QUANTITATIVE MODELING OF SEGMENTAL DURATION

Jan P. H. van Santen

AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill, NJ 07974-0636, U.S.A.

ABSTRACT

In natural speech, durations of phonetic segments are strongly dependent on contextual factors. Quantitative descriptions of these contextual effects have applications in text-to-speech synthesis and in automatic speech recognition. In this paper, we describe a speaker-dependent system for predicting segmental duration from text, with emphasis on the statistical methods used for its construction. We also report results of a subjective listening experiment evaluating an implementation of this system for text-to-speech synthesis purposes.

1. INTRODUCTION

This paper describes a system for prediction of segmental duration from text. In most text-to-speech synthesizer architectures, a duration prediction system is embedded in a sequence of modules, where it is preceded by modules that compute various linguistic features[1] from text. For example, the word "unit" might be represented as a sequence of five feature vectors: (< */u/, word - initial, monosyllabic*, $\cdots$, >) $\cdots$ (< $/t/_{burst}$, *word - final, monosyllabic*, $\cdots$, >). In automatic speech recognition, a (hypothesized) phone is usually annotated only in terms of the preceding and following phones. If some form of lexical access is performed, more complete contextual feature vectors can be computed.

Broadly speaking, construction of duration prediction systems has been approached in two ways. One is to use general-purpose statistical methods such as CART[2] or neural nets. In CART, for example, a tree is constructed by making binary splits on factors that minimize the variance of the durations in the two subsets defined by the split [2]. These methods are called "general purpose" because they can be used across a variety of substantive domains.

There also exists an older tradition exemplified by Klatt [3, 4, 5] and others [6, 7, 8, 9] where duration is computed with duration models, i.e., simple arithmetic models specifically designed for segmental duration. For example, in Klatt's model the duration for feature vector $\mathbf{f} \in \mathbf{F}$ is given by

$$\mathrm{DUR}(\mathbf{f}) = s_{1,1}(f_1) \times \cdots \times s_{1,n+1}(f_{n+1}) + s_{2,n+1}(f_{n+1}). \quad (1)$$

Here, f_j is the j-th component[3] of the vector $\mathbf{f}$, the second subscript (j) in $s_{i,j}$ likewise refers to this component, and the first subscript (i) refers to the fact that the model consists of two *product terms* numbered 1 and 2. The parameters $s_{i,j}$ are called *factor scales*. For example, $s_{1,1}$ $(stressed)$ $=$ 1.40. All current duration models have in common that they (1) use factor scales, and (2) combine the effects of multiple factors using only the addition and multiplication operations. The general class of models defined by these two characteristics, *sums-of-products models*, has been found to have useful mathematical and statistical properties [10].

Briefly, here is how these two standard approaches compare with ours. We share with general-purpose statistical methods the emphasis on formal data analysis methods, and with the older tradition the usage of sums-of-products models. Our approach differs in the following respects. First, although we concur with the modeling tradition that segmental duration data – and in particular the types of interactions one often finds in these data – can be accurately described by sums-of-products models, this class of models is extremely large so that one has to put considerable effort in searching for the most appropriate model.[4] The few models that this tradition has generated make up a vanishingly small portion of a vast space of possibilities, and because they have not been systematically tested against these other possibilities [11] we should consider the search for better models completely open. Second, in contrast with the general-purpose methods approach, the process by which we construct our prediction system is not a one-step procedure but is a multi-step process with an important role being played by various forms of exploratory data analysis.

[1] We define a *factor*, F_j, to be a partition of mutually exclusive and exhaustive possibilities such as {*1-stressed, 2-stressed, unstressed*}. A feature is a "level" on a factor such as *1-stressed*. The *feature space* $\mathbf{F}$ is the product space of all factors: $F_1 \times \cdots \times F_n$. Because of phonotactic and other constraints, only a small fraction of this space can actually occur in a language; we call this the *linguistic space*.

[2] Classification and Regression Trees [1].

[3] In its original form, the Klatt model uses p for the phonetic segment factor where we use f_{n+1}.

