The Timing of Phones and Transitions: Toward a Nucleus-Based Model of English Duration

Susan R. Hertz

This paper presents a brief overview of a new duration model for General American English, which I am currently developing in connection with work being done at both Cornell University and Eloquent Technology, Inc. in speech synthesis by rule. In addition to pursuing the standard goals of synthesis rule development, such as high intelligibility and naturalness, we strive in this work to develop a linguistically realistic system with which we can gain insights into the timing structure of English and other languages.

The paper expands upon Hertz (1991), which describes and motivates a model of speech timing called the *phone-and-transition model.* While the phone-and-transition model provides an appropriate framework for synthesis of any language or dialect (see Hertz, 1990a), this paper will focus on its application to General American English (henceforth "GA"), presenting the basic rule algorithm for determining the internal timing structure of stressed syllable nuclei in GA.

The first section gives a brief outline of the phone-and-transition model. The second section describes the methodology being used to develop the duration model for GA. The third section discusses the duration model, focussing on the rules that compute the durations of the sub-units of stressed syllable nuclei. A final section provides some concluding remarks.

1. The Phone-and-Transition Model

The duration rules build on a phonetic model in which *phones* and *transitions* are explicit units that can be manipulated independently by rules (Hertz, 1991). Consider, for example, the following representation of the word *tot* in GA, as uttered in the frame *Say ____ for me*:

(1) *tot:*

```
phone:        |t    |       |a    |       |t    |
F2:           |1800 |       |1300 |       |1700 |
transition:   |     |trans  |     |trans  |     |
millisecond:  |95   |70     |85   |50     |85   |
```

This structure, which is produced by our current speech synthesis rules for GA, is called a *delta*, because it consists of multiple interconnected *streams* (e.g., `phone`, `F2`,

transition, and millisecond) like a river delta. (In this and other sample deltas below, only the streams relevant to the example are shown. Deltas produced by our synthesis rules contain many other streams as well.)

In this delta, each unit in the phone stream (i.e., each phone) is synchronized by the vertical bars (*sync marks*) with the formant values (*targets*) that are a direct result of the articulation of the phone itself.[1] Intervening transitions represent the movement of the formants from the target of one phone to the target of the next. The formant values during the transitions between phones are computed by interpolating between the formant targets of adjacent phones over the duration specified for the transition (e.g., by interpolating between 1800 Hz and 1300 Hz over 70 ms for the transition between the phones [t] and [a]). (Although not shown in this example, a phone sometimes has two targets with different values at the beginning and end of the phone, in which case there is interpolation within the phone as well.)

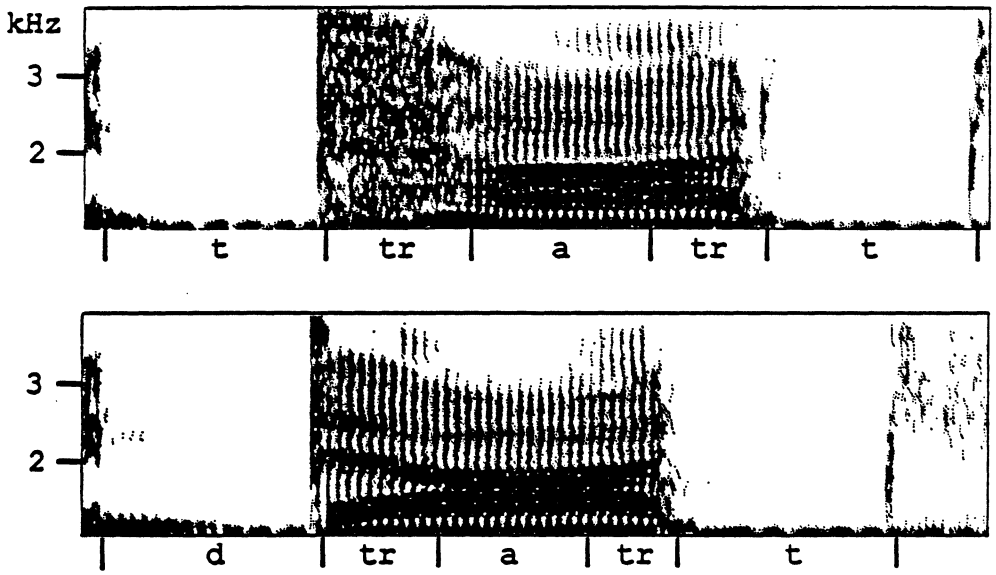
A model based on explicit phones and transitions leads to more straightforward timing rules than conventional models in which the transitions are not treated as independently manipulable units. Consider, for example, the spectrograms of *tot* [tat] and *dot* [dat] in Figure 1. In these spectrograms, we see that the aspiration after the [t] of *tot* is superimposed on the transition between the [t] and the following [a]. In *dot*, on the other hand, the transition from [d] to [a] is voiced. Within the phone-and-transition model, we can easily capture the generalization that stop aspiration aligns with the transition following the stop, as shown below for *tot*:

