follow the common practice of using F2 movements as the primary basis for segmentation. F2 generally gives us a clearer and more consistent basis for segmentation than other formants, which are often harder to read on spectrograms. Furthermore, F2 is generally more responsive to articulator movements, especially those of the tongue blade and tongue body, and shows more rapid variation within a larger range of frequency values. Perhaps for these reasons, we have found that segmentations based on F2 provide a better basis for expressing generalizations about vowel timing patterns than segmentations based on the other formants (see also Ren 1986). When formant structure is not visible, as between adjacent voiceless stops or at stop-fricative boundaries, other criteria must be used, such as acoustic zeroes, bursts or other rapid spectral changes.

The segments in Figure 4 can be analyzed into phones and transitions as follows:

1. The first segment is a *phone* consisting of 85 milliseconds (ms) of silence, corresponding to the time interval during which the articulators maintain a vocal tract configuration appropriate for the voiceless alveolar closure of [t]. This segment ends in a short burst, corresponding to plosive release.
2. The second segment is a 115-ms *transition* characterized by formant movements linking the plosive burst to the steady-state formant structure of the following [a]. Most of this transition is aspirated, while the last 25 ms is voiced. The transition segment corresponds to the return of the glottis from an open configuration to a position appropriate for voicing, and to the simultaneous movement of the tongue to

a position appropriate for [a].

3. The next segment is a *phone* consisting of a well-defined steady-state formant pattern corresponding to the articulation of the low vocoid [a], with a duration of 85 ms. The first formant (F1) maintains a relatively high frequency of about 750 hertz (Hz), while the second formant (F2) maintains a relatively low frequency of 1350 Hz. These values will be called the F1 and F2 targets for the [a]. The dark vertical striations throughout the phone correspond to the opening and closing of the vocal folds, indicating that the [a] is phonetically voiced. Note that the steady-state portions of the F1 and F2 patterns are not perfectly aligned, since the F1 target extends beyond the edge of the phone into the following transition (segment 4). We here use F2 as the basis for our segmentation, for the reasons discussed above.
4. Following the [a] is a *transition* of about 75 ms characterized by a rising F2 pattern. This transition corresponds to the movement of the tongue from the articulatory position of [a] to the articulatory position of [y]. Unlike the aspirated segment following [t], this transition is voiced.
5. The next segment is a 20-ms-long voiced *phone* corresponding to the articulation of the high front vocoid [y], whose F1 value is about 480 Hz and whose F2 value is about 2150 Hz.
6. Following [y] is a 10-ms-long voiced *transition* during which F1 and F2 drop slightly as the tongue moves from the position of [y] to that appropriate for [d].
7. The final segment is a 65-ms-long *phone* corresponding to the articulation of the voiced alveolar stop [d], characterized throughout most of its duration by low-frequency energy (the voice bar) resulting from vocal fold vibration during the stop closure.

Most current segmentation strategies differ from ours in assigning transitions to phones, rather than treating them as independent units. We might call such approaches "all-phone" segmentation strategies. Under such strategies, for example, the aspirated transition numbered 2 in the spectrogram would be assigned either to phone 1 (as in e.g. Chen 1970, Hertz 1982) or to phone 3 (as in e.g. Lehiste and Peterson 1961, Fant 1970, Klatt 1979). A diphthong such as [ay] (segments 2-5 in the spectrogram) necessarily has to be treated either as a single phone or as two phones arbitrarily divided at a replicable segmentation point, such as halfway through the transition.

The distinction between a phone-and-transition strategy and an all-phone strategy is not merely notational, but has empirical consequences. Hertz (1990a, 1990b, 1991, 1992) has discussed several instances in which the failure to treat transitions as explicit units, distinct from phones, results in arbitrary and inconsistent segmentations, and obscures a number of important generalizations about speech. For example, we have observed that intervocalic [h] has no formant targets of its own, but overlays the transition between the two vowels, contributing little or no duration to the utterance (Clements and Hertz 1991; see also Keating 1988). This observation can be directly expressed in a phone-and-transition strategy, but not in an all-phone model. We shall see in section 5.6 below that the rules governing vowel and diphthong duration may crucially treat phones differently from transitions.

3.2.2 A model of acoustic structure

Let us now consider how we can represent the acoustic structure of *tide* in a partially-specified, multitiered representational system (the IRS). For purposes of illustration we will focus on the F2 pattern, which is displayed schematically in Figure 5:

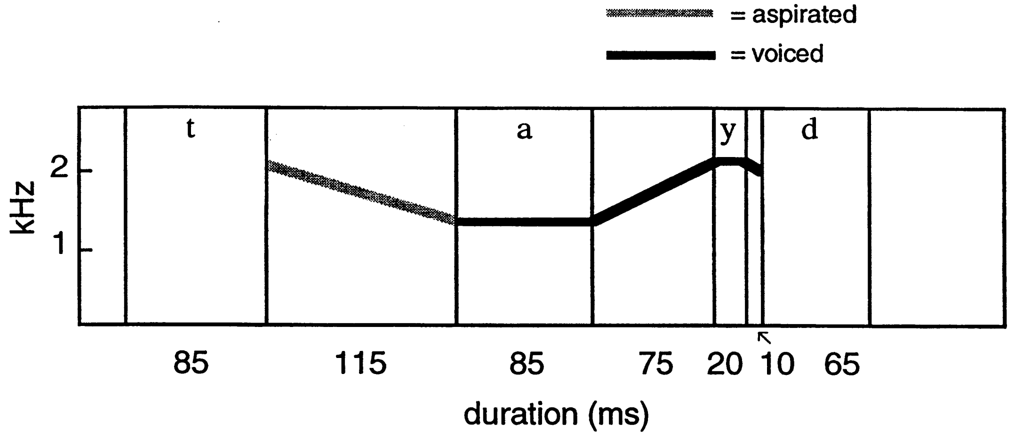


Figure 5. F2 pattern of *tide* [t^hayd], segmented into phones and transitions.

This diagram shows the time course of F2. The shaded portion of the bar represents the fact that this formant is aspirated during the transition from [t] to [a], while the solid portion represents the fact that it is voiced thereafter. This diagram abstracts away from certain acoustic details. For one, we have not indicated the brief period of voicing at the end of the 115-ms transition, since the presence of such voicing varies from one repetition of the utterance to the next and is perceptually insignificant. For another, we have represented F2 as a straight line within each acoustic segment, even though spectrograms show curved transitions at segment junctures. Although we could easily apply a smoothing algorithm to mimic the observed F2 pattern more closely, straight-line interpolation is more than adequate for the purposes of speech synthesis (Holmes 1983), and serves to illustrate the formant patterns under discussion without introducing unnecessary complexity.

We will now show how we can represent this same information in the form of a multi-tiered representation. We first introduce two new tiers to the representation given earlier in Figure 3, a duration tier and an F2 tier. We then specify the duration of each phone and transition on the duration tier and the F2 target values for each phone on the F2 tier. The F2 pattern of Figure 5 can then be generated by interpolation between adjacent target values. The resulting representation is shown below in Figure 6, from which syllable structure has been omitted for convenience; recall that the symbols "t", "a", "i", "d" are shorthand notations for root nodes dominating features on further phonological tiers, not shown here. The new duration and F2 tiers can be thought of as lying on a new plane of structure which intersects the phonological planes at the root tier.

```
skeleton:        x         x          x       x
                 |         |          |       |
root tier:       t         a          y       d
                 |         |          |       |
duration tier:  85   115   85    75   20   10  65
                 |         |          |       |
F2:            2100       1350       2150    2040
```

Figure 6. Multi-tiered representation of duration and F2 values for *tide*.

Observe that the information in this diagram is strictly equivalent to that in Figure 5, as far as F2 and duration are concerned. A duration value (in milliseconds) has been assigned to each root node and each transition, in accordance with the durations shown in Figure 5. A separate tier contains the F2 values (in Hz) for each phone. In the case of phones with no visible formants, such as [t] and [d], we take the F2 targets to be the values that occur at their edges. A given node is *characterized* by the properties that it dominates in the tree; for example, the root node "t" is characterized by a duration of 85 ms and by an F2 target of 2100 Hz. Any root node characterized by at least one formant target is called a *phone*; thus, the phones in Figure 6 are the root nodes designated by the symbols "t", "a", "y", and "d", in conformity with our earlier description. *Transitions* are represented as duration values lying between phones; by definition, they have no independent formant targets of their own (although as we shall see in a moment (section 3.2.3), they may share formant targets with neighboring phones).

While the information in the above representation is sufficient for deriving the shape of the F2 pattern, a complete specification of the information included in Figure 5 must include aspiration and voicing. This information can be specified on additional tiers, as shown in Figure 7 (the skeleton is omitted for convenience). In this figure, values on the voicing tier represent voicing amplitude in decibels (dB). A value of 0 dB, interpreted as absence of voicing, is linked to the [t] and to the transition following it. (Notice that we cannot leave such segments unspecified for voicing values, since if they were unspecified, the interpolation algorithm would assign them intermediate values.) A value of 55 dB is assigned to all phonologically [+voice] phones and to the transitions between them. While in this case we could have left the intervening transitions unspecified and derived their values by interpolation, no interesting regularities are accounted for in this way, and we will assume generally that multilinked acoustic values may not skip intervening nodes.

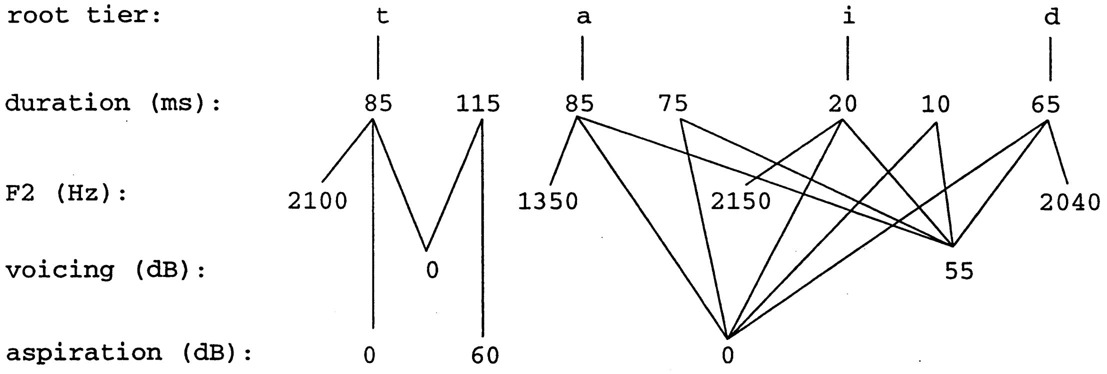


Figure 7. Multi-tiered representation of *tide*, including duration, F2, voicing, and aspiration tiers.

Values on the aspiration tier represent aspiration noise amplitude in a similar fashion. Thus, the aspiration value of 60 dB linked to the transition between "t" and "a" represents the fact that this transition is aspirated, while the value of 0 dB linked to all other segments indicates that they are unaspirated.

Notice that we have represented acoustic target values that are common to a sequence of segments as a single value linked to all members of the sequence. Thus, for example, a single value of 55 dB on the voicing tier is linked to each of the final five units on the duration tier. This mode of association follows from a convention, reminiscent of the Obligatory Contour Principle in phonology (McCarthy 1986), that prohibits identical adjacent values. This convention eliminates the possibility of drawing a formal distinction between two segments sharing a single value for some acoustic parameter and two segments each bearing separate, identical values for it.

The duration tier in representations like Figure 7 has the function of coordinating and sequencing the various acoustic values characterizing a phone, much as the root tier coordinates and sequences sets of phonological features characterizing a segment in the phonological part of the representation. In this sense, the duration tier functions something like the "spine" of the acoustic representation. Duration values, to the extent that they are language-particular, represent the regular temporal variation that can in principle distinguish the phonetic forms of one language, dialect or idiolect from another, such as differences in intrinsic vowel and consonant duration. We will see in our discussion of contour phones (section 3.2.3) that more than one duration value may be assigned to a single phone; and we will see in our discussion of duration rules (section 5.6.1) that durational values may be assigned not only to phones as shown here, but also to the syllable nucleus (among perhaps other, higher-level prosodic units).

We see, then, that by extending the formalism of multi-tiered phonological representation into the acoustic phonetic domain, we can provide precise characterizations of selected aspects of the acoustic patterns observable in spectrograms, and relate them directly to the units of the surface phonological representation that they express. A full representation of the acoustic properties of *tide* would of course include additional tiers representing other acoustic information, such as fundamental frequency, F1, F3, frication, and so on.

We next consider two cases in which multi-tiered representations prove particularly appropriate.

3.2.3 Acoustic overlap

One of the major problems in phonetic segmentation is what we will call *acoustic overlap*: the fact that the different acoustic properties characterizing a phone do not always align neatly with each other. For example, in our spectrogram (Figure 4) the voicing of the phone [a] actually begins toward the end of the preceding transition, the steady state portion of its F1 extends some 40 ms into the following transition, and the final [d] is devoiced toward its end. In carrying out a principled segmentation in such cases, we must establish clear-cut criteria for defining the beginning and end of each phonetic segment.

Our representational system is powerful enough to handle acoustic overlap. Figure 8 shows how it can represent the overlap between F1 and the [a]-to-[i] transition in *tide*.

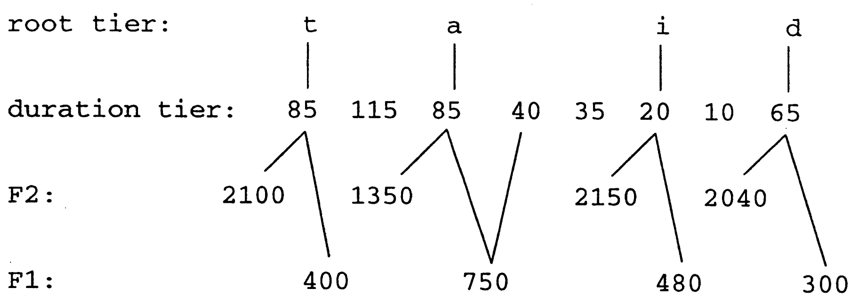


Figure 8. Partial representation of *tide*, illustrating a misalignment between F1 and F2.

Here the 750 Hz F1 target of the phone [a] is shown as extending onto the first 40 ms of the following transition. The second part of this transition (35 ms) bears no F1 target value, and will be realized with values interpolated between the 750 Hz target to its left and the 480 Hz target to its right. This analysis directly captures the overlap between F1 and the following transition in our spectrogram.

Acoustic overlap as shown in this figure provides strong support for a representational system in which values for each acoustic parameter are placed on separate tiers, instead of on a single tier (e.g. in the form of a two-dimensional matrix). This is because only a multi-tiered mode of representation permits the straightforward expression of overlap. By representing overlap as a one-to-many association between a single acoustic parameter value on one tier and two units of duration on the duration tier, we are able to express the generalization that transitions between phones do not bear independent target values of their own; whenever a transition bears a formant target in an acoustic representation, this target has originated in a neighboring phone from which it has spread or shifted through anticipation or lag, as is formally represented by one-to-many association. If all acoustic parameter values were instead aligned with segments in a one-to-one fashion, as in a two-dimensional matrix, we would have no satisfactory way of expressing the distinction between spreading target values and adjacent, independently-specified target values which happen to be identical, and thus we would have no nonarbitrary way of expressing the fact that transitions are, by definition, intervals with no formant target values of their own.

In the case we have just examined, overlap involved an acoustic parameter that extends partway into a transition. There are also cases in which an acoustic property extends partway into a phone. For example, in our sample spectrogram for *tide* the phone [d] is phonetically devoiced for 10 ms at its end, which is a regular feature of pronunciations of *tide* by this speaker. We can represent this partial devoicing by assigning a 0-Hz voicing value to the final 10 ms of the phone, leaving a positive voicing value of 55 dB associated with the remainder of its duration. This representation is shown in Figure 9.

```
root tier:        ...          d
                              / \
duration tier:    ... 10   55   10
                        \ /       |
voicing tier:     ...    55      0
```

Figure 9. Representation of phrase-final [d] in *tide*, illustrating partial final devoicing.

Spectrograms typically show very many cases of acoustic overlap, and it is necessary to establish a criterion for determining how much overlap to model in our representations. One criterion might be that of perceptual detectability: any case of overlap is represented as long as it can be detected by speakers. In our experience listening to pairs of synthesized utterances varying in minimal respects, not all formant misalignments of the type illustrated above are perceptually detectable. However, there are many marginal cases, and it is not always straightforward to decide which information is detectable and which is not. Furthermore, even when a certain acoustic property is undetectable by itself (in the sense that acoustic "minimal pairs" differing only in its presence or absence sound identical), the accumulation of several such properties often results in a notable improvement of the overall speech output; in such cases, it is arbitrary to determine which properties should be omitted and which should not. An alternative criterion, consistent with our overall goals, might be to retain only those cases of overlap that represent the application of language-particular rules and principles. However, this criterion proves to be unworkable in practice, as in most cases we simply do not know which types of overlap are language-particular and which are determined by language-independent mechanisms. As a result of these considerations, we have provisionally adopted a third criterion according to which *all acoustic properties that are regularly present in a certain utterance are represented in its acoustic score*, whether they are individually perceptible or not; thus, only distinctions that are often absent in the speech of a single speaker are omitted. We leave it to further research to determine whether there is a principled way of further eliminating some of the less salient acoustic detail from acoustic scores.