[4] For example, for two factors there are already five models: $s_{1,1} \times s_{1,2}$, $s_{1,1} + s_{2,2}$, $s_{1,1} \times s_{1,2} + s_{2,1}$, $s_{1,1} \times s_{1,2} + s_{2,2}$, and $s_{1,1} \times s_{1,2} + s_{2,1} + s_{3,2}$ (note the use of subscripts).

2. PROPERTIES OF SEGMENTAL DURATION DATA

In this section, we first discuss properties of segmental duration data that pose serious obstacles for prediction, and next properties that may help in overcoming these obstacles.

2.1. Interactions between contextual factors

A first reason for duration prediction being difficult is that segmental duration is affected by many interacting factors. In a recent study, we found eight factors to have large effects on vowel duration [12], and if one were to search the literature for all factors that at least one study found to have statistically significant effects the result would be a list of two dozen or more factors [13, 14, 15].

Segment	Unstressed/ Stressed	Stressed/ Unstressed	Difference	Percent
/s/	149	112	37	33
/f/	126	101	26	25
/t/$_{burst}$	71	9	62	716
/p/$_{burst}$	61	18	43	238
/d/$_{burst}$	12	7	5	67
/b/$_{burst}$	9	8	1	12
/t/$_{closure}$	75	20	55	274
/p/$_{closure}$	90	68	22	33
/n/	63	39	24	62
/m/	75	62	14	22

Table 1: Durations (in ms) of intervocalic consonants in two stress conditions: *unstressed/stressed* and *stressed/unstressed.*

These factors interact in the quantitative sense that the magnitude of an effect (in ms or in percent) is affected by other factors. Table 1 shows durations of intervocalic consonants in two contexts defined by syllabic stress: *preceding vowel unstressed / following vowel stressed* (/f/ in "before"); and: *preceding vowel stressed / following vowel unstressed* (/f/ in "buffer"; /t/ is usually flapped in this context). The Table shows that the effects of stress are much larger for some consonants than for others: a *consonant* × *stress* interaction. Other examples of interactions include *postvocalic consonant* × *phrasal position* and *syllabic stress* × *pitch accent* [12].

These interactions imply that segmental duration can be described neither by the additive model [9] (because the differences vary) nor by the multiplicative model [7] (because the percentages vary).[5] In contrast, the Klatt model was specifically constructed to describe certain interactions, in particular the *postvocalic consonant* × *phrasal position* interaction. However, in an effort to use the Klatt model for text-to-speech synthesis it became clear that this model needed significant modifications to describe interactions involving other factors [5]. Recent tests further confirmed systematic violations of the model [11].

Thus, the existence of large interactions is undeniable, but current sums-of-products models have not succeeded in capturing these interactions. General-purpose prediction systems such as CART, of course, can handle arbitrarily intricate interactions [16].

2.2. Lopsided sparsity

Because there are many factors – several of which have more than two values – the feature space is quite large. The statistical distribution of the feature vectors exhibits an unpleasant property that we shall call "lopsided sparsity". We mean by lopsided sparsity that *the number of very rare vectors is so large that even in small text samples one is assured to encounter at least one of them.*

Sample Size	Type Count	Lowest Type Frequency
20	18	13
320	254	≈ 1
5,120	1,767	< 1
81,920	5,707	< 1
1,310,720	11,576	< 1
22,249,882	17,547	< 1

Table 2: Type counts and lowest type frequencies (per million) of contextual vectors for various sample sizes.

Table 2 illustrates the concept. We analyzed 797,524 sentences, names, and addresses (total word token count: 5,868,172; total segment count 22,249,882) by computing for each segment the feature vector characterizing those aspects of the context that we found to be relevant for segmental duration. This characterization is relatively coarse and leaves out many distinctions (such as – for vowel duration – the place of articulation of post-vocalic consonants). Nevertheless, the total feature vector type count was 17,547. Of these 17,547 types, about 10 percent occurred only once in the entire data base and 40 percent occurred less than once in a million.