(2) *tot:*

```
phone:        |t    |      |a    |      |t    |
F2:           |1800 |      |1300 |      |1700 |
aspiration:   |     |70    |     |      |     |
voicing:      |0          |60           |0    |
transition:   |     |trans |     |trans |     |
millisecond:  |95   |70    |85   |50    |85   |
```

All vertical bars in the same column represent the same sync mark. Thus, 70 decibels of aspiration amplitude is synchronized with the transition after the phone [t], and voicing (i.e., 60 dB of voicing amplitude) starts at the beginning of the phone [a]. This structure

1. This example only shows the second formant targets. Targets for all of the formants are generally aligned with each other. In positioning the targets, we have abstracted away from small timing differences between formants in natural speech, where the formants for a phone do not always reach their targets at precisely the same point (Kewley-Port, 1982). In our experience, these differences are not important for intelligibility, though some of them may result in improved naturalness. We believe that such refinements can be made by low-level adjustments to the positioning of the relevant targets after all the basic duration rules have applied. Any such adjustments that might be needed are a minor concern compared to the considerable advantages for rule development of using a uniform structure for the timing of formant targets.



tr = transition; 1 inch = 105 ms

Figure 1. Spectrograms of *tot* and *dot*.

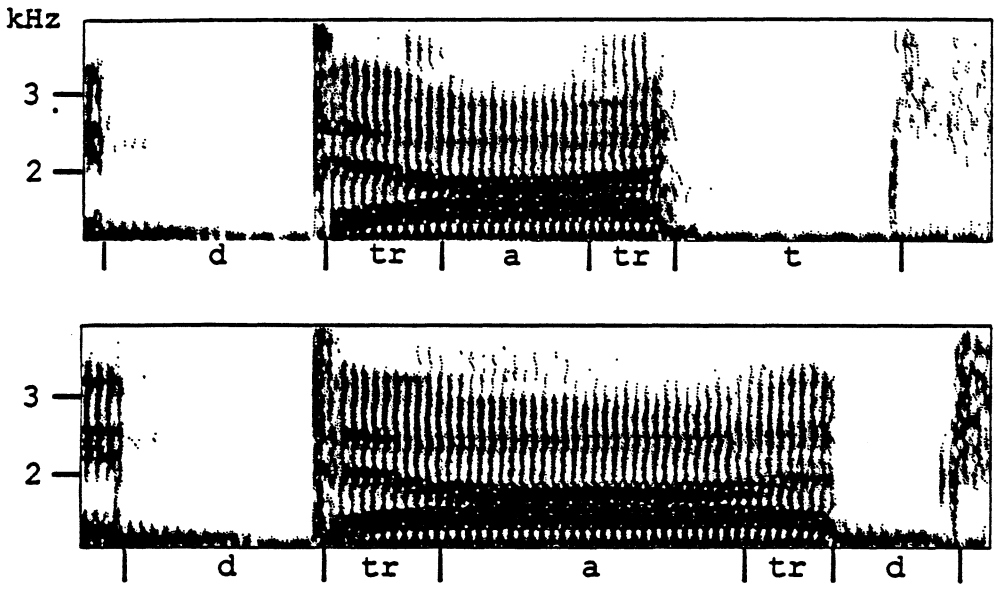
may be contrasted with that of *dot*, in which voicing starts at the beginning of the transition following [d], and there is no aspiration:

(3) *dot:*

phone:	|d		|a		|t	
F2:	|1800|		|1300|		|1700|
aspiration:	|0							
voicing:	|0	|60				|0	
transition:	|	|trans|		|trans|		
millisecond:	|85	|50	|55	|50	|95	

With independent phones and transitions, then, we can express the timing behavior of aspiration and voicing straightforwardly. More conventional models, without independent transitions (e.g., Klatt, 1979; Hertz, 1982), have led to arbitrary treatments of aspiration and to unnecessarily complicated rules, as discussed in Hertz (1991).

