A REPRESENTATIONAL BASIS FOR MODELLING ENGLISH VOWEL DURATION

G. N. Clements, CNRS, UA 1027, Paris
Susan R. Hertz, Eloquent Technology, Inc. and Cornell University, Ithaca, N.Y.
Bertrand Lauret, Université de Paris III, UA 1027, Paris

ABSTRACT

This paper proposes a representational basis for modelling the durational behavior of syllable nuclei in General American English. It examines two lengthening patterns, one in which all portions of the nucleus are affected uniformly, and another in which primarily the beginning and end portions are affected. On the basis of this distinction, a classification of nuclei into one-phone vs. two-phone nuclei is proposed.

1. AN INTEGRATED REPRESENTATIONAL BASIS FOR PHONOLOGY AND PHONETICS

Our aim in this paper is to present the broad outlines of a working model of the phonology-phonetics interface, with an illustration from certain facts of General American English. Our approach is based on the premise that phonetics should be viewed as an essential component of the theory of grammar, and that as such, it can be studied in terms of much the same type of theoretical modelling that we find elsewhere in linguistic theory. In this view, which received a preliminary formulation in Clements and Hertz [1], the phonetic component of a grammar does not consist of descriptions of physical patterns as such, but involves a symbolic representational system defined at a level of some abstraction from physical data.

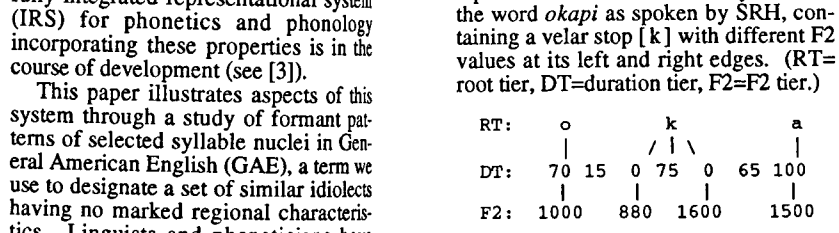
Specifically, we propose that the categorial feature representations of the phonological level are projected directly into the acoustic phonetic level, where they provide the basis for specifying acoustic parameter values in terms of which speech output can be accurately modelled. Like phonological representation, acoustic representation involves partially-specified, multitiered arrays of units related by often complex patterns of association. Acoustic representation differs from phonological representation primarily in introducing new acoustic and duration tiers, required to account for language- and speaker-specific regularities in the acoustic output. By allowing acoustic units to be only partially specified, we allow rising and falling ramps between extrema to be modelled in terms of a target-and-interpolation model [2], while the use of multiple tiers allows for the description of regular patterns of overlap within and across segments. A fully integrated representational system (IRS) for phonetics and phonology incorporating these properties is in the course of development (see [3]).

This paper illustrates aspects of this system through a study of formant patterns of selected syllable nuclei in General American English (GAE), a term we use to designate a set of similar idiolects having no marked regional characteristics. Linguists and phoneticians have long disagreed on the classification of the long gliding vowels of words like *beat, boot, bait,* and *boat*, some treating them as a single segment and others as two. Researchers have also disagreed as to whether the liquids in words like *belt* and *Bart* should be treated as part of the syllable nucleus, or assigned to the margin.

We address these questions within the framework of the integrated approach to phonological and phonetic analysis just outlined. One component of this approach is the phone-and-transition segmentation strategy outlined by Hertz [2]. This strategy is based on the view that speech sounds ("phones") are not necessarily adjacent to each other in phonetic representations, but may be separated from each other by time intervals ("transitions") during which the articulators (lips and tongue) move from the target configuration appropriate for one sound to that appropriate for the next. Phones appear on spectrograms as the time intervals that correspond to such target configurations, while transitions are the time intervals that connect them.

Following these assumptions, we may represent the acoustic structure of an utterance as follows. The root nodes of the phonological representation constitute the *phone tier* of the acoustic representation. Root nodes dominate duration values on a *duration tier*, whose function is to assign each phone a certain duration. Between any two root nodes having different oral tract places of articulation we introduce *transitions*, formally represented as duration values unlinked to root nodes. Duration values dominate appropriate acoustic parameter values on further tiers (F0, F1, F2, aspiration, voicing, etc.). These values can serve as a basis for interpolation across segments unspecified for these parameters.

We illustrate this model with a partial representation of the first two syllables of the word *okapi* as spoken by SRH, containing a velar stop [k] with different F2 values at its left and right edges. (RT= root tier, DT=duration tier, F2=F2 tier.)



This graph represents a pattern with (i) a 70-msec F2 steady state at 1000 Hz characterizing the [o], (ii) a 15-msec transition to the [k] during which F2 falls continuously to a target value of 880 Hz, (iii) a 75-msec period of silence during the [k], (iv) a 65-msec transition to the [a] during which F2 falls from 1600 Hz to 1500 Hz, and (v) a 100-msec F2 steady state at 1500 Hz characterizing the [a]. This representation treats this [k] as a "contour phone", analogous to the contour segments of phonology.