Two aspects of the table are of interest. The second column shows that once sample size exceeds 5,000 the type count increases linearly with the logarithm of the sample size, with no signs of deceleration. In other words, although the linguistic space is certainly much smaller than the feature space, it is unknown whether its size is 20,000, 30,000, or significantly

[5] In the additive model DUR(**f**) = $s_{1,1}(f_1) + \cdots + s_{n,n}(f_n)$; in the multiplicative model DUR(**f**) = $s_{1,1}(f_1) \times \cdots \times s_{1,n}(f_n)$.

larger than that. The third column shows that even in samples as small as 320 segments (the equivalent of a small paragraph) one can be certain to encounter feature vectors that occur only once in a million segment tokens.

It is often suspected that general-purpose prediction systems can have serious problems with frequency imbalance in the training set, in particular when many feature vectors are outright missing. Experiments performed with CART confirmed this suspicion. In a three-factor, 36-element feature space, with artificial durations generated by the Klatt model, we found that removing 66 percent of the feature vectors from the training set produced a CART tree that performed quite poorly on test data. Similarly, neural nets can have the property that decision boundaries are sensitive to relative frequencies of feature vectors in the training sample (e.g., [17]), thereby leading to poor performance on infrequent vectors.

The key reason for these difficulties is that the ability to accurately predict durations for feature vectors for which the training set provides few or no data points is a form of *interpolation*, which in turn requires assumptions about the general form of the mapping from the feature space onto durations (the *response surface*). Precisely because they are general-purpose, these methods make minimal assumptions about the response surface, which in practice often means that the duration assigned to a missing feature vector is left to chance. For example, in CART an infinitesimal disturbance can have a major impact on the tree branching pattern. Even when this has little effect on the fit of the tree to the training data, it can have large effects on which duration is assigned to a missing feature vector. In subsection 2.4, we will argue that the response surface for segmental duration can be described particularly well by sums-of-products models, so that these models are able to generate accurate durations for (near-) missing feature vectors.

It should be noted that for certain applications, in particular automatic speech recognition, poor performance on infrequent feature vectors need not be critical because lexical access can make up for errors. Current implementations of text-to-speech synthesis systems, however, do not have error correction mechanisms. Having a seriously flawed segmental duration every few sentences is not acceptable.

2.3. Text-independent variability

A final complicating aspect of segmental duration is that, given the same input text, the same speaker (speaking at the same speed, and with the same speaking instructions) produces durations that are quite variable. For example, we found that vowel duration had a residual standard deviation of 21.4 ms, representing about 15 percent of average duration. This means that one needs either multiple observations for each feature vector so that statistically stable mean values can be computed, or data analysis techniques that are relatively insensitive to statistical noise.

In large linguistic spaces, text-independent variability implies that training data may require tens of thousands of sentences, even if one uses text selection techniques that maximize coverage such as greedy algorithms[20]. And even such texts will still contain serious frequency imbalances.

2.4. Ordinal patterns in data

A closer look at the interactions in Table 1 reveals that they are, in fact, quite well-behaved, as is shown by the following patterns:

Pattern 1. The durations in the first column are always larger than those in the second column.
Pattern 2. The effects of stress – whether measured as differences or as percentages – are always larger for alveolars than for labials in the same consonant class (i.e., having the same manner of production and voicing feature).
Pattern 3. Within alveolars and labials, the effects of stress (measured as differences) have the same order[6] over consonant classes (voiceless stop bursts largest, voiced stop bursts smallest).
Pattern 4. However, the order of the durations of the consonants is not the same in the two stress conditions. For example, /t/ is longer than /n/ in the first column, but much shorter in the second column.

This pattern of *reversals* and *non-reversals*, or *ordinal pattern*, can be captured by the following sums-of-product model:

$$DUR(C, P, S) =$$

$$s_{1,1}(C) \times s_{1,2}(P) \times s_{1,3}(S) + s_{2,1}(C) \times s_{2,2}(P) \qquad (2)$$

Here, *C* is consonant class, *P* place of articulation, and *S* stress condition; it is assumed that factor scales have positive values only. It is easy to show that this model implies Patterns 1–3 (for differences). Pattern 4 is not in any way *implied* by the model, but can be *accommodated* by appropriate selection of factor scale values. This accommodation would not be possible if the second term had been absent.