3.2.4 Contour phones

In the examples discussed so far, each phone has been characterized by a single target value for each formant throughout its duration. However, we frequently observe that some phones have different values for one or more formants at their beginnings and ends, which cannot be predicted by spreading or interpolation from neighboring target values. We call such phones *contour phones*.

We offer a simple example here. It is well known that the F2 target value at the release of a consonant may vary depending upon the identity of the following vowel. In general, this value is higher before front vowels and lower before back rounded vowels. These differences can be understood as an effect of coarticulation: at the moment of release of a consonant in a CV sequence, the tongue and lips typically anticipate the articulatory position of the following vowel, to the extent that the consonant and vowel articulations are compatible (Wood 1991). Similar remarks hold for the F2 target value at the closure of a consonant in VC sequences, which varies analogously depending on the preceding vowel.

It follows that when a consonant occurs in a VCV context, the formant targets at its closure and release will typically be somewhat different. For example, in the sequence [oka] (as in *okapi*) as spoken by SRH, the F2 value at the left edge of [k] averages around 880 Hz, while the value at the release of [k] averages around 1600 Hz. These values constitute "targets" in the sense of a target-and-interpolation model, since they cannot be predicted from adjacent F2 values. The F2 pattern of [oka] is shown in Figure 10:

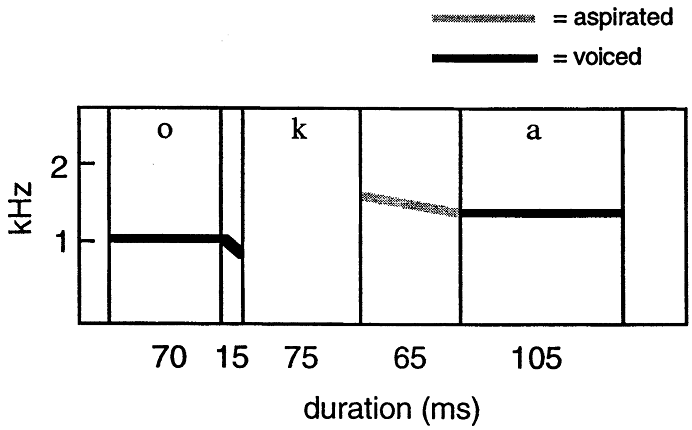


Figure 10. F2 pattern of [oka], showing different F2 targets at the left and right edges of [k].

We can account for such "contour phones" in our model by allowing phones to be characterized by two different formant target values at their left and right edges. This characterization can be achieved by inserting zero-duration "place-holders" at the left and right edges of phones on the duration tier, to which potentially different formant target values can be associated. This structure is shown in Figure 11 below, for the [k] in [oka]:

```
root tier:                   k
                           / | \
duration tier:            0  75  0
                          |      |
F2 tier:                 880   1600
```

Figure 11. Schematic representation of a contour phone.

This representation allows us to assign an F2 value of 880 Hz at the left edge of [k] and a value of 1600 Hz at its right edge. Other values must of course be assigned in other vowel contexts. By assigning each of these targets a zero duration, we treat them as simple "inflection points" in the F2 time course. These targets are not visible on a spectrogram during the closure portion of [k], but shape the observable F2 transitions into the neighboring vowels. When the left and right targets happen to have the same value (as may happen when the consonant is flanked by identical vowels), these values can be conflated into a single token and associated to a single, non-zero duration token, as we did in our earlier representations of the [t] in *tide* (Figures 6-8). Contour phones can of course involve sequences of other acoustic parameter values, and can be created by overlap with neighboring phones or transitions, as was shown earlier in Figure 9.

Contour segments offer further evidence for a multi-tiered model of acoustic representation. If all acoustic and durational values were listed on root nodes (or other higher-level nodes, such as the units of the skeleton) in the form of a two-dimensional matrix, we would not be able to express the temporal order of sequenced parameter values within a contour phone. Moreover, our later phonetic examination of English vocalic nuclei in section 5 will show that we must draw a crucial distinction between genuine contour phones and phone sequences, a distinction which would be lost under such a proposal.

An alternative way of incorporating multitiered phonetic representations into the grammar would be to make use of the Delta representational system, an approach explored by these writers in previous work (Clements and Hertz 1991). In this approach, autosegmental representations at the phonological level are interfaced with two-dimensional Delta "multistream" representations at the phonetic level, viewed as a new plane of representation intersecting the phonological family of planes at the level of the root tier. A Delta representation of the [k] in *oka* is given below, for comparison with Figure 11:

```
root stream:      |          k           |
duration stream:  |  0    | 75     | 0     |
F2 stream:        |  880  |       | 1600  |
```

Figure 12. Delta representation of a contour phone.

In this representation, vertical bars, called "synch marks", align simultaneous acoustic events. It will be seen that this type of representation is equally capable of describing the acoustic pattern of Figure 10, while preserving the one-to-many relation between the phone [k] and the sequence of acoustic segments that characterize it. Delta notation is, in fact, largely equivalent to autosegmental notation in its expressive power, and is easily interconvertible with it (Hertz 1990a). While a "hybrid" model mixing autosegmental and Delta notation is workable, it has the disadvantage, from our point of view, of creating a purely artifactual notational difference between the phonological and phonetic levels. By unifying the phonological and phonetic levels in terms of a common representational system, we can directly explore the relations between phonological and phonetic structure.

In particular, one of our guiding assumptions, which we term the *Congruence Hypothesis*, is that given an appropriately integrated framework for phonological and phonetic description, the optimal representations required for the expression of generalizations at both levels, phonological and phonetic, will be similar. That is, we expect to find no substantial mismatch between the surface representations given by the phonological analysis and the phonetic representations required by the phonetic analysis, the latter consisting largely of a fuller specification of the former through the addition of new tiers. The strongest possible version of this hypothesis is, of course, that there is no mismatch at all. If this hypothesis (in either a strong or a weak version) proves to be true, alternative descriptions of any given phenomenon can be directly compared and evaluated, in part, in terms of the complexity and arbitrariness of the restructuring operations they require to convert surface phonological representations into appropriate phonetic representations.

We now explore the Congruence Hypothesis by examining a case study in English phonology and phonetics involving the analysis of long vowels and diphthongs. Our ultimate goal will be to determine whether the representations required by the phonology can provide optimal input for the expression of acoustic regularities at the phonetic level.

4 Phonological analyses of long vocalic nuclei in English

We address a long-standing problem in the phonology of the English vowel system. Phonologists and phoneticians have long disagreed over the phonological analysis of long vowels and diphthongs in GAE (General American English, a term we use to designate a class of dialects showing no marked regional characteristics).[4] Most have agreed that the vowels in words like *bit*, *put*, *bet*, *but*, *bat* and *pot* consist of single, phonologically short segments, as suggested by the single graphemes usually used to transcribe these vowels: [ɪ, ʊ, ɛ, ʌ, æ, ɑ]. Moreover, many linguists and phoneticians (though not all) would also agree that heavily diphthongized vocalic nuclei, such as those of *bite, bout, and boy*, comprise two segments, analyzing the nucleus of *bite*, for example, as the sequence [a] + [y] at the surface phonological level.

In contrast, there has been continuing disagreement regarding the surface phonological analysis of the phonetically diphthongized long vowels in words like *beet*, *bait*, *boot*, and *boat*. In some analyses, these nuclei have been analyzed as single phonological segments; in others, they have been analyzed as bisegmental sequences, on a structural par with the bisegmental nuclei of *bite* and *bout*. In a two-level generative framework which distinguishes underlying and surface representation, these nuclei have typically been analyzed as single segments underlyingly and as two segments on the surface, where they result from a general rule of diphthongization.

For purposes of the following discussion, we will refer to these three groups of syllabic nuclei mnemonically as *SV-nuclei, D-nuclei, and LV-nuclei*, respectively. (This terminology is intended only for convenience, and should not be understood as reflecting a bias toward any particular analysis.) In (3) we list several representative previous analyses, according to their treatment of the two types of long syllabic nuclei.[5]

[4] Although we focus on GAE, the variety of English which we have studied most carefully, an analogous problem of analysis arises in many other varieties.

[5] We do not consider the long nuclei of words like *few* and *fir* in this study, which show different characteristics and require separate discussion.

(3) Some previous analyses of long syllabic nuclei in GAE:

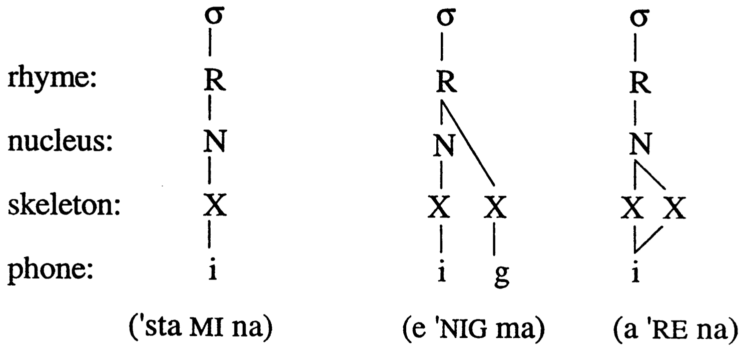
	LV-nuclei	D-nuclei
Swadesh 1935	1 segment	1 segment
Trager and Bloch 1941, Trager and Smith 1951	2 segments	2 segments
Pike 1947	1 segment	2 segments
Chomsky and Halle 1968, Halle and Mohanan 1985	1 segment (underlying), 2 segments (surface)	1 segment (underlying), 2 segments (surface)
???	2 segments	1 segment

Further subclassifications have occasionally been proposed, especially in the phonetic literature. For example, some phonetic studies have proposed a three-way classification into long tense nuclei [i u], gliding nuclei [e^{I} o^{U}], and two-target nuclei [ay aw ɔy] (Lehiste and Peterson 1961). Others have classed [e^{I} o^{U}] (but not [i u]) together with [ay aw ɔy] (Holbrook and Fairbanks 1962). As the final entry in (3) indicates, however, we know of no linguist or phonetician who has proposed to analyze LV-nuclei as two segments and D-nuclei as one.

The earlier of the analyses summarized above have assumed a unilinear or "beads-on-a-string" mode of phonological representation, in which phonological length is measured strictly in terms of the number of successive segments in the string. In such models, a short nucleus consists of one segment and a long nucleus of two. In contrast, more recent analyses of English have recognized a skeletal or timing level independent of the "melodic" tier, as discussed in section 3.1 (Halle and Mohanan 1985; Pike 1947 may also be understood as assuming some such distinction). In this framework, there are two potential ways of measuring phonological length: (i) by counting timing units, or (ii) by counting root nodes. Let us call these two types of length *metrical length* and *melodic length*, respectively. Although Halle and Mohanan analyze all long vocalic nuclei as one melodic segment in underlying representation, they link this segment to two skeletal positions as shown in (2) above, thereby accounting for its phonological length.

There is, in fact, strong evidence supporting the view that English long vocalic nuclei count as two units in a metrical sense, as proposed by Halle and Mohanan (see also Anderson and Jones 1977, Halle 1977, and Selkirk 1982, among others, for relevant further discussion in earlier frameworks). For example, it is well known that the long nuclei behave analogously to VC closed syllables in terms of their ability to attract stress. This observation was first formulated by Chomsky and Halle 1968 in terms of the distinction between "strong" and "weak" clusters. They noted that main stress characteristically falls on the penultimate syllable of nouns if it either contains a long tense vowel, as in *arena* and *Arizona*, or if it is closed, as in *enigma* and *charisma*. These two syllable types, often called "heavy syllables", form a natural class with respect to the stress rules. In contrast, if the penultimate syllable is light (that is, if it ends in a short lax vowel), it is usually unstressed, as in *stamina*, *retina*, and *Canada,* although it is sometimes stressed, as in *manila* or *regatta*. The generalization here is that a penultimate heavy syllable attracts stress, with very few exceptions.[6] This generalization can be captured under the assumption that long vowels are bipositional—that is, linked to two units of the skeleton. Given this analysis, closed syllables and syllables containing long vowels both have two skeletal units in their rhymes. We may now characterize the class of heavy syllables as those that have as least two skeletal units in their rhyme. This analysis is illustrated in (4) with representations of the penultimate syllable rhymes (shown in capital letters in the keywords underneath) of *stamina*, *enigma*, and *arena*: [7]

(4) light syllable: heavy syllables:



Given such representations, we may formulate the stress rule in terms of a single configurational property shared by all heavy syllables. If we had instead taken a feature like [±tense] as the property underlying the distinction between the long and short vocalic

[6] The rare exceptions to this generalization, such as *character* and *calendar*, usually end in syllabic consonants. In some analyses, these consonants are treated as extrasyllabic at the point at which main stress is assigned, making the heavy syllable final.

[7] The representation of *arena* shows an intermediate level of representation, following Vowel Shift but preceding Diphthongization; these rules are further discussed below.

nuclei, we would have been unable to characterize all heavy syllables as a natural class.

While a considerable amount of evidence converges on the view that the long vocalic nuclei count as two units at the level of the phonological skeleton, there is much less evidence bearing on their count at the melodic level. Indeed, most of the evidence which has led linguists in the past to view them as consisting of two units has involved stress and syllable structure patterning, which bear on "metrical" length but not on "melodic" length. What type of evidence might lead us to analyze a given vocalic nucleus as consisting of two melodic units (or "phonemes", for short) at the surface phonological level? We may consider the following criteria:

1. *Observational adequacy*: This criterion requires us to account for the fact that all the nuclei in question are realized as phonetic diphthongs, at least in certain contexts, at some level of our analysis. However, we have no a priori reason for preferring a phonological analysis to a phonetic one.
2. *Patterning evidence*: It is often supposed that a phonetically complex segment AB can be analyzed into a phoneme sequence AB if A and B exist as independent phonemes in the language. This is true of the English vocalic nuclei in question; for example, the two components of the diphthong [ay] resemble the initial phonemes in *art* [art] and *yard* [yard], respectively. But this criterion is not sufficient for a biphonemic analysis: for example, the aspirated initial sound of *tie* [t^hay] resembles the phoneme sequence / t / + / h /, yet [t^h] does not count as two phonemes in *tie*.
3. *Edge effects*: If a phonetically complex segment [AB] patterns phonologically like phoneme A to its left and phoneme B to its right, but not vice versa, it can be regarded as the phoneme sequence AB (see Anderson 1976 for relevant discussion). However, English long vocalic nuclei display no known edge effects.
4. *Melody association*: Given a morphological process that assigns phonemes to positions in a prosodic template in a one-to-one fashion, a phonetically complex segment [AB] counts as two phonemes /AB/ if it is assigned to two template positions. However, English has no processes of this type.

Thus some of the more familiar criteria for determining "melodic" count are either inconclusive or inapplicable in English. However, a further criterion is of more interest:

5. *Cluster-like behavior*: If a phonetically complex sequence [AB] behaves like a cluster with respect to phonological rules applying at the melodic level, it can be analyzed as the phoneme sequence AB.

Potential evidence of this type can be drawn from the familiar pattern of synchronic alternations created by the historical Great Vowel Shift. These alternations have traditionally been analyzed, within the generative tradition, in terms of a rule of Vowel Shift which respecifies the features [high] and [low], together with a rule of Diphthongization, which creates two derived vowels from single long vowels. Since Diphthongization applies only to long vowels, it does not affect metrical length but only melodic length. If we can show that Diphthongization feeds a phonological rule that treats its output as a cluster at the melodic level, we will have established its phonological status.

Before we can examine this point, we must provide a brief background. Typical Vowel Shift alternations are illustrated in (5). Our surface transcriptions follow those of Kenyon and Knott (1944), except that we follow the widespread practice of substituting [y] for [ɪ] and [w] for [ʊ] in the transcription of diphthongs.