Explicit phones and transitions also lead to a straightforward account of vowel lengthening. For example, when a vowel lengthens before a tautosyllabic voiced stop, the lengthening occurs almost exclusively in the phone, with the adjacent transitions lengthening relatively little or not at all, as shown by the spectrograms of *dot* [dat] and *Dodd* [dad] in Figure 2. It is interesting to note that although the utterances represented in these spectrograms were articulated very carefully, with the lengthening for [dad] even slightly exaggerated, the lengthening is nonetheless restricted primarily to the phone [a], with the initial transitions in both utterances having virtually identical durations.



tr = transition; 1 inch = 105 ms

Figure 2. Spectrograms of *dot* and *Dodd.*

While the final transitions have different durations, we attribute the difference to early cessation of voicing during the movement from [a] to [t], rather than to lengthening before [d], as discussed further in Section 3.

The following deltas of *dot* and *Dodd* clearly indicate that the lengthening before [d] occurs in the phone:

(4) *dot:*

```
phone:       |d  |      |a   |      |t  |
transition:  |   |trans|    |trans|   |
millisecond: |85 |50   |55  |50   |95 |
```

Dodd:

```
phone:       |d  |      |a   |      |d  |
transition:  |   |trans|    |trans|   |
millisecond: |85 |50   |110 |50   |50 |
```

The durations in the deltas are the values produced by our rules, not the precise values in the example spectrograms. As discussed in Section 2, our rules are based on general patterns observed across many tokens, not properties of any specific utterance.

Based on observations about timing patterns such as those for aspiration and vowel lengthening, we have adopted independent phones and transitions as the basis for the duration rules, which account for a wide range of timing patterns in GA. We have found similar considerations to motivate independent phones and transitions for other American English dialects and for other languages. For example, we have also observed relatively

stable consonant-vowel transitions in Japanese and Hindi syllables with phonemically (and phonetically) long and short vowels. Hertz (1990a) illustrates how independent phones and transitions allow us to make generalizations across dialects and languages.

2. Methodology

The present duration rules for GA are based on an extensive study of the speech of a female speaker of GA, although the basic principles underlying the model have all been observed for other speakers of the dialect. The primary methodology being used to develop the model is to formulate and test hypotheses cyclically, alternating between analysis and synthesis. In the analysis phase of each cycle, we study natural speech data, primarily by examining spectrograms, and develop hypotheses concerning the underlying phonetic structure of the utterances. To help develop these hypotheses, we segment spectrograms into phones and transitions (see below), and measure and compare their durations. (Spectrograms are made with the Kay real-time DSP Sona-Graph (model 5500) and Entropic System's Waves/ESPS speech analysis software running on a Sun Workstation.)

The synthesis phase of the rule development methodology is carried out with Eloquent Technology's Delta System, a software system for phonology and phonetics that is centered around the multi-stream delta data structure for representing utterances (Hertz, 1988a, 1990b; Charif, Hertz, & Weber, 1992). The Delta System includes an interactive environment with which deltas can be built "by hand," and a programming language for writing formal rules. In the synthesis phase, we use both components of the system to write specific rules and build utterance representations that embody the hypotheses formed during the analysis phase. Speech is then synthesized on the basis of the deltas, and the synthetic output is evaluated. Among other things, evaluation includes listening to the synthetic speech back-to-back with natural speech, and visually comparing spectrograms of synthetic and natural speech. We also administer periodic intelligibility tests.

To expedite rule development in the future, we are currently implementing a multi-dialect linguistic/acoustic database, which we can query to extract generalizations for a number of American English dialects and a number of languages. The database contains both linguistic information (syllable structure, phonemic structure, degrees of stress, etc.) and acoustic values (durations of phones and transitions; formant values at the edges of phones; periods of voicing, aspiration, noise, and nasalization; and so on) for a large number of utterances. We are using the Waves/ESPS speech analysis software to

segment spectrograms into phones and transitions, to mark other relevant information, and to automatically extract formant values. We then enter the durations of the phones and transitions and the other information obtained with Waves/ESPS into the database, using a program that we wrote. The database is currently implemented with Borland's Paradox relational database software on an IBM PC networked to the Sun Workstation on which the initial speech analysis is done.

3. The Duration Model

The duration rules are based on the premise that a unit that we tentatively call the *acoustic nucleus*, which contains specific phones and transitions as sub-units (see below), serves a basic organizational role within the durational system. For example, within the acoustic nucleus, there is a trading relationship among the phone and transition durations (which we believe holds cross-linguistically) such that the total of the phone and transition durations will yield a relatively constant duration for a nucleus of a given type. Similarly, certain processes that change duration in English, such as lengthening before voiced obstruents, operate on the component phones (and in some cases also the transitions) in the acoustic nucleus in such a way that the duration of the nucleus as a whole is modified by the appropriate amount.