There are many other factors that exhibit similarly regular ordinal patterns [11, 12, 18]. In general, factors often interact, but the interactions tend to be well-behaved so that the response surface can be described by simple sums-of-products models.

Now, showing that an ordinal pattern can be captured by a sums-of-products model does not imply that there aren't many other types of models that can accomplish the same.

[6] Except for one minor reversal: 22 ms vs. 26 ms for $/p/_{closure}$ vs. $/f/$.

Intuitively, it would appear that ordinal patterns are not terribly constraining. However, there exist powerful mathematical results that show this intuition to be wrong [19]. For example, there are results showing that if data exhibit a certain ordinal pattern then we can be *assured* that the additive model will fit. Similar results have been shown for certain classes of sums-of-products models (see [19], Ch. 7). Taken together these results make it quite plausible that when data exhibit the types of ordinal patterns often observed in segmental duration, some sums-of-products model will fit the data.

To really make the case for the importance of ordinal patterns, we must make the further key assumption that the ordinal patterns of the response surface discovered in the training data base can be fcund in the language in general (restricted to the same speaker and speaking mode). This is based on the belief that the structure discovered in the data is the result of stable properties of the speech production apparatus. For example, the non-reversal of the syllabic stress factor can be linked to the supposition that stressed syllables are pronounced with more subglottal pressure, increased tension of the vocal chords, and larger articulatory excursions than unstressed syllables. A systematic by-product of these differences would be a difference in timing.

3. SYSTEM CONSTRUCTION

We now describe construction of a duration prediction system based on sums-of-products models.

3.1. Training data

The data base is described in detail elsewhere [12]. A male American English speaker read 2,162 isolated, short, meaningful sentences. The utterances contained 41,588 segments covering 5,073 feature vector types. Utterances were screened for disfluencies and re-recorded until none were observed. The database was segmented manually aided by software which displays the speech wave, spectrogram, and other acoustic representations. Manual segmentation was highly reliable, as shown by an average error of only 3 ms (this was obtained by having four segmentors independently segment a set of 38 utterances).

3.2. Category structure

First, we have to realize that modeling segmental duration for the entire linguistic space with a single sums-of-products model is a lost cause because of the tremendous heterogeneity of this space in terms of articulatory properties and phonetic and prosodic environments. For example, the factor "stress of the surrounding vowels" was shown to be a major factor affecting durations of intervocalic consonants; however, this factor is largely irrelevant for the – barely existing – class of intervocalic vowels. Thus, we have to construct a *category structure*, or *tree*, that divides the linguistic space into categories and develop separate sums-of-products models for these categories. In our system, we first distinguish between vowels and consonants. Next, for consonants, we distinguish between intervocalic and non-intervocalic consonants. Non-intervocalic consonants are further divided into consonants occurring in syllable onsets vs. non-phrase-final syllable codas vs. phrase-final syllable codas. Finally, all of these are split up by consonant class. Note that construction of this category structure is not based on statistical analysis but on standard phonetic and phonological distinctions.

3.3. Factor relevance and distinctions

For each category (e.g., non-intervocalic voiceless stop bursts in syllable onsets), we perform a preliminary statistical analysis to decide which factors are relevant and which distinctions to make on these factors (see [12] for details).

3.4. Model selection

We already hinted that the number of distinct sums-of-products models increases sharply with the number of factors; for example, for five factors there are more than 2 billion sums-of-products models, and for the eight factors we used for modeling vowel duration there are more than 10^{76} models.[7] Thus, in cases with more than three or four factors it is computationally unattractive to fit all possible models and select the one that fits best. Fortunately, there are methods that allow one to find the best model with far less computational effort [10, 11] – requiring only 31 analyses (each the computational equivalent of an analysis of variance) for five factors. These methods are "diagnostic" because they can detect trends in the data that eliminate entire classes of sums-of-products models from consideration.