(5) a. ay ~ ɪ: divine/divinity, elide/elision
b. i ~ ɛ: serene/serenity, supreme/supremacy
c. e ~ æ: sane/sanity, inflame/inflammable
d. o ~ ɑ: verbose/verbosity, cone/conic
e. (y)u ~ ʌ: reduce/reduction, presume/presumption
f. aw ~ ʌ: profound/profundity, pronounce/pronunciation

In typical generative analyses, the alternating vowels in the related pairs of (5a-d) are derived from underlying vowels whose phonological (i.e. metrical) length is identical to that of the first member of the pair and whose height is identical to that of the second member. In such analyses, for example, the underlying vowel of *divine ~ divinity* is long / ī / and the underlying vowel of *serene ~ serenity* is long / ē /. The stressed vowel in the first member of each such pair undergoes Vowel Shift, and the corresponding vowel in the second member undergoes Shortening, which takes precedence over Vowel Shift (for discussion of Shortening, see Myers 1987, Burzio 1993). This analysis gives us the following classification of the alternating English vowels and diphthongs illustrated in (5):

(6) Phonological analysis of selected syllabic nuclei in GAE (Halle and Mohanan 1985):

	short nuclei:			long nuclei:		
	underlying	phonetic		underlying	phonetic	
a.	i	ɪ	pit	ī	ay	bite
	ɨ	yu	venue	ɨ̄	aw	bout
b.	e	ɛ	pet	ē	i	peat
	ʌ	ʌ	putt	ʌ̄	yu	view
c.	æ	æ	pat	ǣ	e	bait
	ɒ	ɑ	pot	ɒ̄	o	boat

The effects of Vowel Shift can be seen by comparing the first and second columns in the right half of this diagram, under the heading "long nuclei". As the diagram shows, long high vowels are lowered and diphthongized (6a), long mid vowels are raised to high (6b), and long low vowels are raised to mid (6c). In the SPE analysis (Chomsky and Halle 1968), these processes were expressed in terms of three rules. Vowel Shift proper reverses the value of the feature [±high] in nonlow vowels ([-low, αhigh] → [-αhigh]) and the value of the feature [±low] in nonhigh vowels ([-high, αlow] → [-αlow]). Diphthongization, ordered before Vowel Shift, turns all long vowels (here analyzed as tense) into bisegmental sequences. Finally, a later rule of æ-Backing backs the low vowel [æ] to [a] before a glide. As Chomsky and Halle point out, this set of rules receives further support from a set of alternations involving vowel lengthening before suffixes of the form /-CiV/. Thus, for example, the underlying short / ɒ / of *Milton* (cf. *Milt*[ɒ]*nic*) is lengthened before the suffix /-iən/, and then raised to [o] (*Milt*[o]*nian*). As Chomsky and Halle show, the rules required to account for the alternations in (6) extend straightforwardly to these further alternations.

In their original analysis, Chomsky and Halle embedded Diphthongization deeply in their rule system, ordering it before Vowel Shift. Subsequently, however, Halle (1977) proposed a new analysis in which Diphthongization is ordered after Vowel Shift, consistently with its nonlexicalized and phonologically exceptionless nature.[8] In the later analysis

[8] Diphthongization does not, in fact, exhibit the characteristics of a lexical rule as proposed by Kiparsky 1982 and Mohanan 1986. For example, Diphthongization does not show Strict Cyclicity effects, has no morphological conditions or lexical exceptions, and crucially precedes no other lexical rule. Interestingly, it also shows no unambiguous characteristics of a postlexical phonological rule; for example, it is not a neutralizing rule, as it does not merge any underlyingly distinct representations. Just below we consider

of Halle and Mohanan (1985), which incorporates this revision, Diphthongization follows Vowel Shift but still crucially precedes the rule of æ-Backing. This ordering is required to allow derivations like that of *tide*, which now goes as follows: / tīd / (underlying) → tǣd (by Vowel Shift) → tæyd (by Diphthongization) → [tayd] (by æ-Backing). As this derivation shows, æ-Backing treats the output of Diphthongization as a cluster at the melodic level, changing [æ] to [a] before the glide [y]. By the criterion of Cluster-like Behavior suggested earlier, then, this analysis entails the biphonemic status of the diphthongs at the phonological level.

The rule of æ-Backing, which plays a central role in this argument, is motivated by the need to adjust what would otherwise be an incorrect *[æy] resulting from Vowel Shift.[9] Its ordering after Diphthongization allows us to maintain the generalization that the glide inserted by Diphthongization always agrees with the vowel in its value for the feature [±back]. If we reformulated æ-Backing as a context-free rule directly shifting derived [ǣ] to [ā], for example, and ordered it before Diphthongization, Diphthongization would have to insert a glide *disagreeing* in its value for [±back] in this case only. Such an analysis would be both more complicated and less natural. Halle and Mohanan's analysis provides some justification, then, for the treatment of at least the diphthong [ay] as a cluster at the melodic level. Since Diphthongization is formulated to apply to all long vowels and not just derived [ǣ], it follows in their analysis that *all* long vocalic nuclei are bisegmental in the surface phonology.

This review completes our presentation of the Halle/Mohanan classification of LV- and D-nuclei as summarized in (3). There is, however, another approach to these facts which does not have the same consequences for the analysis of surface long vowels and diphthongs. Various aspects of this approach have been independently suggested by several writers including Schane 1984, Falk 1991, Sluyters 1992, and Labov 1994. We shall here review the analysis proposed by Falk (1991), as it is carried out in an autosegmental framework comparable to that assumed by Halle and Mohanan.

In Falk's analysis, the synchronic residue of Vowel Shift is analyzed into two central processes. First, a breaking rule of I-Delinking diphthongizes underlying long [ī] by delinking its association to the first skeletal slot; this rule feeds a further rule which fills in

the one argument for the phonological status of Diphthongization known to us, based on its interaction with the rule of æ-Backing.

[9] Note that æ-Backing has no further motivation. While it also could account for the general absence of [æ] before glides in the surface phonology, this gap is only a subcase of a more general constraint prohibiting all vowels other than [a ɔ] before tautosyllabic glides.

the empty position with the feature [+open] (eventually realized as [a]). Second, a raising rule of Aperture Specification raises long nonhigh vowels in scalar fashion, shifting mid vowels to high and low vowels to mid. This analysis is summarized in the following derivations of [ay] and the LV-nuclei [i], [e], and [o]:

```
(7)        [ay]    [i]     [e]     [o]

    a.     X X     X X     X X     X X      underlying representation
           \ /     \ /     \/      \ /
            i       e       æ       ɒ

    b.     X X                              I-Delinking, [+open]
           | |      -       -       -           insertion
           a i

    c.             X X     X X     X X      Aperture Specification
            -      \ /     \ /     \ /
                    i       e       o
```

In this analysis, only I-Delinking creates a bisegmental sequence ([ay]) in its output; Aperture Specification leaves the melodic count of the LV- nuclei [i e o] unchanged.[10] In fact, Falk's analysis "ignores the glide of [ey]" and the other LV-nuclei (p. 492), and leaves their ultimate status open. What is significant about this analysis, from our point of view, is that it provides *nonparallel* accounts of the LV-nuclei and the D-nuclei. The D-nuclei are all bisegmental on the surface, while no claim is made regarding the L-nuclei, which remain unisegmental in the output of Falk's rule system.

We have seen, in summary, that there is considerable disagreement among linguists regarding the surface analysis of the long vocalic nuclei of GAE. As far as more recent analyses are concerned, one (that of Halle and Mohanan 1985) treats all long nuclei in parallel as bisegmental sequences at the surface level, while another (that of Falk 1991) treats only D-nuclei as bisegmental melodic sequences on the surface, leaving the surface phonological representation of LV-nuclei open, and thus potentially allowing their diphthongal quality to be derived in the phonetics.

[10] Falk treats [aw] and [ɔy], which do not exhibit robust vowel shift alternations, as underlying diphthongs.

5 Durational properties of vocalic nuclei

Our concern now is to consider the analysis of long vocalic nuclei from a phonetic perspective. Our aim will be to find out whether a closer examination of the phonetics of vocalic nuclei in GAE can shed further light on the monosegmental or bisegmental status of the long vocalic nuclei (D-nuclei, LV-nuclei). Evidence that a given nucleus type behaves as two segments in the phonetics will support a phonological treatment of Diphthongization, since a phonological treatment creates two roots nodes, and hence two melodic segments, directly in its output. Evidence that a nucleus type behaves as one segment in the phonetics will, in contrast, support a single-root-node analysis, and will consequently argue against a phonological treatment of Diphthongization. We shall be especially interested to see whether D- nuclei and LV-nuclei behave in parallel.

Our strategy will be as follows. After a brief discussion of our methodology (section 5.1), we will examine durational characteristics of SV-nuclei, which are uncontroversially single melodic segments in the phonology (section 5.2). With this background, we then consider the durational properties of D-nuclei (section 5.3) and LV-nuclei (section 5.4) in turn, with the aim of determining whether they behave like single segments or sequences. On the basis of our results, section 5.5 proposes a formal analysis of the representation of vocalic nuclei. Section 5.6 presents a model of duration and formant target values assignment, and section 5.7 discusses the issue of phonological/phonetic mismatch.

5.1 Methodology

We take the position that theory construction at the phonetic level follows the same principles that apply elsewhere in linguistic theory. Accordingly, our approach to phonetic analysis draws upon a similar methodology, which involves studying the behavior of any subsystem within the context of the larger system of which it forms a part, describing this behavior in terms of an explicit theoretical model, and formulating rules and principles that make testable predictions throughout the system as a whole.

Our approach is centered in hypothesis formulation and testing, involving both analysis and synthesis carried out in repetitive cycles. The *analysis phase* of each cycle involves studying natural speech data, mainly by examining spectrograms, segmenting them into phones and transitions, comparing and measuring their durations, and formulating generalizations about the observed patterns.[11] The *synthesis phase* is carried out in the

[11] Currently, Hertz and her associates are developing a multi-dialect (and eventually, multi-language) database of linguistic and phonetic information (Hertz et al. 1994). By querying the database, it is possible to obtain durations and acoustic values in specific contexts within and across dialects and languages.

framework of the Delta System, a speech synthesis rule development tool centered around a multi-tiered utterance representation that can accommodate both phonological and phonetic information (cf. Figure 12; see Hertz 1988, 1990a, 1990b, 1991, and Hertz and Huffman 1992 for more information). This phase involves first writing synthesis rules embodying our hypotheses, and then implementing the rules and evaluating their synthetic output. Evaluation is both visual and auditory. Visual evaluation involves comparing spectrograms and other acoustic records of natural and synthetic versions of the same utterance. Auditory evaluation is carried out by synthesizing an utterance, listening to it, comparing the result with natural versions of the same utterance, and modifying the synthesis rules as required. Informal intelligibility and naturalness tests are administered periodically using naïve listeners.

We must call attention to an important limitation on our study. Ideally, any phonetic study of a language or dialect should be based on data drawn from several representative speakers. In the present case, data have been collected in the context of an extensive study of the acoustic system of a single, representative speaker of GAE (one of the authors, SRH) conducted over nearly twenty years. Consequently, our generalizations and analysis hold strictly only for this subject, and may not extend in all details to other GAE speakers. Further study of other subjects is currently in progress (preliminary results for one GAE speaker are reported in Clements, Hertz and Lauret 1995). In addition, Hertz and her associates are extending their database to include speakers of other American English dialects (note 11). These studies will help to determine how many of the generalizations reported here hold for other speakers and dialects as well.

5.2 Durational properties of SV-nuclei

We first consider the acoustic structure of SV-nuclei. Many writers describe the SV-nuclei as short monophthongal vowels. Figure 13a illustrates one such pronunciation of the SV-nucleus [ɪ] of *bid* as produced in the test frame *Say __for me*:

Multi-tiered utterance representations for synthesis with the Delta System can be constructed automatically on the basis of information extracted by querying the database. One of the main goals of this work is to separate rules and generalizations specific to certain dialects or languages from those of more general (perhaps universal) validity.

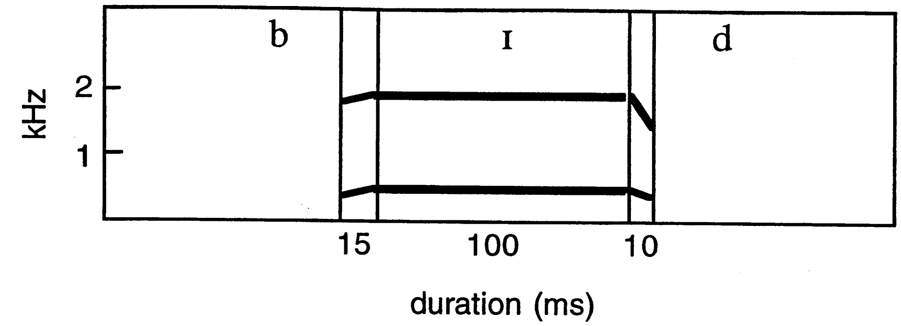


Figure 13a. Monophthongal realization of the short vowel of *bid* [bɪd].

An important exception to the monophthongal realization of SV-nuclei must be made for / æ ɔ /, usually treated as short for phonological reasons (Halle and Mohanan 1985), but which are raised and produced with salient "ingliding"—that is, a movement of all formants to a roughly schwa-like target configuration—by many speakers (Labov 1994). In addition to these, we have found that some GAE speakers (including one of the authors, GNC) typically produce the short vowels of *bid* and *bed* with phonetic ingliding in many contexts, for example in stressed syllables before labial and (non-liquid) alveolar phones. This phenomenon has been reported for GAE by other researchers (e.g. Wells 1982: 485, Bailey 1985, Allen, Hunnicutt and Klatt 1987), and is widely reported in other dialects as well (see e.g. Kurath and McDavid 1961, Foley 1972). This type of realization is illustrated schematically in Figure 13b, showing a representative F1 and F2 pattern of the word *bid* as spoken in the same test frame by a spęaker of the "ingliding" dialect.

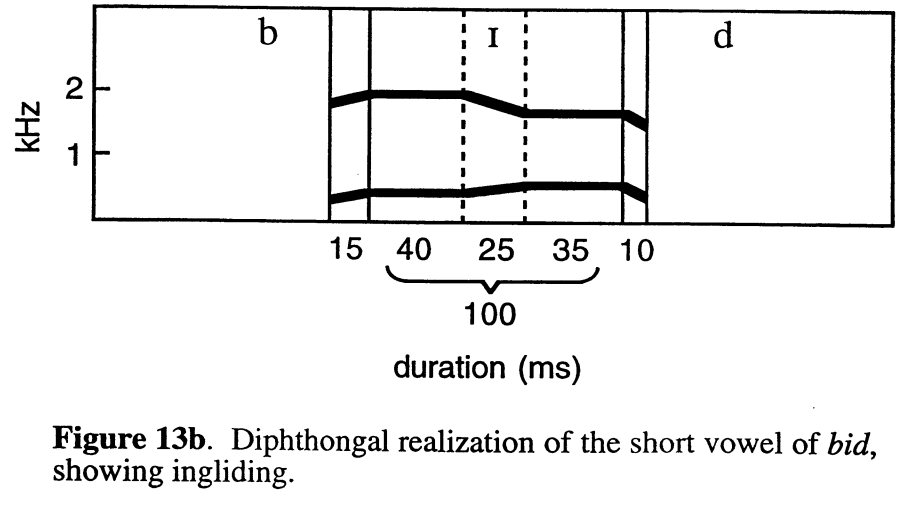


Figure 13b. Diphthongal realization of the short vowel of *bid*, showing ingliding.

We have segmented critical intervals in this figure with solid lines, delimiting phone-transition boundaries, and dashed lines, delimiting intervals within the vowel phone. It will be noted that the vowel [ɪ] has two steady state portions. The formant values of the first, 40-ms-long steady state are typical of those of a high front lax vowel, while those of the second, 35-ms-long steady state are typical of those of an upper mid central vowel. Between these steady state portions lies an interval of 25 ms. Note that while this interval is an interpolation much like the transitions preceding and following the [ɪ], it differs from these in lying entirely inside a phone. We call such an interval an *internal* transition. Later we will show that internal transitions have different durational properties from external transitions (transitions between phones), such as those linking [b] to [ɪ] and [ɪ] to [d] in this example. We will see that these differences can be used as a criterion for distinguishing between one-phone and two-phone nuclei.

In the past, some researchers have viewed these short "diphthongal" vowels as having only a single target portion, corresponding to what we have called the initial steady state, with the following formant movements constituting a single "offglide", or transition to the stop (see, for example, Lehiste and Peterson 1961). However, inspection of many examples shows that this analysis cannot be maintained. First, it cannot explain the existence of the second steady-state portion which frequently appears at the end of such vowels, as shown in Figure 13b. Second, a long final transition would be highly unusual between a front vowel and an alveolar or labial stop—contexts in which transitions are generally short. (Transitions are short between front vowels and alveolar stops due to the small articulatory distance between them, and they are short between front vowels and labial stops due to the fact that adjacent tongue and lip articulations tend to overlap.) Third, the rising F1 pattern following the first target of [ɪ] (also visible in the spectrograms reproduced in Lehiste and Peterson 1961, p. 269) cannot be explained as a transition to the following consonant, given that labials and alveolars have inherently low F1 targets; this pattern can only be due to a second formant target within the vowel. Furthermore, close examination of spectrograms often reveals a short, 5-to-10-ms-long transition between the end of the vowel phone and the beginning of the consonant, as shown in Figure 13b, which cannot be explained under a one-target analysis.

Yet another problem with a one-target analysis stems from the fact that other speakers do not realize these vowels with inglides but as steady-state phones, as shown in Figure 13a earlier. In such realizations, we find that the duration of the single steady state tends to be equal to the sum of the duration of the two steady states plus the internal transition in typical "ingliding" realizations, and that moreover, the transition to the following consonant has the same duration as the transition following the second target of the "ingliding" realization (compare Figure 13a and Figure 13b). If we were to adopt the one-target analysis, we would have to recognize a significantly longer post-[ɪ] transition in Figure 13b than in Figure 13a; elsewhere, however, we have found that the durations of external transitions in any given context do not vary greatly from one speaker to another. As a final point, we have observed that GAE speakers who use "ingliding" realizations in real words also tend to use them when asked to utter the same vowels in isolation, while speakers who use single-target realizations in real words also generally do so in isolation.