This section presents an overview of the rules that generate durations for stressed syllable nuclei in our current synthesis program for GA. The first subsection motivates the acoustic nucleus unit on which the rules are based; the second subsection discusses the actual rules.

3.1. The Acoustic Nucleus

As a point of departure, consider the duration of vowel phones and transitions in monosyllables with the structure C_1VC_2, where C_1 is a voiced consonant, V a vowel, and C_2 a voiceless stop (henceforth "final voiceless stop monosyllables"). In such syllables, we observe that the longer the adjacent transitions are, the shorter the vowel phone is, with the total duration of the transitions and the phone remaining relatively constant in a given context, as can be seen in the spectrograms of *beat* and *wheat* in Figure 3.

In both *beat* and *wheat*, the total duration of the phone [i] and the transitions on either side is about 100 ms. This duration, however, is distributed differently among the transitions and the [i] in the two cases. In *beat*, the transition from [b] to [i] is about 10 ms long, and [i] itself is 85 ms. In *wheat*, on the other hand, the transition from [w] to [i] is about 40 ms long, and [i] itself is 55 ms. The final transitions into [t] are about 5

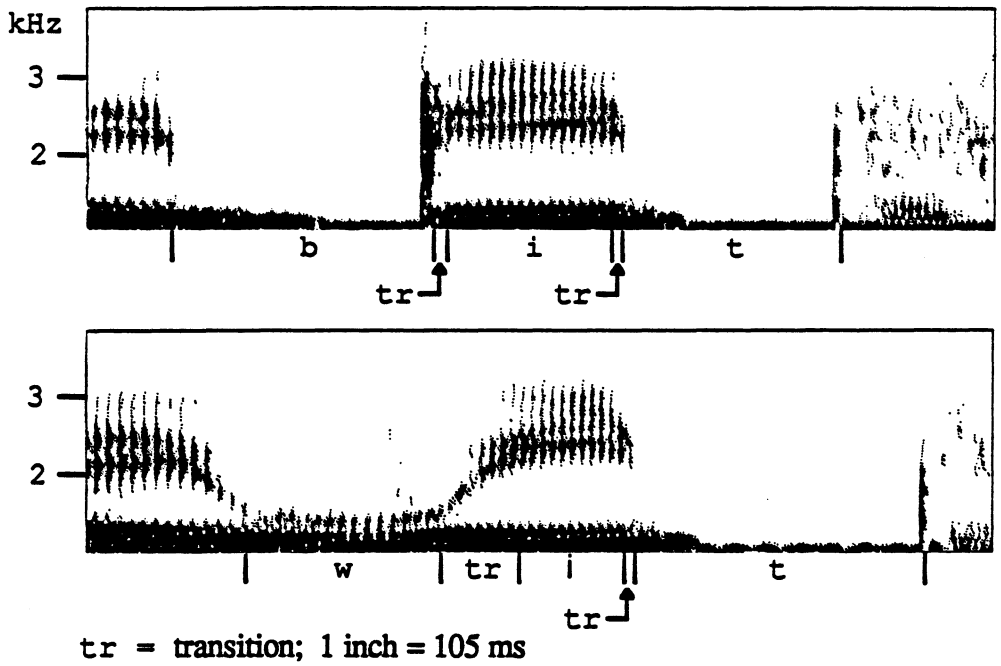


Figure 3. Spectrograms of *beat* and *wheat*.

ms long in each case. Thus, in *beat*, the total transition duration is 30 ms longer than in *wheat*, and the [i] is 30 ms shorter.

Since the transitions together with the vowel phone have a relatively constant duration, it is convenient to group the transitions plus the phone into a higher level unit of duration, which, as mentioned above, we call the *acoustic nucleus*, or simply *nucleus* (although we recognize that our use of this term differs from its various senses in phonology). Consider, for example, the following deltas for *beat* and *wheat*:

(5) *beat*:

nucleus:		nuc				
phone:	b		i		t	
transition:		trans		trans		
millisecond:		10	85	5		

100

wheat:

nucleus:		nuc				
phone:	w		i		t	
transition:		trans		trans		
millisecond:		40	55	5		

100

Note the different distribution of the total nucleus duration among the components of the nucleus in the two cases.

With an explicit nucleus unit, then, we can express the trading relationship between phones and transitions straightforwardly. In addition, we can easily write rules that modify the durations of phones and transitions differentially. For example, we can model lengthening before voiced obstruents with a rule that stretches the phone in the nucleus, but not the neighboring transitions, when the nucleus precedes a voiced obstruent (see Section 1 above).