2. DURATIONAL ASPECTS OF ENGLISH SYLLABLE NUCLEI

With this background, we report on a preliminary study of a variety of syllable nucleus types in GAE. Our goal is to find out whether their durational behavior can help us decide whether a given nucleus consists of a single unit or two.

It is well known that GAE syllable nuclei are often lengthened before voiced obstruents, especially phrase-finally [5]. What is less well understood is whether all nuclei lengthen in a uniform manner, or whether they show different patterns of lengthening. Our hypothesis is that if GAE contains a distinction between one-segment nuclei and two-segment nuclei at the phonological level, this distinction might be reflected in different patterns of lengthening at the phonetic level. Such differences, if they exist, might help to answer the questions concerning the analysis of GAE nuclei raised above.

To test this hypothesis, we collected data on four sets of paired monosyllabic words differing only in the voicing of the final consonant: *bit/bid, bait/bade, bite/bide, felt/felled.* One female and three male speakers of GAE were recorded; we report on data from one of the latter (GNC) here. Each test word was pronounced in the frame *say ___ for me*, and the full sequence was repeated ten times. Recordings were digitized at 16 kHz and analyzed by means of the CSRE formant tracker. Aberrant values were discarded, accounting for the occasional gaps in the formant tracks. Segmentation was carried out mainly on the basis of second formant (F2) tracks, since we have found these to provide the most consistent basis for analyzing the temporal properties of vowels and diphthongs.

Representative F2 tracks are displayed in Figure 1. All graphs are reproduced at the same scale. Each one displays an overlay of F2 tracks extracted from the first five tokens of each word. The ordinate represents formant frequency in Hz, and the units of the abscissa represent 8-msec time intervals. Overall, we see that all items were produced with considerable consistency from token to token.

All pairs in Figure 1 have rising or falling F2 ramps, showing that they are phonetic diphthongs. The diphthongal nature of the nucleus of *bit/bid* for this speaker is confirmed by the fact that the F1 track (not shown here) rises as F2 falls, showing that the nucleus ends in a central offglide. Although discussed in [3], the difference in formant values at the beginnings and ends of the nuclei in these words cannot be attributed to coarticulation with the neighboring consonants.

All pairs of nuclei in Figure 1 exhibit F2 lengthening in their second member. However, the first and second columns show distinct patterns. The F2 tracks in the first column resemble a straight line, with some examples showing a sharp drop at the very end. Disregarding these drops, the tracks can be modelled to a fair approximation by positing two duration-less target points at their beginnings and ends, and perfoming a straight-line inter-

polation between them. For these diphthongs, lengthening does not change the overall shape of the F2 track, although its slope is somewhat reduced in the lengthened form.

The diphthongs in the second column display a different F2 lengthening pattern. For these diphthongs, lengthening visibly changes the shape of the F2 track, especially at the ends. Comparing *bite* and *bide*, we see that the durationless initial target of *bite* is replaced by a quasi-plateau some 80 msec long in *bide*; its final steady state is also somewhat longer. In contrast, the F2 transition between the initial and final extrema has about the same duration in both cases (the sharper rise in *bite* can be attributed to its higher final target value). Similar remarks hold of the second pair. The durationless initial target of *felt* is replaced in *felled* by a steady state approximately 30 msec long, and the final portion expands similarly (in three tokens, final F2 values were too low in amplitude to be read). The transitions between these relatively stable portions have about the same duration in both cases.

3. DISCUSSION

We propose that these two patterns can be analyzed as one-segment and two-segment diphthongs, respectively. Note first that the nucleus of *bit*/*bid* is uncontroversially a single vowel at the phonological level, while that of *felt*/*felled* just as clearly constitutes a two-segment sequence. We can explain the fact that the nucleus of *bait*/*bade* patterns with that of *bit*/*bid* by treating them both as one-segment nuclei, and the fact that *bite*/*bide* patterns with *felt*/*felled* by treating both as two-segment nuclei. In addition, if we considered the first-column nuclei to have a two-segment structure, we would have to allow that their first segments are durationless even in lengthening contexts, contrary to our observations elsewhere. Further arguments in support of this analysis are given in [3].

Using the representational system outlined earlier, we can interpret the nuclei [ɪ] and [e] as single phones (= root nodes), even though they show different F2 target values at their left and right edges. Their analysis parallels that of the "contour" phone [k] in *okapi*, observed earlier. Typical values for the [e] of *bade*, for instance, are shown below:

```
RT:        e
          / | \
DT:     0 140 0
        |     |
F2:  1600   1850
```

The representation of [ɪ] differs from that of [e] both in its choice of F2 values and in the fact that its root node is linked to one, instead of two positions on the phonological skeleton (not shown here).