How can short contour vowels be modeled in a multi-tiered representational system? In most "ingliding" vowels, one or both of the F2 targets exhibit some duration. However, we have found that speakers are not consistent as to how they distribute the duration among the vowel targets and the transition between them, even from one token of an utterance to the next. In fact, in some utterances one or the other of the targets has no duration at all, appearing simply as an "inflection point" between transitions. A realization with two zero-duration targets is schematized in Figure 14.

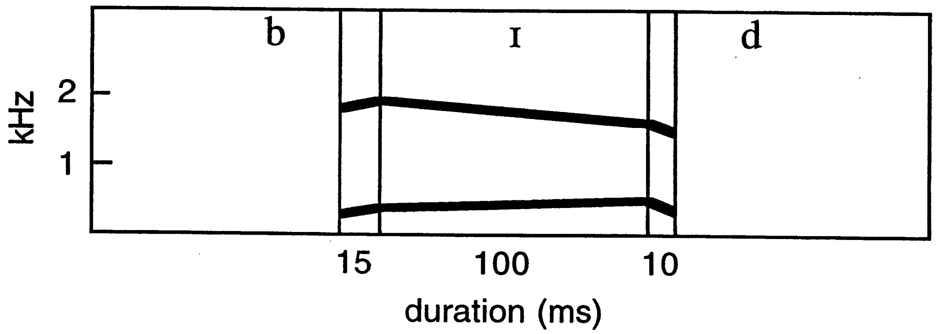


Figure 14. Diphthongal realization of the short vowel of *bid*, with two zero-duration targets.

Such variability is not surprising. In our experience, the realizations in Figure 13b and Figure 14 are perceptually indistinguishable as long as the total duration of [ɪ] remains the same. Given the criterion established earlier (section 3.2.3), we may abstract away from this variability in our acoustic scores. Since from a perceptual point of view we can assign short contour vowels durationless initial and final formant targets without affecting their intelligibility or naturalness in any way, we may represent all such targets as durationless in our scores.[12]

In accordance with our earlier analysis of contour phones (section 3.2.4), then, we will assign contour vowels three duration values: two zero-ms values at each edge to serve as "placeholders" for target values, and a non-zero duration for the internal transition between them. Figure 15 shows a representation of the duration and F2 tiers for an ingliding pronunciation of *bid* [bɪd].

```
root tier:        b           i           d
                  |         / | \         |
duration tier:    x   15   0  100  0  10  x
                  |        |       |      |
F2:               y      1960     1780    y
```

Figure 15. Partial representation of diphthongal *bid*, with two zero-ms target durations.

We might briefly consider an alternative analysis in which the "ingliding" vowels are represented as contour segments in the phonology, rather than in the phonetics. In such an analysis, these vowels are represented with *two* root nodes linked to a single skeletal position, as has sometimes been proposed for the representation of affricates or prenasalized stops (see Clements and Keyser 1983 for such an analysis of short diphthongs in Spanish). Under this analysis, the ingliding [ɪ] would be represented with two root nodes representing the sequence [ɪə] linked under one skeletal unit, as shown in Figure 16.

```
skeleton:           X
                   / \
root tier:        I   ə
                  |   |
duration tier:    x   y
```

Figure 16. Alternative representation of *bid*, with two root nodes (x, y = any values)

Here it is a skeletal unit that branches, instead of the root node. While this representation is conceivable, there is no independent motivation for a two-segment analysis of short vowels

[12] In our work on synthesis, we have found that not only [ɪ], but all contour vowels can be realistically modelled with two zero-ms targets, one at each edge. This analysis allows a desirable simplification of our representations, though it is not crucial to the analysis of GAE vocalic nuclei proposed below.

in the phonological system of English. For example, the ingliding short vowels do not show the phonological "edge effects" characteristic of phonological contour segments such as prenasalized stops, or satisfy any of the other criteria for a two-segment analysis outlined in section 4. We will also see that such an analysis is contradicted by generalizations at the phonetic level (section 5.4). For these reasons, we maintain the phonetic analysis of short "ingliding" vowels shown in Figure 15.

5.3 Durational properties of D-nuclei

We next consider the durational structure of the D-nuclei. We have so far implicitly assumed that the D-nuclei [ay aw] of words like *tide* and *profound* consist of two phones (i.e. root nodes), a structure consistent with most phonological analyses (cf. (3)). Let us now consider some facts that support this analysis from a phonetic point of view. We will focus our discussion on [ay] due to the fact that its formant movements are more prominent than those of [aw], allowing a more consistent segmentation.

D-nuclei behave differently from other types of nuclei in several respects. A first difference concerns their strong tendency to preserve their diphthongal structure in all contexts. Let us use the term "edge values" to refer to all formant target values aligned with the beginning or end of any phone. We find that D-nuclei preserve different edge values at their beginnings and ends across different contexts and speech rates, even when they are relatively short in duration (see also Pike 1947). In this respect D-nuclei stand in contrast to SV-nuclei like [ɪ] as well as to LV-nuclei like [e] (to be discussed below), all of which readily simplify to monophthongs in certain contexts. The robustly two-target structure of D-nuclei follows directly from a two-phone analysis, since as a general rule each phone in a string is characterized by a set of independent formant targets.[13]

Second, in any given word and for any given speech rate, each of the phones of a D-nucleus has a relatively stable duration which does not vary substantially from one instance of an utterance to the next for a given speaker. This behavior stands in contrast to that of the two other types of nuclei. In the preceding section we observed highly variable formant target durations in the case of ingliding SV-nuclei like [ɪ], and we will see that the same is true for LV-nuclei like [e]. The durational consistency of D-nuclei lends further support to the notion that these nuclei consist of two phones, since separate phones generally

[13] However, laryngeal glides such as [h] generally do not show formant targets of their own (Keating 1988, Hertz 1990). This fact follows from a phonological analysis in which they fail to bear place nodes (Clements 1985, Steriade 1987). For this and further reasons, we treat laryngeal glides not as phones, but as aspirated transitions between phones (Clements and Hertz 1991).

exhibit relatively predictable durations in any given context.

The most compelling evidence for a two-phone analysis of D-nuclei, however, comes from a comparison of the durational behavior of vocalic nuclei in different segmental contexts. It is well-known that stressed vocalic nuclei are longer before voiced consonants than before voiceless ones in many contexts (see e.g. Peterson and Lehiste 1960, Chen 1970). Thus, for example, in phrase-final position, the nuclei of the second members of the following pairs are longer than those of the first:

(8)	bit	bid
	bat	bad
	bait	bade
	bite	bide

However, D-nuclei behave differently from the other types of nuclei in that lengthening is distributed across them in unequal fashion, affecting their steady states but not the transitions between them (Hertz 1991, 1992). Comparing the [ay] of *bide* with that of *bite*, for example, we find that the lengthening affects both steady states, the first one (that of [a]) somewhat more than the second one (that of [y]). In contrast, we find little or no lengthening in the intervening transition.

These observations are illustrated in Figure 17, which displays typical F2 patterns of the nuclei of *bite* and *bide* as uttered in the frame *Say* __ *for me*.[14] It will be noticed in these examples that while the [a] nearly quadruples its duration in *bide*, the transition between [a] and [y] has the same duration (75 ms) in both words. We also observe that the [y], though relatively short to begin with, doubles its duration. The lengthening of [y], though less dramatic than that of [a], appears to have some perceptual importance: in informal synthesis we have observed that even a 10-ms lengthening of the [y] helps to cue the [+voice] feature of the following consonant.

[14] This diagram, and all those which follow, are based on durations typical of the speech of our primary subject (SRH).

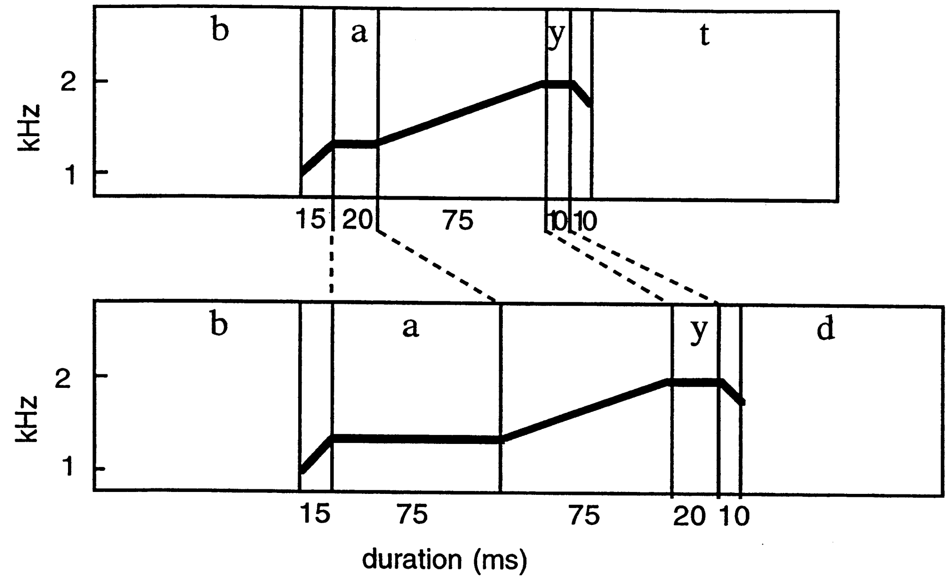


Figure 17. The F2 patterns of *bite* and *bide* compared, showing durational differences.

The lengthening behavior of D-nuclei, as just summarized, stands in contrast to the behavior of other nuclei showing rising or falling formant values throughout most of their duration. The phones of SV-nuclei in “ingliding” pronunciations of words like *bit* and *bid*, for example, often lengthen and shorten as a unit, with not only the initial and final formant targets but also the internal transition between them participating in the lengthening. (We discuss the behavior of LV-nuclei, which is analogous, below.) These differences can easily be accounted for by analyzing D-nuclei as two phones and SV-nuclei as one; we can then express the lengthening rule as one that applies to all phones in the nucleus, to the exclusion of transitions. (We discuss the lengthening principles in more detail in section 5.6.1.)

As shown in Figure 17, not only the transition between [a] and [y] but also the C-to-[a] and [y]-to-C transitions are stable in duration.[15] In general, we find that in stressed

[15] Figure 17 (and Figures 18-9 following) show the final transition before a voiceless obstruent as identical in duration to the final transition before the comparable voiced obstruent. In fact, as noted by Summers 1987, such transitions often appear to have shorter durations in spectrograms. The explanation for this fact is that transitions often tend to be devoiced toward their end before voiceless obstruents. Klatt has suggested that this devoicing may result from “the natural tendency to make a slightly early glottal opening gesture for a postvocalic voiceless consonant in order to ensure that no low-frequency voicing cue is generated during the obstruent” (Klatt 1976, 1214). Support for this interpretation comes from the fact that formant structure can sometimes be observed to persevere during the voiceless portion of the

syllables spoken in a normal, conversational speaking rate, external transitions do not vary in duration as a function of the voicing of the following consonant, while phones (and internal transitions) often show considerable variation. We call this effect the ***stable transition phenomenon*** (Hertz 1991). This effect provides strong support for an analysis of D-nuclei as two phones and SV-nuclei as one. The relative stability of (external) transitions is also evident when syllables are intentionally exaggerated in length; external transitions tend to lengthen only a little, while phones (and internal transitions) lengthen much more.

Further evidence for a two-phone analysis of D-nuclei comes from the durational behavior of vowel-plus-liquid sequences such as [ɛl] in *felt* and *felled.* We find that they exhibit a stable transition effect quite parallel to that of [ay] (Hertz 1991). Typical F2 patterns of *felt* and *felled* as spoken in the frame *Say* ___ *for me* are shown in Figure 18.

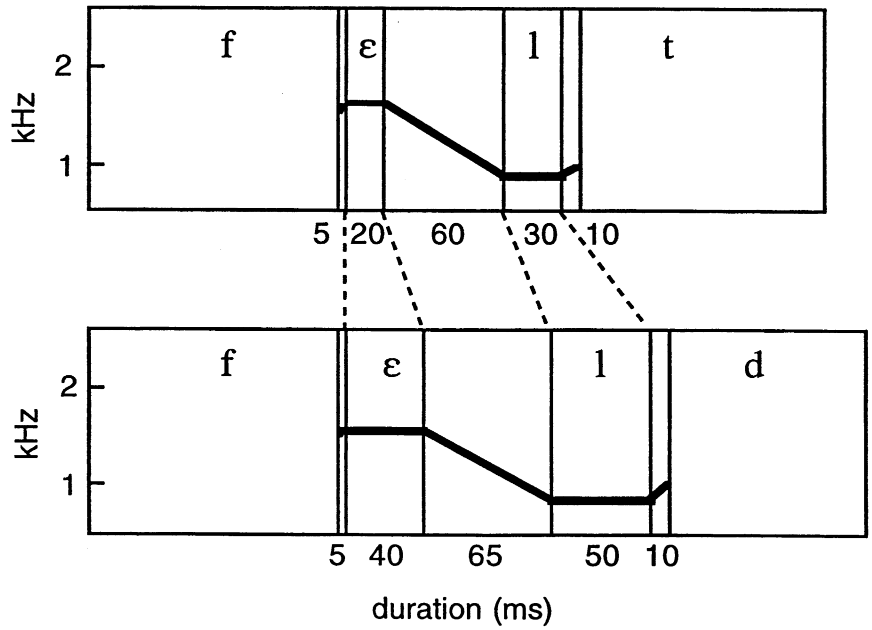


Figure 18. F2 pattern of [fɛlt] and [fɛld], showing lengthening of both [ɛ] and [l].

transition, and that the voiced part of an apparently shortened transition systematically fails to reach the formant targets of the following obstruent, as shown in detail in Summers' data. (See Hertz 1991, 103-5 for further discussion.) For these reasons, synthesis rules modelling this phenomenon must include a rule devoicing preconsonantal transitions toward their end before voiceless consonants. In order not to introduce complexities irrelevant to our main points, our representations abstract away from this devoicing rule.

This observation provides very strong evidence for a two-phone analysis of the D-nuclei. It is uncontroversial that a vowel-plus-liquid sequence like [ɛl] constitutes two phonological segments, and hence two phones (phonetically interpreted root nodes) in the acoustic score. If we are to capture the fact that D-nuclei exhibit the same durational behavior, we must analyze them in the same way.

5.4 Durational properties of LV-nuclei

Let us now consider the formant structure of LV-nuclei. Our goal is to determine whether these nuclei are best modeled with two phones, as in the case of D-nuclei, or with one phone, as in the case of SV-nuclei.

First of all, we have found that while LV-nuclei are frequently diphthongized, they are often realized with steady-state formant structure throughout, especially in durationally short contexts (see also Pike 1947). For example, it is not unusual for a speaker to use a diphthongized realization of [o] in the word-final syllable of words like *evoke* but a steady-state realization of [o] in the shorter, unreduced non-final syllables of words like *vocation, evocation.* This behavior does not resemble that of the robustly diphthongal D-nuclei, as noted above. Moreover, some speakers of GAE regularly monophthongize the high LV-nuclei [i u] (Lehiste 1964) and even the mid vowel [o] (Peterson and Coxe 1953) in all contexts. (Monophthongized realizations of all four LV-nuclei are of course widely found in dialects outside GAE (e.g. Irish, West Indian), while the D-nuclei are consistently realized as diphthongs (Wells 1982).)