It is not only vowel phones, however, that lengthen in syllables ending in a voiced obstruent, but also any following tautosyllabic glides and liquids, as discussed in Hertz (1991; see also Chen, 1970). For example, in words like *wide* [wayd], *willed* [wIld], and *wild* [wayld], the vowel *and* the following [y] and/or [l] lengthen.[2] We thus include in the nucleus not only the vowel of the syllable, but also any following glides and liquids (along with the relevant transitions), as illustrated by the following deltas for the words *white* and *wilt*:

(6) *white*:

```
nucleus:     |   |nuc                          |   |
phone:       |w  |      |a  |      |y  |      |t  |
transition:  |   |trans|   |trans|   |trans|   |
```

wilt:

```
nucleus:     |   |nuc                          |   |
phone:       |w  |      |I  |      |l  |      |t  |
transition:  |   |trans|   |trans|   |trans|   |
```

Given such structures, the lengthening rule can be expressed quite simply as a rule that lengthens all phones within a nucleus that precedes a voiced obstruent. Additional motivation for the nucleus comes from the fact that the precise degree of lengthening is constrained by the overall nucleus duration, as discussed in Section 3.2.

3.2. The Rules

The duration rules begin by assigning to each acoustic nucleus a duration typical of the duration of the nucleus in a final voiceless stop monosyllable uttered in the frame *Say ____for me.* (Syllables occurring in this frame will henceforth be referred to as "phrase-medial.") The rules then modify the starting durations according to the actual context. Since we have observed consistent differences among the durations of nuclei before different voiceless stops (e.g., nuclei before [p] tend to be shorter than nuclei before [t]

2. The transitions seem to be more stable in some sonorant sequences than others, but the general principle that most of the lengthening is in the phones holds across the various sequences. We are still examining the factors that determine whether transitions will lengthen and by how much.

or [k]), we have arbitrarily selected monosyllables ending in [t] for the starting durations.[3]

The starting duration of the nucleus (transitions plus phones) depends on the particular phones in the nucleus. For example, it is well-known for single-phone nuclei that a nucleus containing a mid tense vowel ([e] or [o]) is longer than a nucleus containing a high tense vowel ([i] or [u]), all other factors being equal. Similarly, a nucleus containing a tense vowel is longer than a nucleus containing the corresponding non-tense vowel (e.g., a nucleus with [e] is longer than one with [E]), and nuclei containing low vowels are longest, everything else being equal (Peterson & Lehiste, 1960).

The rules distribute the nucleus duration among the component phones and transitions as follows. First, durations are assigned to the transitions and to any non-vowel phones (i.e., glides and/or liquids) in the nucleus, with rules that depend on the features of the constituent phones. Then the vowel of the nucleus is assigned a duration by subtracting the total duration of the transitions and non-vowel phones (henceforth the "total non-vowel duration") from the total nucleus duration.

Consider, for example, the nucleus of *white*, which contains the two phones [a] and [y]. First, the rules assign to the nucleus a duration of 145 ms, the default starting duration assigned to two-phone nuclei. (A few two-phone nuclei—particularly those consisting of tense vowels followed by [l]—are assigned longer durations.) Next, the rules assign specific durations to the transitions and to the phone [y], as shown below:

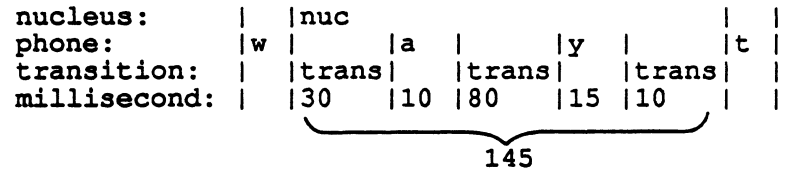
(7) *white:*

nucleus:	|	|nuc					|	|
phone:	|w |		|a |		|y |		|t |
transition:	|	|trans|		|trans|		|trans|		
millisecond:	|	|30	|	|80	|15	|10	|	|.