The nuclei [ay] and [ɛl], in contrast, are analyzed as phone sequences, as shown below for the [ay] of *bide*:

```
RT:     a      y
        |      |
DT:    80  70  20
        |      |
F2:  1030    1670
```

Given these analyses, we may state the following generalization: lengthening before voiced stops affects all phones within the syllable nucleus, but affects the transitions between them little, if at all (see [2], [3], [4] for fuller discussion). We can directly explain the fact that [l] lengthens in *felled* by considering that it, too, belongs to the syllable nucleus.

These preliminary observations are offered in illustration of our approach to the study of the phonetics/ phonology interface. Our current project is to examine a fuller set of data, involving further nucleus types, more contexts, and other speakers, in order to determine the generality of these observations, and refine and improve them as necessary.

ACKNOWLEDGEMENT

This work has been supported in part by grant NIDCD R44 DC00758 to Eloquent Technology, Inc..

SELECTED REFERENCES

[1] Clements, G.N. and S.R. Hertz (1991) "Nonlinear Phonology and Acoustic Interpretation", *Proc. of the 12th Int. Congress of Phonetic Sciences*, Aix-en-Provence, vol. 1, 364-73.

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

[3] Clements, G.N. and S.R. Hertz (1995) "An Integrated Model of Phonetic Representation in Grammar," ms.

[4] 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., *Proc. of ICSLP 92*, Edmonton, vol. 2, 1171-4.

[5] Chen, M. 1970. "Vowel Length Variation as a Function of the Voicing of the Consonant Environment," *Phonetica* 22, 129-59.

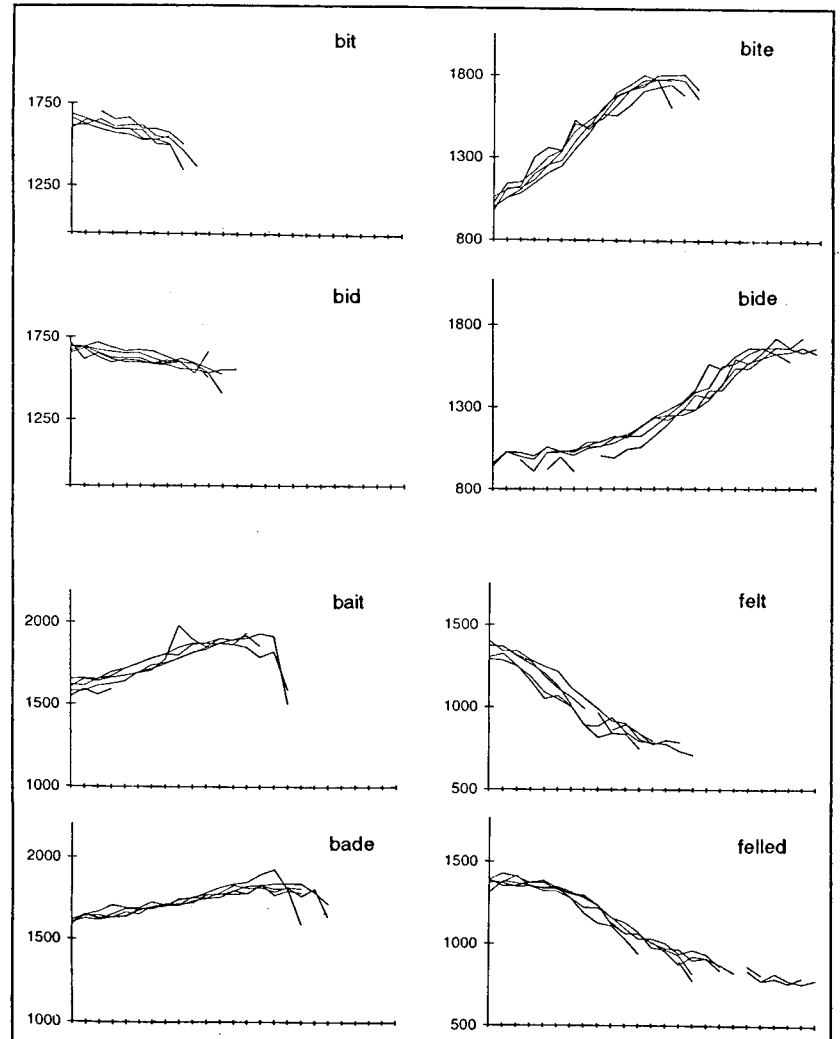


Figure 1.
Lengthening patterns of four types of GAE syllable nuclei, spoken in the context *say __ for me*. Graphs show overlaid F2 tracks of five typical tokens for one speaker.