Second, again unlike D-nuclei (but like SV-nuclei), we find that the initial and final formant targets of LV-nuclei are variable in duration within and across GAE speakers, even when produced in the same word and at the same speech rate, and are often realized with no appreciable duration (i.e. no steady state) at all. Just as for SV-nuclei (but unlike D-nuclei), we have found that we can model LV-nuclei with zero-duration formant targets at their beginnings and ends with virtually no perceptual consequences.[16]

[16] See also Bond 1982, who reports that LV-nuclei are reliably identified in test stimuli even if their steady-state portions are short or absent. Our description of the LV-nuclei is, however, somewhat different from that of Lehiste and Peterson 1961. These authors examined CVC monosyllables spoken by six speakers. For what we here term D-nuclei, they found two well-defined targets, just as we have, but for the LV-nuclei [e o] they reported only one target, occurring toward the end of [e] and the beginning of [o]. This description is not actually at variance with ours, however, since these writers further observed that [e] typically showed an early inflection point in the F2 curve, located at about 1800 Hz, while [o] showed a late inflection point, with a value of about 800 Hz; these inflection points correspond to what we call initial and final zero-duration targets, respectively. (Lehiste and Peterson did not describe

Third, LV-nuclei lengthen as a single unit in lengthening contexts. When we examine the formant patterns of pairs like *bait* and *bade*, we find that the extra length in *bade* generally involves all portions of the phone, including the relatively long internal transition between the initial and final formant target values. Given that we can model LV-nuclei with zero-duration formant targets, the lengthening pattern can be represented as in Figure 19, which models the F2 patterns of *bait* and *bade* as found in the context: *Say ___for me.*

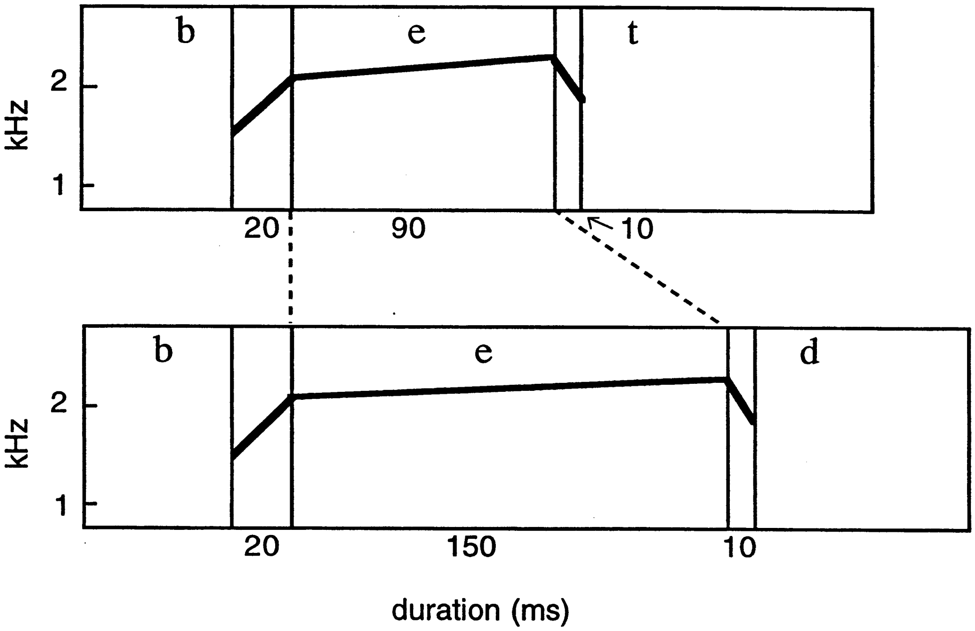


Figure 19. F2 patterns of *bait* and *bade* compared, showing durational differences.

Thus the LV-nuclei do not exhibit the stable transition effect, as far as their internal transition is concerned. In this respect they contrast with D-nuclei (see Figure 17), but

these inflection points as targets, since they reserved the term "target" for steady-state formant patterns lasting for at least 20 msec.)

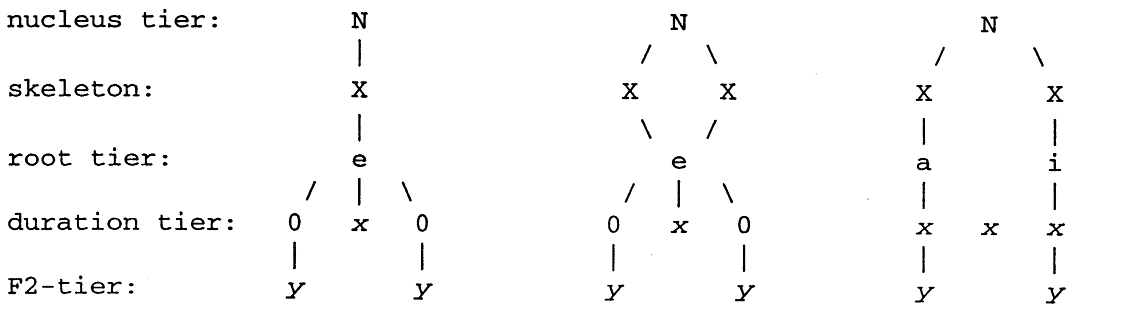
In a study of a New York City dialect, Gay (1968) generally found two steady-state targets in these vowels in various CVC contexts, in both slow and moderate speech rates, though one or both steady state targets were often very brief, and sometimes absent (i.e., durationless); see his Tables I and III. Some of the GAE speakers we have observed resemble Gay's in this respect.

resemble SV-nuclei.[17]

We are now in a position to return to a question left open at the end of our discussion of short "ingliding" SV-nuclei, when we briefly considered the possibility of representing these nuclei as contour segments (i.e. two sequenced phones) at the phonological level, as shown in Figure 16. We now see that this type of representation would incorrectly predict that these nuclei should resemble those of D-nuclei in their durational behavior, while in fact they pattern similarly to LV-nuclei.

5.5 The representation of vocalic nuclei

We can now bring our phonetic analysis to bear on the phonological analysis discussed earlier (section 4). There we cited phonological evidence supporting the view that SV-nuclei are linked to one skeletal slot and LV- nuclei to two in phonological representation. In this section we have reviewed phonetic evidence showing that SV- and LV-nuclei behave as single phones (phonetically interpreted root nodes), while D-nuclei consist of two. Putting these results together, we arrive at the composite phonological/phonetic representations of all three nucleus types shown in Figure 20. (In this figure, *x* and *y* are variables over appropriate duration and formant values, respectively.)



a. SV-nucleus [ɛ] b. LV-nucleus [e] c. D-nucleus [ay]

Figure 20. Representations of the three types of vocalic nuclei.

This figure combines several phonological tiers (the nucleus, skeleton, and root tiers)with

[17] These results are similar in large part to those reported in Gay's 1968 study of a New York City dialect, mentioned in the previous note. Gay found, as we did, that the longer duration of [e o] before voiced obstruents was reflected primarily in changes in transition duration, and that initial and final steady-states were not even present in many cases. In contrast, the longer duration of [ay ɔy] in the same context was accomplished primarily by a lengthening of the first steady-state target.

several phonetic tiers (the duration and F2-tiers). These representations summarize our analysis. They assign SV- and LV-nuclei identical structure at the level of the duration tier, accounting for their parallel durational behavior, and LV- and D-nuclei identical structure at the level of the skeleton, accounting for their parallel phonological behavior.

We can now replace the mnemonic labels SV-nuclei, LV-nuclei, and D-nuclei with the more straightforward terms *short vowels*, *long vowels*, and *diphthongs*, respectively. These terms, as we shall use them, are defined by the representations in Figure 20a, 20b, and 20c, respectively. Since these representations integrate phonological and phonetic information, these definitions hold for both the phonological and phonetic levels of analysis. A short vowel at either level of analysis is defined as a syllable nucleus having a single timing unit and a single root node, a long vowel as a nucleus with two timing units linked to one root node, and a diphthong as a nucleus consisting of two timing units linked independently to two root nodes.

Other phonetic analyses of GAE long vocalic nuclei that have appeared in the literature are mostly consistent with our account, as we have remarked in footnotes. However, some phoneticians have occasionally preferred a one-phone analysis of diphthongs like [ay]. For example, Lehiste (1964) and Gay (1970) have maintained that the analysis of [ay] as a phoneme sequence / ay / is contradicted by the observation that the formant values of the steady state found at the end of this nucleus are not strictly identical to the allophones of / y / found other contexts, such as prevocalic position. This phonetic observation is correct, as will be confirmed by the values for [ay] and [y] that we give in the next section. However, as these writers acknowledge, phonemes often differ in realization according to context, and prevocalic [y] and the final target of [ay] can plausibly be analyzed as contextually-determined realizations of the phoneme / y /.

Gay (1970) advances a further argument against a two-phone analysis of [ay ɔy aw]. He cites perceptual tests suggesting that the presence of steady state targets may not be essential to the identification of these diphthongs. He also finds that they do not always have steady states of more than 20 ms in his acoustic data, an observation again confirmed by our own data (see e.g. Figure 17 for the diphthong of *bite*). However, it is not unreasonable to expect that vocoids may shorten in vocoid clusters, just as consonants do in consonant clusters. In the next section, we will interpret this shortening effect as a special case of a more general principle which predicts (among other things) that all else being equal, a vowel will be shorter in a diphthong than it is in a one-vowel nucleus of similar duration.

5.6 The assignment of phonetic values

In order to complete our analysis, we will now consider the manner in which actual acoustic scores appropriate for individual speakers can be constructed within the integrated representational system (IRS). The following discussion draws upon a complete set of speech synthesis rules currently under development by Hertz and her associates, which generate durations and acoustic values for any arbitrarily-selected GAE sentence. These rules are designed to simulate the speech of a young adult male GAE speaker having no notable idiosyncracies. (Voices of other types of speakers, including women and children, can be simulated by filters that modify these values appropriately.) We first consider duration assignment (section 5.6.1), and then turn to the specification of formant values (section 5.6.2).

5.6.1 Duration values

As is well known, vowel duration is influenced by many interacting factors. Among these we can distinguish *intrinsic* factors such as phonological length and height, and *extrinsic* factors such as the presence vs. absence of stress, features of neighboring consonants (voicing, aspiration, place of articulation, degree of stricture), speaking rate, and various syntactic and semantic considerations such as position in the sentence, word familiarity, etc. (for a general review see Klatt 1976). Since vowel duration is a highly gradient phenomenon, it cannot be assigned in terms of a binary phonological feature [±long] but must be specified at the phonetic level.

In a model distinguishing phones and transition, the unit that proves to be most appropriate for stating duration rules is the syllable *nucleus*. It will be recalled that at the phonological level, the syllable nucleus includes the syllabic peak and a following tautosyllabic glide, if present. For the purposes of duration assignment, the nucleus must be expanded in two respects. First, it must incorporate all voiced transitions adjacent to a nuclear vowel or glide. This means, for example, that the acoustic nucleus of a word like *tide* contains not only the phones [a] and [y], but also the voiced transitions on either side of the [y]. On the other hand, it excludes the voiceless aspirated transition following the [t]. Second, in order to account for the fact that liquids lengthen before tautosyllabic voiced obstruents in much the same way that vowels and glides do (see section 5.3, and below), we must allow the acoustic nucleus to embrace a following tautosyllabic liquid, if one is present.

The acoustic nucleus, defined in this way, corresponds to the high-sonority "core" of the syllable and is coextensive with the maximum uninterrupted stretch of voiced sonorance (or “spontaneous voicing”, in the sense of Chomsky and Halle 1968).[18] A partial acoustic representation of *tide*, including the nucleus, is shown in Figure 21.

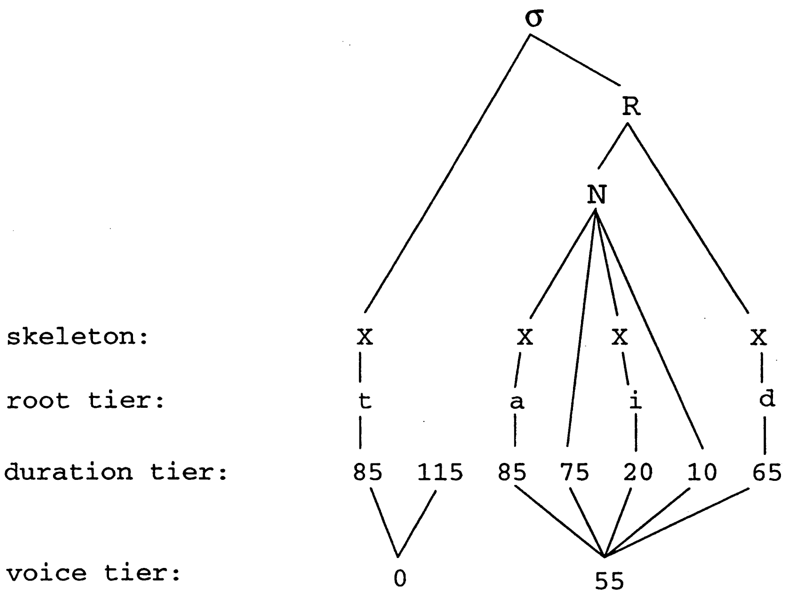


Figure 21. Partial acoustic representation of *tide*, showing the expanded nucleus.

The phonetic duration of any given nucleus can be obtained from its *base duration value* by applying a set of duration adjustment rules reflecting extrinsic contextual factors of the type mentioned above. The base duration value reflects the expected duration of a given nucleus if no contextual factors apply to it. As our rules are presently formulated, the base duration of a given nucleus is derived from the average value it has in stressed CVC monosyllables of the form *C__t,* as spoken at normal conversational speed in a frame such as *Say __ for me*. (This choice receives some preliminary motivation from our observation that English nuclei show more consistent durations before voiceless stops than before voiced ones across dialects, though further study is needed to verify how widely this generalization holds.) Some sample base durations for the speaker modeled in our rules (rounded off to the nearest 10 ms) are given below:[19]

[18] We do not discuss the question of whether nasals should be included in the syllable nucleus or not. We have found the acoustic evidence to be ambiguous in this regard, lending support to both views.

[19] These values are typical of a moderate speaking rate, and must be adjusted for faster or slower tempos.

(9) Base duration values for selected stressed vocalic nuclei:

[i]:	100 ms
[ɪ]:	70 ms
[e]:	120 ms
[a]:	130 ms
[ay]:	130 ms

In general, for our modeled speaker as for many others, nuclei with phonologically short vowels are shorter than nuclei with the corresponding long vowels (e.g. [ɪ] vs. [i]), those with high vowels are shorter than those with corresponding mid vowels (e.g. [i] vs. [e]), and those with mid vowels are shorter than those with corresponding low vowels (e.g. [e] vs. [a]), when compared in similar contexts. Base durations for diphthongs are observed to be similar to those of long mid vowels and low vowels, as is shown by the values for [e a ay] given here and elsewhere in the literature.

The base duration of any given nucleus is modified by rules applying in particular contexts. Thus, for example, stressed vowels are lengthened before voiced stops, in certain contexts. This lengthening is subject to a principle of maximum duration that prevents nuclei containing two or three phones from lengthening beyond a certain "ceiling" value, so that the total lengthening for longer nuclei may be proportionately somewhat less than that of shorter ones. Illustrative examples of phrase-medial lengthening rules for our modeled speaker are given in (10) below.

(10) Phrase-medial Lengthening:

a. lengthen a stressed word-final nucleus 1.5 times before a tautosyllabic [+voice] obstruent, not exceeding a limit determined by the number of phones in the nucleus;

b. lengthen a stressed word-final nucleus 1.2 times before a tautosyllabic fricative.

The effect of the first rule has been shown in Figures 17-19. Note that these rules apply cumulatively. Consequently, both apply before a voiced fricative in words like *ties*, lengthening the nucleus by a total of 1.8 times its base duration. Since the final duration of any nucleus is the product of several interacting rules of this sort, it would be redundant to list all durations found in all contexts; it is simpler to predict the durations by rule.[20]

[20] Speakers vary in the amount of lengthening they apply in the contexts specified in (17), and some

Once a duration has been assigned to the nucleus as a whole, it must be distributed among its constituent phones and transitions. The distribution is accomplished, in our current implementation of the IRS, in accordance with a principle of give-and-take that we call the *Tax Law*. As we have already observed, transitions tend to have relatively fixed, predictable durations between any two given segments, depending mainly on their place of articulation and degree of stricture (see Lehiste and Peterson 1961 for examples); their duration can therefore be taken as a constant. In contrast, the duration of nuclear phones covaries with the duration of the nucleus as a whole, expanding as the nucleus expands and contracting as the nucleus contracts. Our study of [ay] shows that at least for this diphthong, the vowel and the glide phones do not behave identically. While the glide duration always appears to represent a fixed proportion of the nucleus duration (representing about 10% in the case of [y]), the duration of the vowel varies inversely with the combined durations of the other members of the nucleus, all else being equal. We may calculate the duration of the vowel to a close approximation by first calculating the duration of the nucleus as a whole in the way shown above, then calculating the duration of any glide or liquid in the nucleus, and finally subtracting the combined durations of the transitions, glides and liquids from the total duration of the nucleus.

As illustration, consider the case of a nucleus containing only the short vowel [a]. According to table (9), the base duration of such a nucleus as it appears in a word such as *dot* is 130 ms. Recall that this is the duration assigned to the nucleus as a whole, including the adjacent voiced transitions. In a syllable such as *Dodd* [dad], shown at the top of Figure 22,[21] this value is increased by a factor of 1.5 by Phrase-medial Lengthening (10a), giving a total of 195 ms. Since both transitions in this syllable are voiced, both belong to the nucleus. In this particular case, the transitions have a duration of 60 and 35 ms, respectively. By the Tax Law, then, the duration of the vowel [a] is equal to the total duration of the nucleus minus the sum of the durations of its two transitions, yielding a value of 100 ms. Using the tax-law analogy, we may think of the total nucleus duration (d_N) as the vowel's "gross income", the durations of the transitions ($d_{t1} + d_{t2}$) as its "taxes", and the remainder (d_V) as its "net income". (We discuss the second form, *Tod*, below.)

display virtually no phrase-medial lengthening at all. We emphasize that our rules model the behavior of one representative speaker of GAE, and that other speakers may use somewhat different rules.

[21] We point out that in some of the F2 displays to be discussed below, the values in the displays, reflecting typical duration values for SRH, are not identical in all respects to those predicted by our rules. This is because the rules presented in our text are a small subset of those needed to model all observed regularities.