The rules then give the phone [a] the duration needed to bring the total nucleus duration to 145 ms—that is, 10 ms:

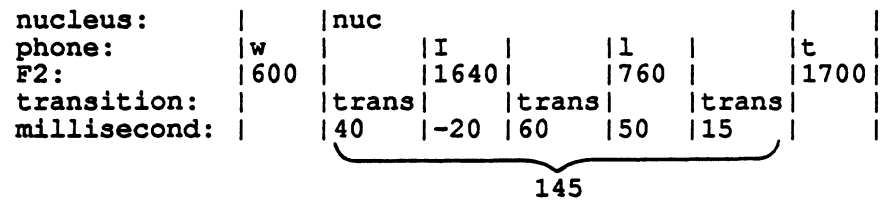
3. While the choice of the starting environment is somewhat arbitrary, there is preliminary evidence from our dialect studies that using the durations in voiceless stop monosyllables will allow for the simplest description of durational differences among dialects. In particular, we have observed in preliminary studies of Black English and some Southern dialects, that the vowel durations in final voiceless stop monosyllables are similar to those in GA, while the vowel durations in other contexts (e.g., before voiced stops) are very different. If this preliminary result turns out to hold across dialects, it will justify considering the durations of acoustic nuclei in voiceless stop monosyllables to be basic in some sense, and hence, to be the appropriate starting durations. The dialects would then be considered to differ primarily in degrees of lengthening of the nuclei, rather than shortening.

(8) *white:*



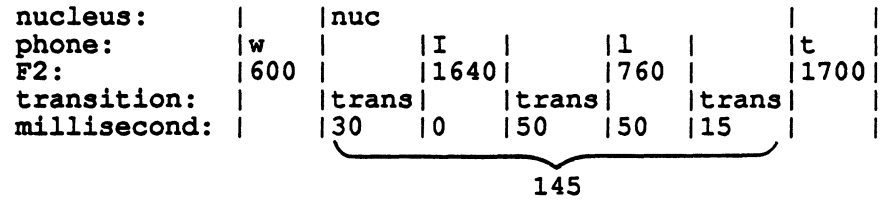
Now consider the word *wilt.* As in *white*, the rules first assign a total nucleus duration of 145 ms. Then they assign durations to the transitions and to the [l]. In this case, however, the total non-vowel duration is 165 ms, which happens to be 20 ms greater than the starting nucleus duration. Thus, when the non-vowel duration is subtracted from the total nucleus duration, the vowel receives a "negative duration" of –20 ms, as shown below:

(9) *wilt:*



Negative durations are overridden with the appropriate positive durations in most lengthening contexts. When a vowel phone still has a negative duration after all the duration rules have applied, the vowel is given a duration of 0 ms, and the rules (still tentative) shorten selected transitions in the nucleus by the appropriate amounts. In the case of *wilt*, for example, no further duration rules apply to the nucleus, so the transitions between [w] and [I] and between [I] and [l] are each shortened by 10 ms, yielding:

(10) *wilt:*



Note that since [I] has a duration of 0 ms, its second formant value of 1640 Hz functions as a durationless target. The second formant pattern moves from the 600 Hz target of the [w] over 30 ms to the 1640 Hz target of [I]. From there it moves immediately over 50 ms to the 760 Hz target of the [l], which is held for 50 ms before the pattern moves on to the 1700 Hz target of the [t].

In all of the examples so far, the transitions have been treated as part of the nucleus.[4] This, however, has been a slight oversimplification, since it is actually only voiced portions of transitions that contribute duration to the nucleus, and are therefore included in the computation of the total non-vowel duration. The transition into the final [t] of *wilt* is only voiced for a small fraction of its duration. (The early cessation of voicing partway through transitions into voiceless obstruents is discussed in Hertz (1991; see also Klatt, 1976). However, since this transition is so short to begin with, this detail does not significantly affect the duration computations presented for *wilt* above.

Let us now compare the words *tot* and *dot*. For each of these words, the rules assign a total nucleus duration of 140 ms, but the total non-vowel duration differs considerably. In *dot*, the initial 50 ms transition is voiced, but only the first 35 ms of the final transition is voiced. The total non-vowel duration is thus 85 ms. Now consider *tot*, which is like *dot* in all respects except that the initial transition is aspirated and not voiced, and therefore does not contribute duration to the nucleus. Thus the total non-vowel duration for *tot* is only 35 ms (the duration of the voiced portion of the final transition), and the vowel receives a starting duration of 105 ms (140 – 35). The different starting durations of the vowels of *tot* and *dot* are illustrated in the following deltas, which also show the cessation of voicing 35 ms into the final transition:

(11) *tot:*

```
nucleus:      |         |nuc      |       |
phone:        |t |      |a    |         |t |
aspiration:   |0 |70    |0                |
voicing:      |0         |60       |0      |
transition:   |   |trans|     |trans  |   |
millisecond:  |   |70    |105 |35  |15 |   |
                           \_____/
                             140
```

dot:

```
nucleus:      |   |nuc            |       |
phone:        |d |       |a   |         |t |
aspiration:   |0                          |
voicing:      |0 |60              |0      |
transition:   |   |trans|    |trans  |   |
millisecond:  |   |50    |55 |35  |15 |   |
                   \_____________/
                         140
```