3.5. Parameter estimation

Once a sums-of-products is selected, parameters are estimated with a weighted least-squares method using a simple parameter-wise gradient technique.

4. RESULTS

4.1. Statistical fit

Forty-two sums-of-products models were constructed – one for each "leaf" of the category tree. Overall, 619 parameters were estimated (32 for vowels, 196 for intervocalic consonants, and 391 for non-intervocalic consonants). On average, each parameter was based on eight data points.

The overall correlation (over all 41,588 segments) between observed and predicted durations was 0.93 (0.90, 0.90, and 0.87, when computed separately for vowels, intervocalic con-

[7] The number of distinct models converges to $2^{2^n-1}-1$, where n is the number of factors.

sonants, and non-intervocalic consonants, respectively).

When we computed average durations for each feature vector in two equal-sized subsets of the data base, and estimated parameters for the sums-of-products model for vowels separately on each subset, the durations predicted from the two parameter sets correlated 0.987. Similarly, when we estimated parameters from data obtained on a second (female) speaker, male durations (feature vector means) were predicted with a correlation of 0.96.

In addition to these correlational findings, we also found that the key interactions were mimicked closely by the predicted durations (e.g., see Figs. 14–16 in [12]).

4.2. Text-to-speech synthesizer evaluation

A new duration module for the AT&T Bell Laboratories text-to-speech synthesizer was written based on the 42 sums-of-products models and their parameter estimates. We then compared the durations generated by the new module with those generated by the old module in a subjective listening experiment using naive listeners (see [20] for details). The old module consists of a list of several hundred duration rules similar to, but somewhat simpler than, the Klatt rules [5]. In the experiment, a listener heard two versions of the same sentence, selected the preferred version, and indicated strength of choice on a 1–6 scale (where 1 denotes complete indifference and 6 the strongest possible preference). All listeners preferred the new version. Across listeners, the new version was preferred on 73 percent of the presentations (80 percent for strength ratings of three or more). On only one of the 200 sentences was there a statistically significant majority of listeners preferring the old version; on 81 percent of the sentences listeners preferred the new version – on 60 percent with a statistically significant majority.

5. DISCUSSION

The approach taken in the paper raises some general issues that we want to briefly touch upon here.

5.1. "With Enough Data"

A general theme in our approach to modeling segmental duration is that this domain has properties distinguishing it from other domains and that this requires special-purpose methods. However, the ever-increasing amount of data that can be collected, processed, and stored, may lead one to believe that in the near future general-purpose prediction systems will be able to outperform any special-purpose system – the "With Enough Data" argument. We submit that this may rest on a misappreciation of the magnitude of sparsity encountered in certain linguistic spaces. When a training set does not provide a good number of data points for every feature vector in the linguistic space, it is unclear how general-purpose methods can be called upon to fill in the holes in the response surface without making explicit assumptions about the phenomena being modeled, or, in other words, without de facto being a special-purpose system.

5.2. Manually vs. automatically generated segment boundaries

Although manually generated phoneme boundaries have some degree of arbitrariness, there is enough overlap between various conventions to produce a remarkable degree of consensus between durational findings obtained in different studies. However, automatic speech recognition systems often produce phoneme boundaries that do not correspond to those produced manually, which may lead to very different durational behavior. For example, we found in a sub-word unit based system that vowels followed by /z/ were quite short, whereas in manually segmented data such vowels tend to be long [12]. Apparently, the training algorithm achieved higher likelihoods by putting the boundary well into the vowel. Mismatches such as these make duration models based on manually segmented data irrelevant for speech recognition. Thus, either one has to develop models for these automatically generated segment durations, or one has to constrain training algorithms to produce boundaries that correspond more closely to those generated manually.