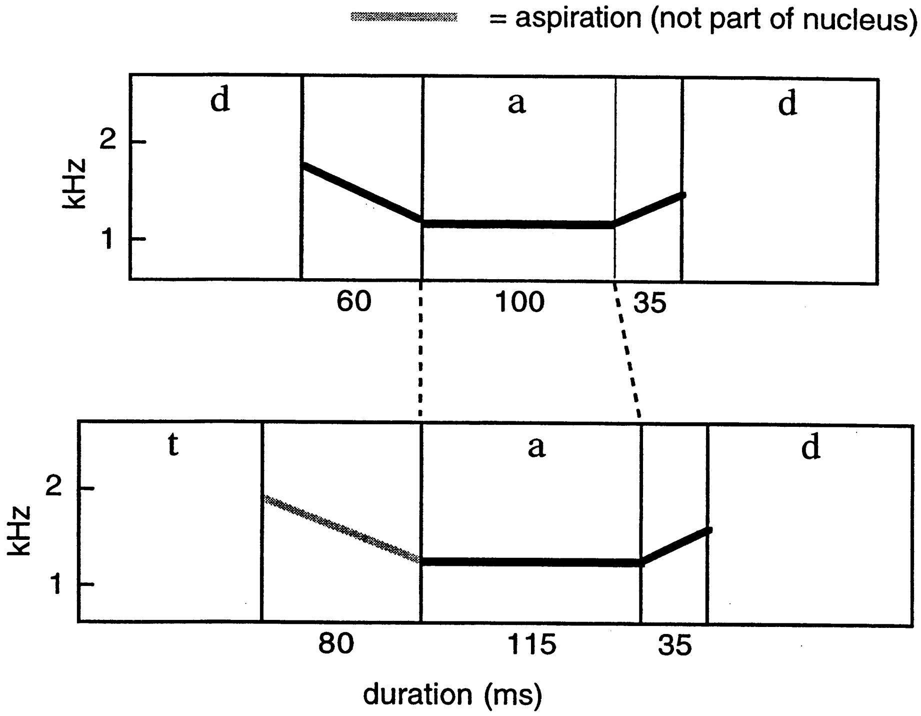


Figure 22. F2 patterns of *Dodd* and *Tod.*

The Tax Law, as stated above, makes three predictions: all else being equal, 1) vowel lengthening is disproportionately greater than nucleus lengthening, 2) vowel duration varies inversely with transition duration, and 3) vowel duration varies inversely with the number of other segments (phones, transitions) in the nucleus. We will now examine each of these predictions in turn and show that they are confirmed in representative examples.

The first prediction can be tested by considering the durational difference between the vowel [a] in *bite* and the same vowel in *bide*, as displayed in Figure 17 above. In *bite*, the total duration of the nucleus is 130 ms, but the vowel [a] takes up only 20 ms of it. This is because the durations of the nonsyllabic segments (glide, transitions) total 110 ms, leaving only 20 ms for the vowel. In *bide*, the duration of the nucleus increases to 195 ms due to Phrase-medial Lengthening (10a), representing an increase of 1.5. Once the total duration of the nonsyllabic segments (120 ms) has been deducted from this nucleus, a net duration of 75 ms is still left for the vowel, representing an increase of 3.75. Here the increase in vowel duration is greatly disproportionate to the increase in nucleus duration. (Note that in both cases, [y] receives roughly 10% of the total nucleus duration, causing its length to approximately double in *bide*.)

We can explain the behavior of vowel + liquid sequences in the same way. We have already discussed the pair *felt* and *felled* shown in Figure 18, and noted that both phones in the sequence [ɛ l] lengthen considerably more before [d] than does the transition between them. Again, the vowel lengthens disproportionately to the liquid as well as to the nucleus as a whole. This behavior is accounted for on our assumption that liquids are incorporated into the expanded nucleus at the phonetic level. Since the sequence [ɛ l] (together with the adjacent transitions) constitutes a nucleus, it undergoes the nucleus lengthening rules of (10), and under the Tax Law, the extra length is assigned to the vowel after the durations of the other segments of the nucleus have been deducted.[22]

The Tax Law further predicts that vowel duration varies inversely with transition durations, all else being equal. This means that within a given nucleus, varying only the identity of a neighboring consonant (and thus of their neighboring transitions), the duration of the vowel increases as the total duration of the transitions decreases, and vice versa. This prediction is also confirmed by our observations. For example, we find that the transition from [w] to a following vowel is typically much longer than the transition following [b]; for speaker SRH, for example, the transition from [w] to [e] has a typical duration of about 70 ms, while the transition from [b] to [e] has a duration of about 20 ms. Figure 23 shows the minimal pair *bait* [bet] and *wait* [wet], as spoken by SRH in the frame *Say___for me*. We observe that the [e] of *wait* is shorter than that of *bait* by an amount equal to the difference between the two transition durations. This follows from the Tax Law: given that the duration of the final transition is identical in both cases, the longer duration of the initial transition in *wait* must be compensated for by the shorter duration of the following vowel.

[22] We have observed similar durational behavior in tautosyllabic VGL (vowel + glide + liquid) sequences, suggesting that these, too, should be treated as forming a single nucleus. For example, we find that the extra duration of [ayl] in *whiled* as compared to *whilst* occurs in all three phones, [a], [y], and [l], while the durations of the transitions are relatively constant.

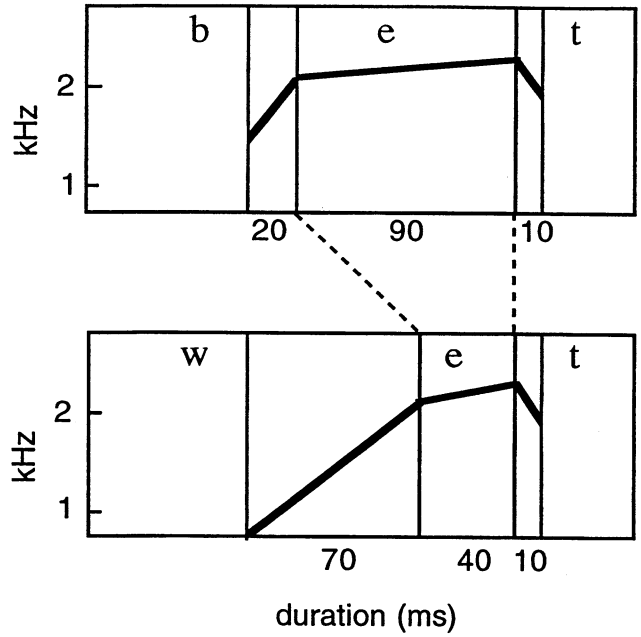


Figure 23. F2 pattern of *bait* and *wait*, showing trading relations between the vowel and the preceding transition.

The third prediction is that vowel duration varies inversely with the number of other segments (phones, transitions) in the nucleus. This prediction can be tested by comparing pairs of words differing only in that one has a nonsyllabic phone which is absent in the other. Compare, for instance, the nucleus of *tide* [t^hayd] (Figure 5) with that of *Tod* [t^had] (Figure 22). Both nuclei have an identical base duration and are subject to Lengthening (10a). The duration of the vowel [a] in *Tod* is derived by deducting the duration of the final transition from that of the nucleus (recall that the first transition is excluded from the nucleus as it is unvoiced). The duration of the same vowel in *tide*, however, is derived by deducting the durations of three segments: the medial and final transitions and the glide [y]. Since their combined duration (105 ms) is greater than that of the single transition of *Tod* (35 ms), the net duration remaining for the vowel is shorter, even though the overall nucleus duration of *tide* is actually longer.

Another way of testing this prediction is to compare pairs of words differing only in that one has an aspirated transition where the other has a voiced transition. Since the aspirated transition does not form part of the nucleus, its duration is not deducted from that of the nucleus, and we expect the vowel to be longer. This prediction is somewhat more difficult to confirm due to the interaction of a further regularity, namely that vowels tend to be shorter than expected after aspirated transitions. This regularity is visible in the comparison of *Dodd* and *Tod* in Figure 22, where the 115-ms duration of the vowel in *Tod* is less than what we would have expected on the basis of the principles established so far. (Figure 22 also illustrates the fact that aspirated transitions tend to be longer than corresponding unaspirated ones.) In spite of this, we find that the [a] of *Tod* is *still* longer than the [a] of *Dodd*, a fact which follows again from the Tax Law.

We stress the importance of the nucleus in obtaining these results. A model which did not recognize the nucleus and which assigned durations to segments alone would require additional, arbitrary rules shortening vowels before tautosyllabic glides and lengthening vowels after aspiration. In the present framework, in contrast, these effects follows from a single, independently-needed principle, the Tax Law.

The issue of duration assignment in speech is extremely complex. As pointed out earlier (note 21), the principles outlined here are a small subset of those that are required to predict durational patterns in their full complexity across the full range of GAE speech (see e.g. Hertz 1992 for discussion of some of the additional factors involved in duration assignment). Furthermore, we remind the reader that the specific durations and durational adjustments given in this section are drawn primarily from the study of a single speaker, and may vary in detail for other GAE speakers that we might have chosen to model. Similar remarks hold for the formant values to be discussed in the next section.

5.6.2 Formant values

Like duration values, formant target values are assigned to phones on the basis of their phonological feature content and contextual factors. It is well known that different speakers of a language differ in terms of the range of formant values characterizing their speech. Much of this variation can be explained by differences in the size and shape of different speakers' vocal tracts. Such variation is not without limit, however, and is constrained by the phonetic definitions of the features characterizing their phoneme inventory. These definitions require, for example, that a [+open] vowel must have a higher F1 value than an otherwise similar [-open] vowel in a similar context, for any given speaker. Given these constraints, we speculate that cross-speaker differences in formant values may be partly predictable in terms of the global "acoustic space" used by each speaker in producing vowels. We conceive of the *acoustic vowel space* of a given speaker as the *n*-dimensional space defined by the total range of frequency variation used by that speaker in producing each formant (F_1, F_2, ... F_n). The target values characterizing any particular vowel or glide for that speaker must be appropriately situated within this space, consistently with the segment's feature definition.

Representative target values for selected GAE vowels as produced by the male speaker modeled in our synthesis rules are shown in (11). These values are typical of those found in lengthening contexts in which local coarticulatory effects are minimal, such as when these vowels are pronounced in the phrase-final context *Say h__d.* (Vowel symbols in brackets stand, as usual, for root nodes in feature trees.)

(11) Illustrative F1 and F2 target values for selected GAE vowels (in Hz):

		F1	F2
[i]:		320	2160
[ɪ]:	left edge:	440	1800
	right edge:	440	1650
[e]:	left edge:	470	1760
	right edge:	350	2120
[a]:		780	1200

These base values are adjusted to account for contextual effects. For example, the formant values of [y] in the diphthong [ay] are more compact than those in other contexts, reflecting a lower and more centralized articulation, as shown in (12):

(12) Illustrative F1 and F2 values for [y]:

	F1	F2
[y] in the context [a __]$_{\mathrm{Nuc}}$:	400	1800
[y] (in many other contexts):	300	2100

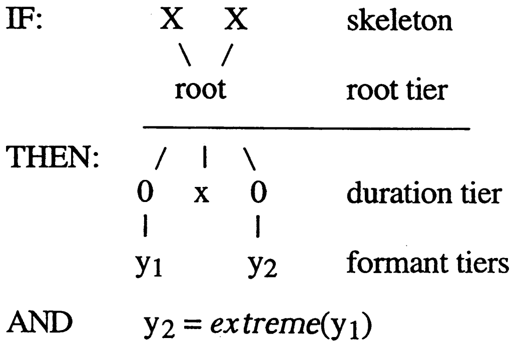
The second row of values would be appropriate for the synthesis of [y] in a word like *yacht*, for example.[23]

With this background, let us return to the treatment of Diphthongization. We have seen evidence that at least for our main subject, Diphthongization should be considered a phonetic phenomenon and not a phonological one. How then can we account for diphthongization at the phonetic level? As far as the long vowels are concerned, the regularity we

[23] We do not discuss the question of whether the [y] of [ay] could be phonologically identified with [i] or [ɪ] instead of [y]. We point out, however, that the formant values of [y] in [ay] are not the same as those found in either of these two vocoids in similar contexts; for example, the F1 value of [y] in *side* [sayd] is typically about 400 Hz, while the F1 value of [i] in (*Port*) *Said* [sa.id] is about 320 Hz.

wish to account for is that diphthongal realizations typically glide toward formant target values lying closer to the outer extremities of each speaker's acoustic vowel space. This regularity can be incorporated into the phonetic grammar in a variety of ways. For concreteness, we state it in terms of the structural implication shown in (13), following Itô 1986, though other formalisms might be equally appropriate.

(13) Phonetic Diphthongization



The first part of the implication states that a long (bipositional) vowel occurring in the surface phonological representation must be characterized by two formant target sets (sets of simultaneous formant targets) in the phonetic representation, one at its left edge and one at its right. The second part of the implication requires the values of the second set of targets, y_2, to lie closer to the extremes of the speaker's acoustic vowel space than the first, y_1, at the region corresponding to the vowel's place of articulation. Thus, for example, the front (palatal) vowel [e] must be realized as a phonetic diphthong whose formants glide toward final target values approximating those of [i], and the back (velar) vowel [o] as one whose formants glide toward values approximating those of [u]. Statement (13) can be understood as a constraint which scans fully-specified acoustic scores and checks them for conformity to these conditions. "Ingliding" short vowels are subject to a similar constraint, except that the final line is replaced by the expression "$y_1 = \mathit{extreme}(y_2)$", requiring final target values to converge in the direction of the central vowel [ə].[24]

[24] The fact that long and short vowels glide in opposite directions can perhaps be explained in terms of dispersion theory (Liljencrants and Lindblom 1972, Bosh 1987). Another generalization suggesting a role for dispersion theory is that the final set of formant target values of long vowels and diphthongs covaries with the initial set. For example, as the initial F1 target values drop successively from 750 Hz in [ay] to 520 Hz in [e] to 240 Hz in [i], the final F1 values drop in parallel from 480 Hz for [ay] to 360 Hz for [e] to 240 Hz for [i].

5.7 GAE long vocalic nuclei and the issue of mismatch

We may now summarize the main results of our review of GAE long vocalic nuclei in this section. We have found phonetic evidence from durational behavior supporting an analysis in which long vowels ("LV-nuclei") consist of one phone in surface phonological representation while diphthongs ("D-nuclei") consist of two. These results bear on the alternative phonological analyses reviewed in section 4. There we compared two phonological analyses, one in which all long vowels are converted into diphthongs in the phonology and another in which mid and high long vowels remain single segments at the end of the phonological derivation and are diphthongized only in the phonetics. If we are correct in believing that there is minimal mismatch between phonological and phonetic analyses (the Congruence Hypothesis), our phonetic description offers support for the second of these two approaches. Otherwise, we would have to introduce an otherwise unnecessary "translation rule" at the transition from the phonetics to the phonology which would conflate nonlow diphthongs into one-segment long vowels to provide an optimal basis for the statement of duration rules.

The Congruence Hypothesis does not provide us with an absolute criterion in phonological analysis, however. Many other, purely phonological criteria come into play in evaluating competing phonological analyses. These will sometimes conflict with the principle of minimizing mismatch, and some may take precedence over it (if this were not so, the Congruence Hypothesis would be unfalsifiable). Mismatches are well-known at other interfaces in grammar, and may have to be tolerated in some cases. Nevertheless, many apparent cases of mismatch turn out not to be such upon further analysis (see e.g. Cohn and McCarthy 1994 for an example from Indonesian), and we believe it a reasonable goal to reduce mismatches between phonetics and phonology to a minimum in linguistic descriptions, to the extent compatible with well-motivated principles of phonological analysis.

6 Further discussion

We have postponed discussion of a number of relevant issues which can be usefully raised at this point.

One concerns some obvious limitations of our study. This study has been based on an intensive study of the speech of one speaker of GAE. It remains to be seen how many of the generalizations reported here will extend to the speech of other GAE speakers, as well as to other dialects and languages. Even for this speaker, we have focused especially on the properties of one subset of vowel nuclei in one lengthening context, and our results may have to be modified or extended when other nucleus types and other contexts are examined in similar detail. All of these questions require further study.

A further issue concerns the status of speaker-dependent numerical values in grammatical description. We have already observed that any two GAE speakers may differ considerably in terms of the absolute acoustic and durational values characterizing their speech, which will vary according to age, gender, and other factors. Should this fact force us to conclude that each member of a given speech community has a unique grammar, differing from that of other members in the values specified in the phonetic component? If that is so, in what sense can it be said that all members of speech community share the same grammar?

One approach to this paradox would be to consider the possibility that the phonetic component assigns *normalized* rather than absolute values for at least some phonetic parameters. This idea could be implemented, for example, in terms of a model making use of multivalued scales; in such a model, the absolute values observed in speech production could be derived by mapping scalar values into the overall range of values used by each speaker. This approach is not simply a revival of earlier proposals to assign scalar values to phonological features (Chomsky and Halle 1968, Ladefoged 1971), since it is not phonological or phonetic features, but phonetic parameters that are given a scalar interpretation. Another way of abstracting away from speaker-specific differences might involve the representation of phonetic events in terms of sequences of abstract vocal tract configurations (e.g. Mrayati, Carré, and Guérin 1988), which can be given speaker-specific interpretations at the level of acoustic implementation. These or other methods of normalization might make it possible to eliminate the use of absolute numerical values in the phonetic component of grammar. At present, however, there is little agreement among researchers on exactly what methods or models are most appropriate for cross-speaker normalization (see e.g. Syrdal and Gopal 1986, Miller 1989, Johnson 1990: 642-4, and Kent and Read 1992: 91-2 for useful discussion). One important practical advantage of describing vowels in terms of absolute acoustic values, as we have done here, is that it allows us to test and evaluate our models in terms of a easily-observable and well-understood system of reference, allowing direct comparison of our results with those of many other researchers.