In reality, we find that the vowel of *tot* is actually slightly shorter than 105 ms (as shown in Example (2) above), and consequently, the final total nucleus duration in *tot* is

4. We are still investigating how best to treat transitions between vowels of words like *neon*, in which the transition between the [i] and [a] could be grouped with the first or second nucleus, or a portion of the transition could be included with each.

slightly shorter than in *dot*. We attribute this difference to an independent rule that shortens all vowels after aspiration by 20 ms, whether the aspiration results from a voiceless stop or from [h] (cf. Peterson & Lehiste, 1960).

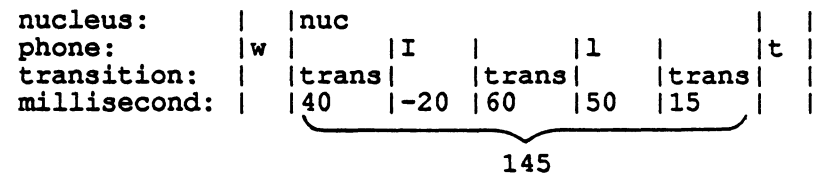
Unlike the lengthening before tautosyllabic voiced consonants, the shortening after aspiration is local to the first phone of the nucleus, regardless of the phone or the overall structure of the nucleus. There are other rules that modify the durations of specific phones in the nucleus, without regard to the overall nucleus structure. For example, word-final vowels, liquids, and glides are lengthened.

Let us now consider in more detail how syllable nuclei lengthen before tautosyllabic voiced stops. In general, the amount of lengthening of the nucleus depends on its structure. Nuclei that have relatively short starting durations (e.g., those consisting of a single non-low vowel) tend to be about 1.5 times as long as they are before a tautosyllabic voiceless stop, while long nuclei (e.g., those consisting of three phones or a tense vowel followed by [l]) tend to be about 1.3 times as long. Rather than handling these cases with different lengthening rules, we posit maximum durations for different nuclei before voiced stops, and stretch the nuclei to 1.5 times their starting duration or to their maximum duration, whichever is less. While the linguistic evidence does not help us choose between these two approaches, a maximum duration is more manageable from a programming point of view, since we can put durations measured directly from spectrograms into the rule program as the maximum durations, rather than having to compute the appropriate percentages of lengthening for the various individual cases. Maximum durations also allow for easier comparisons of duration rules across dialects.

It is important to note that a maximum duration is specific to a given context. Thus, for example, a three-phone nucleus before a tautosyllabic voiced stop stretches to a maximum duration that is less than 1.5 times its starting duration, but before a tautosyllabic voiced fricative, it stretches to more than 1.5 times its starting duration.

In general, a nucleus is stretched by lengthening the phones in it to attain the appropriate nucleus duration. Consider, for example, the nuclei of *wilt* and *willed*. Initially, both words are assigned the durations shown below for *wilt*, in accordance with the strategy discussed above (see Example (9)).

(12) *wilt:*



In *willed*, the desired nucleus duration is 210 ms, the maximum duration for phrase-medial nuclei consisting of lax vowels followed by [l] before a tautosyllabic voiced stop. This represents an increase in duration of 65 ms over the 145 ms duration of the same nucleus before a tautosyllabic voiceless stop. Since, for all intents and purposes, all of the lengthening occurs in the phones, the total duration of the phones [I] and [l] must be increased from the initial total duration of 30 ms (see the above delta) to 95 ms (30 + 65) to achieve the appropriate nucleus duration, as shown below:

(13) *willed:*

nucleus:	|	|nuc							|	|
phone:	|w |		|I	|		|l	|		|d |
transition:	|	|trans|			|trans|			|trans|		|
millisecond:	|	|40	|25	|60		|70	|15		|	|

210

We have written a preliminary set of rules that distributes the total phone duration among the individual phones in a complex nucleus in a lengthening environment. Since these rules are still undergoing revision as we explore more data, they are not discussed further in this paper.[5]