Another important issue, to which we cannot do full justice here, involves the relations between acoustic patterns and the articulations that underlie them. Any acoustic description of facts involving speech timing and coarticulation involves much complex and at first sight arbitrary detail. For example, we have noted that consonants often show different sets of formant target values at each edge, which vary according to the identity of the adjacent vowels (cf. Figure 10). While this variation adds much complexity to the acoustic description, it makes good sense from the point of view of articulatory coarticulation. Again, the durations of the transitions between any two phones depends on their stricture type and place of articulation in a way that is arbitrary from a purely acoustic point of view. Many other examples of this type could be added to the list.

These observations do not invalidate our approach, but point to the limitations of any purely acoustic approach to phonetic description. The nature of preferred acoustic patterns in speech depends intimately on the structure of the vocal tracts that produce them. For this reason, we view the model of acoustic representation presented here as constituting just one level of the full phonetic grammar, which must eventually be integrated into a fuller model including an articulatory level. The surface representations provided by the phonology are interpreted at both these levels, which are coordinate and interconnected (Figure 24).

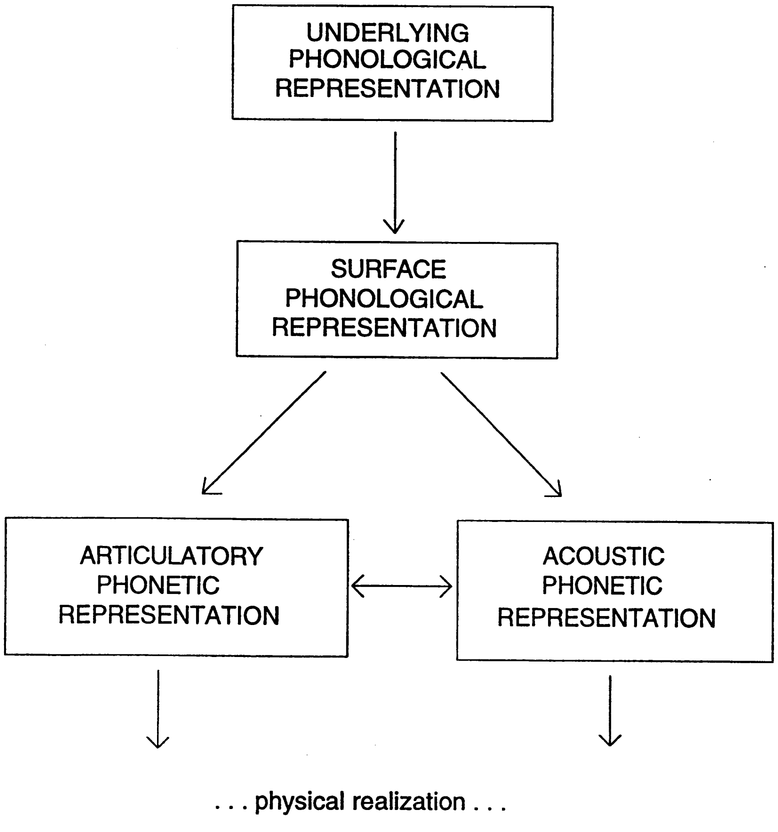

UNDERLYING PHONOLOGICAL REPRESENTATION
SURFACE PHONOLOGICAL REPRESENTATION
ARTICULATORY PHONETIC REPRESENTATION
ACOUSTIC PHONETIC REPRESENTATION
. . . physical realization . . .

Figure 24. A simplified model of phonology/phonetics relations.

Since any given articulatory configuration determines a specific acoustic output, and since a given acoustic representation narrowly limits the set of articulatory configurations that can be associated with it, the representations of each level constrain those of the other, as we have suggested by the double arrow. Given a model of this type, it may be possible to explain many apparently arbitrary aspects of acoustic representation in terms of physical constraints on articulatory production. For these and other reasons, we view the acoustic level of representation as part of a fuller model of the phonetic component of grammar which must be developed in conjunction with research on articulation, aerodynamics, perception, and other factors influencing the nature of speech production.[25]

7 Summary and conclusions

This paper has proposed that phonetic structure forms a grammatical system in much the same sense as syntax, phonology, and other traditional domains of grammatical study. Like other components of grammar, the phonetic component involves both language-independent and language-specific principles, the latter of which determine the ways in which languages differ from one another in their phonetic realizations. The phonetic component comprises (among others) an acoustic level of representation, distinct from both surface phonological representation and the physical speech signal. Given a target-and-interpolation model of phonetic interpretation, acoustic representations can be viewed as having the same formal structure as phonological representations. We have defined and motivated a phone-and-transition model of acoustic segmentation, which characterizes acoustic representations as quasi-segmental in the sense that they can be exhaustively analyzed as sequences of discrete phones separated by transitions.

The above proposals have been motivated in terms of a study of GAE long vocalic nuclei. Drawing upon the representational properties of our model, we have argued that the diphthongization of the long vowels [e o i u] is a phonetic rather than a phonological phenomenon, at least for the idiolect studied here. We have shown that a phonetic analysis of diphthongization simplifies the phonology by removing the need for an otherwise unnecessary rule, and simplifies the phonetics by providing an optimal basis for the statement of duration rules. Phrase-medial Lengthening (10) is also a phonetic rule, as it assigns gradient values and does not eliminate phonological contrasts. We have concluded that there need be no mismatch at the phonology/phonetics interface, at least as far as the facts

[25] See Fujimura 1990 for discussion of many other components involved in a full model of speech production.

discussed here are concerned, since the representations provided by the phonology are identical to those required for the simplest statement of phonetic rules.[26]

The research described in this paper has been conducted in the light of the Congruence Hypothesis, which holds that the most highly valued representations of an utterance provided by the phonology and the phonetics will be largely isomorphic. This hypothesis has been implemented in terms of a specific formal framework which we have termed the IRS. Thus, for example, we have proposed a definition of three types of vocalic nuclei which hold equally at the phonological and phonetic levels (Figure 20), if we allow that phonological representations are "underspecified" with respect to acoustic parameter values that are introduced only at the level of phonetic interpretation.

It has previously been thought that phonetic considerations should play no role in phonological analysis, a consequence of the principle of "separation of levels" widely endorsed in mid-century. This view appears increasingly arbitrary. While phonology and phonetics form separate components of the grammar, they are not entirely independent of each other, any more than are phonology and morphology, or morphology and syntax. Indeed, much research over past years has indicated that different grammatical components are more closely interrelated than had previously been supposed. It follows that the simplicity of any phonetic or phonological description must be evaluated in terms of the overall simplicity of the grammar of which it forms a part. The inclusion of an integrated model of phonetic representation into grammatical theory allows phonological and phonetic descriptions to be evaluated in terms of a global evaluation measure, placing an important phonological constraint on phonetic description and an equally important phonetic constraint on phonological description.

[26] This statements is, of course, theory-dependent The optimal formulation of the duration rules required that the vowel peak, any following approximant (glide or liquid), and adjacent voiced transitions form a *nucleus* constituent in the phonetic representation. This constituent is provided at the phonological level by most rhyme-based syllable theories, but must be excluded from mora-based ones, in which it is superfluous (Steriade 1990). If mora-based theories are to account for durational generalizations in terms of the Tax Law or similar principles, they will have to trade in moras for nuclei at the transition from the phonology to the phonetics. We have seen that nucleus-based models of syllable structure directly provide the structure we need for the simplest expression of duration rules, with the proviso that the nucleus must be expanded to include liquids at the phonetic level, for the reasons discussed earlier.

8 Acknowledgements

Earlier versions of this paper have been presented at the Workshop on the Phonology-Phonetics Interface (Ohio State University, 1993), the Conference on Interfaces in Phonology (Berlin, 1995), and the Conference on Phonological Theory (Abbaye de Royaumont, 1995), as well as at several university colloquia. A preliminary version appears as Clements and Hertz (1996). We would like to express our thanks to audiences on these occasions for their many stimulating questions and comments, as well as to Abby Cohn, Ken de Jong, François Dell, Morris Halle, Janet Pierrehumbert, Jim Scobbie, and Lisa Zsiga for reading and offering detailed comments on earlier versions; we naturally accept full responsibility for the views expressed here as well as for all errors and omissions. Our research was supported in part by research grant number 5 R44 DC 00758-03 from the National Institute on Deafness and Other Communication Disorders, National Institutes of Health, and by a travel grant from the CNRS under a cooperative program with the NSF.

9 References

Allen, J., M. S. Hunnicutt, and D. Klatt (1987) *From Text to Speech: the MITalk System.* Cambridge: Cambridge University Press.

Anderson, S.R. (1976) Nasal Consonants and the Internal Structure of Segments. *Language* 52, 326-344.

Anderson, J.M. and C. Jones (1977) *Phonological Structure and the History of English.* Amsterdam: North-Holland.

Bailey, C.-J. N. (1985) *English Phonetic Transcription.* Dallas, Texas: Summer Institute of Linguistics.

Blevins, J. (1995) The Syllable in Phonological Theory. In J. Goldsmith (ed.) *A Handbook of Phonological Theory.* Oxford and Cambridge, Ma: Basil Blackwell.

Bond, Z.S. (1982) Experiments with Synthetic Diphthongs. *Journal of Phonetics* 10, 259-64.

Bosh, L.F.M. ten (1987) About Diphthongs: an Implementation into Dispersion Theory. *Proceedings of the Institute of Phonetic Sciences, University of Amsterdam*, 1-14.

Browman, C.P. and L. Goldstein (1986) Towards an Articulatory Phonology. *Phonology Yearbook* 3, 219-252.

Browman, C.P. and L. Goldstein (1989) Articulatory Gestures as Phonological Units. *Phonology* 6.2, 201-251.

Browman, C.P. and L. Goldstein (1990) Tiers in Articulatory Phonology, with Some Implications for Casual Speech. In J. Kingston and M. Beckman (eds.) *Papers in*

Laboratory Phonology 1: Between the Grammar and the Physics of Speech, Cambridge University Press, Cambridge, 341-76.

Burzio, L. (1993) English Stress, Vowel Length, and Modularity. *Journal of Linguistics* 29.2, 359-418.

Chen, M. (1970) Vowel Length Variation as a Function of the Voicing of the Consonant Environment. *Phonetica* 22, 129-59.

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

Clements, G.N. (1985) The Geometry of Phonological Features. *Phonology Yearbook* 2, 225-252.

Clements, G.N. (1992) Phonological Primes: Features or Gestures? *Phonetica* 49, 181-93.

Clements, G.N. and S.R. Hertz (1991) Nonlinear Phonology and Acoustic Interpretation. In *Proceedings of the XIIth International Congress of Phonetic Sciences*, Aix-en-Provence, vol. 1, 364-73.

Clements, G.N. and S.R. Hertz (1996) An Integrated Approach to Phonology and Phonetics. In J. Durand and B. Laks (eds.) *Current Trends in Phonology*, vol. 1. CNRS, Paris-X, and University of Salford: University of Salford Publications, 143-74.

Clements, G.N., S.R. Hertz, and B. Lauret (1995) A Representational Basis for Modeling English Vowel Duration. In *Proceedings of the XIIIth International Congress of Phonetic Sciences*, Stockholm, vol. 1, 46-49.

Clements, G.N. and S.J. Keyser (1983) *CV Phonology.* Cambridge, Ma.: MIT Press.

Clements, G.N. and E. Hume (1995) The Internal Structure of Speech Sounds. In John Goldsmith (ed.) *Handbook of Phonological Theory.* Oxford and Cambridge, Ma.: Basil Blackwell, 245-306.

Cohn, A.C. (1990) Phonetic and Phonological Rules of Nasalization. *UCLA Working Papers in Phonetics* 76. Los Angeles: Dept. of Linguistics, UCLA.

Cohn, A.C. (1993) Nasalization in English: Phonology or Phonetics? *Phonology* 10, 43-81.

Cohn, A.C. and J. McCarthy (1994) Alignment and Parallelism in Indonesian Phonology. Ms., Cornell University, Ithaca, and University of Massachusetts, Amherst.

Coker, C.H. (1976) A Model of Articulatory Dynamics and Control. *Proceedings of the IEEE* 64, 452-60.

Davis, K. (1994) Stop Voicing in Hindi. *Journal of Phonetics* 22, 177-93

Dixon, N.R. and H.D. Maxey (1968) Terminal Analog Synthesis of Continuous Speech using the Diphone Method of Segment Assembly. *IEEE Trans. Audio Electroacoust.* 16, 40-50.

Falk, Y.N. (1991) Shifty Vowels. *Folia Linguistica* XXV/3-4, 483-513.

Fant, G. (1970) *Acoustic Theory of Speech Production* (2nd edition). The Hague: Mouton.

Foley, L.M. (1972) *A Phonological and Lexical Study of the Speech of Tuscaloosa County, Alabama.* Publications of the American Dialect Society No. 58. University, Alabama: University of Alabama Press.

Fourakis, M. and R. Port. 1986. "Stop Epenthesis in English." *JoP* 14.2, 197-221,

Fujimura, O. (1990) Methods and Goals of Speech Production Research. *Language and Speech* 33.3, 195-258.

Fujimura, O. and J. Lovins (1978) Syllables as Concatenative Phonetic Elements. In A. Bell and J.B. Hooper (eds.) *Syllables and Segments.* New York: North-Holland, 107-120.

Gay, T. (1968) Effect of Speaking Rate on Diphthong Formant Movement. *JASA* 44, 1550-1573.

Gay, T. (1970) A Perceptual Study of American English Diphthongs. *Language and Speech* 13, 65-88.

Goldsmith, J.A. (1976) *Autosegmental Phonology.* MIT Ph.D. dissertation. (Published 1979 by Garland Publishing, N.Y.)

Goldsmith, J.A. (1990) *Autosegmental and Metrical Phonology.* Oxford and Cambridge, Ma.: Basil Blackwell.

Goldsmith, J. A. (ed.) (1995) *A Handbook of Phonological Theory.* Oxford and Cambridge, Ma.: Basil Blackwell.

Halle, M. (1977) Tenseness, Vowel Shift, and the Phonology of Back Vowels in Modern English. *Linguistic Inquiry* 8, 611-625.

Halle, M. and K.P. Mohanan (1985) The Segmental Phonology of Modern English. *Linguistic Inquiry* 16.1, 57-116.

Halle, M. and K.N. Stevens (1991) Knowledge of Language and the Sounds of Speech. In J. Sundberg, L. Nord and R. Carlson (eds.) *Music, Language, Speech, and Brain.* (Wenner-Gren International Symposium Series, vol. 59.) Houndmills, Basingstoke, Hampshire, and London: MacMillan Academic and Professional Ltd., 1-19.

Hertz, S.R. (1979) An Interactive Speech Synthesis System for Linguistics. PhD dissertation, Cornell University, Ithaca, N.Y.

Hertz, S.R. (1982) From Text to Speech with SRS. *JASA* 72.4, 1155-1170.

Hertz, S.R. (1988) Delta: Flexible Solutions to Tough Problems in Speech Synthesis by

Rule. *The Official Proceedings of Speech Tech 88*. New York: Media Dimensions Inc.

Hertz, S.R. (1990a) The Delta Programming Language: an Integrated Approach to Non-linear Phonology, Phonetics, and Speech Synthesis. In J. Kingston and M. Beckman (eds.) *Papers in Laboratory Phonology 1: Between the Grammar and the Physics of Speech*. Cambridge: Cambridge University Press, 215-57.

Hertz, S.R. (1990b) A Modular Approach to Multi-dialect and Multi-language Speech Synthesis Using the Delta System. In *Proceedings of the ESCA Workshop on Speech Synthesis*, Autrans, France, 225-228.

Hertz, S.R. (1991) Streams, Phones, and Transitions: Toward a New Phonological and Phonetic Model of Formant Timing. *Journal of Phonetics* 19, 91-109.

Hertz, S.R. (1992) The Timing of Phones and Transitions: a Nucleus-based Model of English Duration. In *Working Papers of the Cornell Phonetics Laboratory* 7, 135-150.

Hertz, S.R. and M. Huffman (1992) A Nucleus-based Timing Model Applied to Multi-dialect Speech Synthesis by Rule. In J.J. Ohala et al. (eds.) *Proceedings of ICSLP 92*, vol. 2 Edmonton, Alberta: Department of Linguistics, University of Alberta, 1171-4.

Hertz, S.R., J. Kadin, and K. Karplus (1985) The Delta Rule Development System for Speech Synthesis from Text. *Proceedings of the IEEE* 73, No. 11, 1589-1601.