So far we have seen examples of phrase-medial syllables, in which only the phones of the nucleus are lengthened. In phrase-final position, in contrast, not only the phones, but also the transitions tend to be lengthened, though transitions are not lengthened nearly as much as phones. As a first approximation, the rules currently lengthen all transitions in a phrase-final nucleus by 40%. They then lengthen the nucleus to 2.5 times its present duration (i.e., its duration after any other lengthening rules have applied, such as the rule for voiced stops), though since nuclei with particular structures are given maximum durations, the effective lengthening is sometimes less. While in other contexts, only a few maximum durations need to be posited, and the cases are quite general, phrase-finally there are several very specific nucleus structures that seem to require different maximum durations. For example, a phrase-final nucleus consisting of the vowel [I] followed by a voiceless stop has a maximum duration of 120 ms. That is, although it is very short to begin with, it nevertheless is not lengthened beyond a certain

5. In the model described in Hertz (1991), the phones of complex nuclei (which were never assigned negative durations) were lengthened before voiced stops by different percentages, depending on their linguistic features, without regard to the overall nucleus duration. The result of this strategy was that the final durations of nuclei such as those of *bead* and *weed* were often quite different. In our current model, the nuclei of such words would have the same duration, which much more closely accords with the facts. The average duration of the nuclei may actually differ by about 5 ms, but certainly not by as much as our earlier model would have predicted. As we put more GA data in the database, we will refine our present model to correctly predict any systematic small differences that we may have missed from our more informal analyses of spectrograms.

point. There are several other cases that are similar, often involving high vowels and vowels before voiceless stops.

While we have focussed on rules for stressed monosyllables, there are a number of rules that adjust nucleus durations in other contexts. Among these are rules that shorten the nuclei of stressed monosyllables that are not word-final, and rules that shorten unstressed syllables and function words.

4. Final Remarks

This report has presented the basic premises of a new model of duration for General American English. This model is centered around the notion of an acoustic nucleus organized into phones and transitions. The model is leading to much simpler synthesis rules than the more conventional model that we employed in our previous synthesis rules for English (Hertz, 1982, 1988b). It is also serving as a fruitful basis for our cross-dialect and cross-language investigations, in which we are trying to separate universal rules from language-specific ones (Hertz, 1990a).

The development of the specific duration rules for General American English is still in progress, and future analysis and synthesis cycles are likely to lead to new or revised rules, so specific rules and algorithms discussed in this paper should be regarded as tentative. We do not, however, expect to make changes in the basic premises of the model.

Acknowledgments

Many thanks to Marie Huffman, Abby Cohn, Allard Jongman, and David Lewis for helpful comments on drafts of this paper. The development of the duration model has been supported in part by (1) the U.S. Dept. of Education, under Contracts RS89071002 and RS90087003 to Eloquent Technology, Inc., the author's company; (2) the NIH under Grant 1 R43 DC00758-01 to Eloquent Technology, Inc; and (3) the New York State Science and Technology Foundation, under grant RDG 89174 to Cornell University. The content of this publication does not necessarily reflect the views or policies of these agencies.

References

Charif, R., Hertz, S. R., and Weber, T. (1992). *The Delta User's Manual*, Ithaca, New York: Eloquent Technology, Inc.

Chen, M. (1970) Vowel Length variation as a function of the voicing of the consonant environment, *Phonetica* **22**, 129-159, 1970.

Hertz, S. R. (1982) From text to speech with SRS, *Journal of the Acoustical Society of America*, **72**, 1155-1170.

Hertz, S. R. (1988a) 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. (1988b) *SRS Phoneme-to-Speech Rules for English*, Ithaca, New York: Eloquent Technology, Inc.

Hertz, S. R. (1990a) A modular approach to multi-language and multi-dialect speech synthesis, Proceedings of the ESCA Conference on Speech Synthesis.

Hertz, S. R. (1990b) The Delta programming language: an integrated approach to non-linear phonology, phonetics, and speech synthesis. In *Papers in Laboratory Phonology I: Between the Grammar and the Physics of Speech* (J. Kingston & M. Beckman, editors), Cambridge: Cambridge University Press.

Hertz, S. R. (1991) Streams, phones and transitions: toward a new phonological and phonetic model of formant timing, *Journal of Phonetics* **19**, 91-109.

Kewly-Port, D. (1982) Measurement of formant transitions in naturally produced stop consonant-vowel syllables, *Journal of the Acoustical Society of America*, **72**, 379-389.

Klatt, D. H. (1976) Linguistic uses of segmental duration in English: acoustic and perceptual evidence. *Journal of the Acoustical Society of America*, **59**(5), 1208-1221.

Klatt, D. H. (1979) Synthesis by rule of segmental durations in English sentences. In *Frontiers of Speech Communication Research* (B. Lindblom & S. Ohman, editors), New York: Academic Press.

Peterson, G. and Lehiste, I. (1960) Duration of syllable nuclei in English, *Journal of the Acoustical Society of America*, **32**, 693-703.