Hertz, S.R., E.C. Zsiga, K.J. de Jong, P. Gries, and K.E. Lockwood (1994) From Database to Speech: a Multi-dialect Relational Database Integrated with the Eloquence Synthesis Technology. *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis*, New Palz, N.Y., September 12-15, 1994, 45-8.

Hillenbrand, J., D.R. Ingrisano, B.L. Smith, and J.E. Flege (1984) Perception of the Voiced-voiceless Contrast in Syllable-final Stops. *JASA* 76, 18-27.

Holbrook, A. and G. Fairbanks (1962) Diphthong Formants and their Movements. *Journal of Speech and Hearing Research* 5, 38-58.

Holmes, J.N. (1983) Speech Technology in the Next Decades. *Proceedings of the Xth International Congress of Phonetic Sciences*. Dordrecht: Foris Publications, 125-40.

Holmes, J.N., I.G. Mattingly, and J.N. Shearme (1964) Speech Synthesis by Rule. *Languaage and Speech* 7, 127-143.

Huffman, M. (1990) Implementation of Nasal: Timing and Articulatory Landmarks. *UCLA Working Papers in Phonetics* 75. Los Angeles: Dept. of Linguistics, UCLA.

Itô, J. (1986) Syllable Theory in Prosodic Phonology. Ph.D. dissertation, University of Massachusetts, Amherst. (Published 1988 by Garland Publishing, N.Y.)

Jespersen, O. (1909) *A Modern English Grammar on Historical Principles. Part 1: Sounds and Spelling*. Heidelberg: Carl Winter's Universitätsbuchhandlung.

Johnson, K. (1990) The Role of Perceived Speaker Identity in F0 Normalization of Vowels. *JASA* 88(2), 642-54.

Keating, P. (1985) Universal Phonetics and the Organization of Grammars. In V.A. Fromkin (ed.) *Phonetic Linguistics*. New York: Academic Press, 115-32.

Keating, P. (1988) Underspecification in Phonetics. *Phonology* 5.2, 275-292.

Kelly, J. and L. Gerstman (1961) An Artificial Talker Driven from Phonetic Input. *JASA* 33 Suppl. 1, S35.

Kenstowicz, M. and C.W. Kisseberth (1979) *Generative Phonology: Description and Theory*. New York, N.Y.: Academic Press.

Kenstowicz, M. (1994) *Phonology in Generative Grammar*. Oxford and Cambridge, Ma.: Basil Blackwell.

Kent, R. D. and C. Read (1992) *The Acoustic Analysis of Speech*. San Diego: Singular Publishing Group, Inc.

Kenyon, J.S. and T.A. Knott (1944) *A Pronouncing Dictionary of American English* Springfield, Mass.: Merriam.

Keyser, S.J. and K.N. Stevens (1994) Feature Geometry and the Vocal Tract. *Phonology* 11.2, 207-36.

Kingston, J. (1990) Articulatory Binding. In J. Kingston and M. Beckman (eds.) *Papers in Laboratory Phonology 1: Between the Grammar and the Physics of Speech*. Cambridge: Cambridge University Press, 406-34.

Kiparsky, P. (1973) Elsewhere in Phonology. In S. Anderson and P. Kiparsky (eds.) *A Festschrift for Morris Halle*. New York: Holt, Rinehart and Winston, 93-106.

Kiparsky, P. (1982) Lexical Morphology and Phonology. In I.-S. Yang (ed.) *Linguistics in the Morning Calm*. Hanshin, Seoul: Linguistic Society of Korea, 3-91.

Klatt, D.H. (1976) Linguistic Uses of Segmental Duration in English: Acoustic and Perceptual Evidence. *JASA* 59.5, 1208-1221.

Klatt, D.H. (1979) Synthesis by Rule of Segmental Durations in English Sentences. In B. Lindblom and S. Öhman (eds.) *Frontiers of Speech Communication Research*. New York, N.Y.: Academic Press, 287-300.

Klatt, D.H. (1980) Software for a Cascade/Parallel Formant Synthesizer. *JASA* 67, 971-95.

Klatt, D.H. (1987) Review of Text-to-Speech Conversion for English. *JASA* 82, 737-93.

Klatt, D.H. and L.C. Klatt (1990) Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers. *JASA* 87, 820-57.

Kurath, H. and R.I. McDavid, Jr. (1961) *The Pronunciation of English in the Atlantic*

States. Ann Arbor: University of Michigan Press.

Labov, W. (1986) Sources of Inherent Variation. In J.S. Perkell and D.H. Klatt (eds.) *Invariance and Variability in Speech Processes.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Labov, W. (1994) *Principles of Linguistic Change: Internal Factors.* Oxford: Blackwell.

Ladd, D.R. (1993) In Defense of a Metrical Theory of Intonational Downstep. In Harry van der Hulst and Keith Snider (eds.) *The Phonology of Tone: the Representation of Tonal Register.* Berlin and New York: Mouton de Gruyter, 109-132.

Ladefoged, P. (1967) The Nature of Vowel Quality. In *Three Areas of Experimental Phonetics.* London: Oxford University Press, 50-147.

Ladefoged, P. (1971) *Preliminaries to Linguistic Phonetics.* Chicago: University of Chicago Press.

Lehiste, I. (1964) *Acoustical Characteristics of Selected English Consonants.* Bloomington: Indiana University Press.

Lehiste, I. and G. Peterson (1961) Transitions, Glides, and Diphthongs. *JASA* 33, 268-77.

Liljencrants, J. and B. Lindblom (1972) Numerical Simulation of Vowel Quality Systems: the Role of Perceptual Contrast. *Language* 48, 839-862.

Lindau, M. (1984) Phonetic differences in Glottalic Consonants. *Journal of Phonetics* 12, 147-56.

Lubker, J. (1979) The reorganization times of bite-block vowels. *Phonetica* 36, 273-293.

McCarthy, J. (1986) OCP Effects: Gemination and Antigemination. *Linguistic Inquiry* 17, 207-263.

Maeda, S. (1990) Compensatory Articulation During Speech: Evidence from the Analysis and Synthesis of Vocal-Tract Shapes using an Articulatory Model. In W.J. Hardcastle and A. Marchal (eds.) *Speech Production and Speech Modeling*. Dordrecht: Kluwer Academic Publishers, 131-49.

Maeda, S. (1991) On Articulatory and Acoustic Variabilities. *Journal of Phonetics* 19, 321-331

Mermelstein, P. (1973) Articulatory Model for the Study of Speech Production. *JASA* 53, 1070-1082.

Miller, J.D. (1989) Auditory-perceptual Interpretation of the Vowel. *JASA* 85.5, 2114-2134.

Milliken, S. (1988) Protosyllables: a Theory of Underlying Syllable Structure in Nonlinear Phonology. Ph.D. dissertation, Cornell University, Ithaca, N.Y.

Mohanan, K.P. (1986) *The Theory of Lexical Phonology*. Dordrecht, Reidel.

Mrayati, M., R. Carré, and B. Guérin (1988) Distinctive Regions and Modes: a New Theory of Speech Production. *Speech Communication* 7, 257-286.

Myers, S. (1987) Vowel Shortening in English. *Natural Language and Linguistic Theory* 5.4, 485-518.

Ohala, J.J. and H. Kawasaki (1984) Prosodic Phonology and Phonetics. *Phonology Yearbook* 1, 113-28.

Olive, J.P. (1990) A New Algorithm for a Concatenative Speech Synthesis System using an Augmented Acoustic Inventory of Speech Sounds. *Proceedings of the ESCA Conference on Speech Synthesis*, 25-29.

Perkell, J. (1980) Phonetic Features and the Physiology of Speech Production. In B. Butterworth (ed.) *Language Production I: Speech and Talk.* London and New York: Academic Press, 337-72.

Peterson, G.E. and M.S. Coxe (1953) The Vowels /e/ and /o/ in American Speech. *Quarterly Journal of Speech* 39, 33-41.

Peterson, G.E. and I. Lehiste (1960) Duration of Syllable Nuclei in English. *JASA* 32, 693-703.

Peterson, G., W. Wang and E. Sivertson (1958) Segmentation Techniques in Speech Synthesis. *JASA* 30, 739-42.

Pierrehumbert, J. (1980) The Phonology and Phonetics of English Intonation. Ph.D. dissertation, MIT, Cambridge, Ma.

Pierrehumbert, J. (1994) Knowledge of Variation. *Papers from the Parasession on Language Variation, CLS 30.* Chicago: Department of Linguistics, University of Chicago.

Pierrehumbert, J. and M. Beckman (1988) *Japanese Tone Structure*. Cambridge, Ma.: MIT Press.

Pierrehumbert, J., M. Beckman, and D.R. Ladd (in press). Laboratory Phonology. To appear in J. Durand and B. Laks (eds.) *Current Trends in Phonology*. CNRS, Paris-X, and University of Salford: University of Salford Publications.

Pike, K.L. (1947) On the Phonemic Status of English Diphthongs. *Language* 23.2, 151-9

Ren, H. (1986) On the Acoustic Structure of Diphthongal Syllables. *UCLA Working Papers in Phonetics* 65. Los Angeles: Dept. of Linguistics, UCLA.

Schane, S. (1984) Two English Vowel Movements: a Particle Analysis. In M. Aronoff and R. T. Oehrle (eds.) *Language Sound Structure: Studies in Phonology Presented to Morris Halle by his Teacher and Students.* Cambridge, Ma.: MIT Press, 32-51.

Selkirk, E.O. (1982) The Syllable. In H. van der Hulst and N. Smith (eds.) *The Structure of Phonological Representations* (Part 2). Dordrecht: Foris Publications, 338-83.

Sluyters, W.A.M. (1992) Representing Diphthongs. Doctoral dissertation, Katholieke Universiteit te Nijmegen.

Stampe, D. (1980) *A Dissertation on Natural Phonology.* Garland Publishing Co., N.Y.

Steriade, D. (1987) Locality Conditions and Feature Geometry. *Proceedings of NELS 17.* Amherst, Ma.: Department of Linguistics, University of Massachusetts, 595-618.

Steriade, D. (1990) Moras and other Slots. *Proceedings of FLSM 1.* Madison: Department of Linguistics, University of Wisconsin at Madison.

Stevens, K.N. (1972) The Quantal Nature of Speech: Evidence from Articulatory-Acoustic Data. In E.E. David, Jr. and P.B. Denes (eds.) *Human Communication: a Unified View.* New York: McGraw-Hill, pp. 51-66.

Stevens, K.N. (1989) On the Quantal Nature of Speech. *Journal of Phonetics* 17, 3-45.

Stevens, K.N. (1994) Phonetic Evidence for Hierarchies of Features. In P. Keating (ed.) *Phonological Structure and Phonetic Form: Papers in Laboratory Phonology 3.* Cambridge: Cambridge University Press.

Stevens, K.N, S.J. Keyser, and H. Kawasaki (1986) Toward a Phonetic and Phonological Theory of Redundant Features. In J. Perkell and D. Klatt (eds.) *Symposium on Invariance and Variability of Speech Processes.* Hillsdale: Lawrence Erlbaum, 432-469.

Summers, W. Van (1987) Effects of Stress and Final-consonant Voicing on Vowel Production: Articulatory and Acoustic Analyses. *JASA* 82.3, 847-863.

Swadesh, M. (1935) The Vowels of Chicago English. *Language* 11, 148-51.

Syrdal, A. and H. Gopal (1986) A Perceptual model for Vowel Recognition Based on Auditory Representation of American-English Vowels. *JASA* 79.4, 1086-1100.

Trager, G.L. and B. Bloch (1941) The Syllabic Phonemes of English. *Language* 17, 233-46.

Trager, G.L. and H.L. Smith (1951) *An Outline of English Structure.* Norman, Oklahoma: Battenburg Press.

Tranel, B. (1987) *The Sounds of French: an Introduction.* Cambridge: Cambridge University Press.

Wells, J.C. (1982) *Accents of English.* Volumes 1-3. Cambridge: Cambridge University Press.

Wolfe, P.M. (1972) *Linguistic Change and the Great Vowel Shift in English.* Berkeley, Ca.: The University of California Press.

Wood, S. (1979) A Radiographic Analysis of Constriction Locations for Vowels. *Journal of Phonetics* 7, 25-43.

Wood, S. (1991) X-ray data on the temporal coordination of speech gestures. *Journal of Phonetics* 19, 281-92.

Zsiga, E.C. (1993) Features, Gestures, and the Temporal Aspects of Phonological Organization. Ph.D. dissertation, Yale University, New Haven, Conn.

Zue, V. and M. Laferriere (1979) Acoustic Study of Medial /t,d/ in American English. *JASA* 66(4), 1039-1050.

Appendix A: Some further implications of the analysis.

Our analysis of diphthongization has implications that extend beyond the facts examined so far. In the domain of second language acquisition, it is frequently observed that English speakers tend to generalize their diphthongized pronunciations of long vowels to other languages. Thus, English speakers tend to lengthen and diphthongize the French vowels [i, u, e, o] in open syllables, although these vowels are short and monophthongal in French (e.g. Tranel 1987, 41); as early as 1909, Jespersen noted that the vowels [e o] were diphthongized in the normal British pronunciation of French words like *soirée*, *écarté*, *éclat*, *naïveté*, *château*, *chaperon*, *hauteur*, and *dépôt*.. The explanation for this behavior is that only long vowels and diphthongs occur in open syllables in English; thus French vowels, which occur typically in open syllables, are identified with English long vowels. This behavior has been explained in the past by the principle that speakers generalize the productive phonological rules of their own language to other languages (see e.g. Kenstowicz and Kisseberth 1979, Stampe 1980). However, a phonetic analysis of diphthongization can account for these observations just as easily. A word like *soirée* will be assigned the underlying representation / swarē /, and Phonetic Diphthongization (13) will require the long mid vowel to be interpreted with two sets of formant targets. In contrast, nonproductive rules of the phonology are not generalized to foreign words; thus the rules involved in Vowel Shift, which have lexical exceptions (e.g. *oblique/obliquity*; *detain/detention*), are not applied to words like *soirée* and *ami*.

Our analysis also has implications for the issue of how successive historical sound changes are stratified in synchronic rule systems. In Figure 24 we summarize the historical development of contemporary GAE diphthongs, based on the interpretation of early descriptions of English pronunciation offered by Chomsky and Halle (1968) as revised by Wolfe (1972), following the sources listed underneath.[27]

[27] Note that another source of [ey] is ME *ai*, and another source of [ow] is ME *ou*, neither of which is included on the table.

ME:	ī	ẹ̄	ę̄	ɑ̄	ǭ	ọ̄	ū	ẹ̄w	ę̄w
1:	ey	ī	ɛ̄?/ē?	ɑ̄	ɔ̄?/ō?	ū	ow	yu	ew
2:	əy?	ī	ē	ǣ	ō	ū	ʌw	iw?	ew
								\	/
3:	əy	ī	ē	ɛ̄	ō	ū	əw	yu?	
		\	/						
4:	ʌy	iy		ey	ow	uw	ɔw	yuw	
GAE:	ay	iy		ey	ow	uw	aw	yuw	

1 = John Hart (1551, 1569, 1570)
2 = John Wallis (1653-1699)
3 = Christopher Cooper (1687)
4 = T. Batchelor (1809)

Figure A-1. Historical development of selected English long vowel nuclei.

This table suggests the broad outlines of how the system of modern English long vowels and diphthongs has evolved. Except for the roughly contemporary testimonies of Wallis and Cooper, each successive line represents a later historical stage of English. Many details of this development are unclear and subject to debate; in particular, the historical sources do not allow us to sort out the relative chronology of high vowel diphthongization-and-lowering, mid vowel raising, and low vowel raising, at least the first two of which were already in place in Hart's time. However, we must assume that high vowels were phonologically diphthongized, lowered or otherwise changed before the mid vowels were raised, since otherwise the mid vowels would have merged with them. It is pertinent to note that diphthongization was introduced in two well-separated stages. Diphthongization of the ME long high vowels / ī ū / was already present in the 16th-century (line 1). In contrast, diphthongization of the raised reflexes of the ME low and mid vowels / ē ǣ ā ɒ̄ ō / is not reported in the literature until the early 19th century (line 4).[28] Historically, then, these two types of diphthongization were well-separated in time. If our phonetic analysis is correct, and if phonological diphthongization affects only the high vowels as suggested by

[28] Some scholars have suggested, on the contrary, that all long vowels might have had diphthongal realizations as early as ME. However, this remains a minority view, which receives little support in the descriptions of the 16th and 17th century grammarians; see Wolfe 1972, especially pp. 160-6, for a careful review of the relevant evidence.

Falk and others, we find that the contemporary system of GAE preserves the historical order of the two types of diphthongization in the form of synchronic rule stratification, assigning the historically earlier rule to the phonology and the later, 19th century rule to the phonetics.

authors' addresses:

Nick Clements
CNRS (UA 1027)
19 rue des Bernardins
75005 Paris, France
e-mail: clements@ilpga.msh-paris.fr

Susan R. Hertz
24 Highgate Circle
Ithaca, N.Y. 14850
e-mail: hertz@cs.cornell.edu