5.3. Segments vs. other units

The final issue concerns the use of segments vs. larger units, in particular syllables. It has been suggested that not segments but syllables should play a central role in duration prediction [21, 22], the hypothesis being that speakers control durations of syllables more carefully than the durations of the segments that make up a syllable. However, the following three considerations make this proposal somewhat less appealing. First, in our factorial characterization of context, the role of the syllable is as important as that of the segment or the word. To illustrate, we define within-word position in terms of syllables (and segments, but only to distinguish between open vs. closed syllables), and within-phrase position in terms of words, syllables, and segments.

Second, there are implications from research on sub-segmental timing effects [23, 24, 25]. An example of such an effect is that the steady-state part of /a^y/ expands much more than the glide part (comparing "bite" with "bide"); in other diphthongs or in vowels, primarily the final part is stretched. Timing of some of these phenomena appears to be quite precise: Gay [23] found near-identical formant velocities across three different speaking rates. These findings urge close scrutiny of the claim that larger units are timed with more precision than smaller units. They also imply that whatever unit one selects for the lead role, timing must be specified on a fine, sub-segmental scale.

Third, it is not clear how to explain the well-documented fact that phrasal position amplifies the effects of post-vocalic voicing on vowel duration. In Campbell's [21] approach, each segment is characterized by a mean duration and an "elasticity" (variance) parameter to allow for some segments to be stretched more than others when a syllable is stretched by extra-syllabic factors. Because elasticity is assumed to be a context-independent segmental parameter, it cannot explain the amplification effect of phrasal position. Although syllable-based conceptualizations other than Campbell's might be able to address this problem, the challenge of how to specify sub-syllabic timing within a syllabic framework is clearly a serious one.

A possible resolution of the unit issue is that it may not need to be resolved. The timing pattern of speech might be viewed as the resultant of multiple constraints – some computable locally and others, say, at the paragraph level; some being inescapable consequences of the physiology of the vocal tract and others under voluntary control. These constraints could be embedded in a multi-level model where no unit or level is more central than others, but where timing is computed on a sub-segmental scale.

It should also be understood that the very concept of unit tacitly makes the concatenative assumption. This assumption is not shared by approaches based on asynchronous entities such as feature bundles [26] or formant control parameters [27]. In these systems, at any point in time more than one entity can be "on" and their on- and offsets need not coincide.

References

1. Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J., *Classification and regression trees.* Wadsworth & Brooks, Monterey, CA, 1984.
2. Riley, M. D., "Tree-based modeling for speech synthesis", In G. Bailly, C. Benoit, and T.R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pp. 265–273, Elsevier, Amsterdam, 1992.
3. Klatt, D. H., "Linguistic uses of segmental duration in English: Acoustic and perceptual evidence", *Journal of the Acoustical Society of America*, Vol. 59, 1976, pp. 1209-1221.
4. Klatt, D. H., "Review of text-to-speech conversion for English", *Journal of the Acoustical Society of America*, Vol. 82(3), 1987, pp. 737–793.
5. Allen, J., Hunnicut, S., and Klatt, D. H., *From Text to Speech: The MITalk System.* Cambridge University press, Cambridge, U.K., 1987.
6. Coker, C. H., Umeda, N., and Browman, C. P., "Automatic synthesis from ordinary English text", *IEEE Transactions on Audio and Electroacoustics*, AU-21(3), 1973, pp. 293–298.
7. Lindblom, D., and Rapp, K., "Some temporal properties of spoken Swedish", *PILUS*, Vol. 21, 1973, pp. 1–59.
8. Carlson, R., "Duration models in use", In *Proceedings of the XIIth Meeting*, Aix-en-Provence, France. International Congress of Phonetic Sciences, 1991.
9. Kaiki, N., Takeda, K., and Sagisaka, Y., "Statistical analysis for segmental duration rules in Japanese speech synthesis", In *Proceedings ICSLP '90*, 1990, pp. 17–20.
10. van Santen, J. P. H., "Analyzing n-way tables with sums-of-products models", *Journal of Mathematical Psychology*, Vol. 37, 1993 (In press).
11. van Santen, J. P. H., and Olive, J. P., "The analysis of contextual effects on segmental duration", *Computer Speech and Language*, Vol. 4, 1990, pp. 359–391.
12. van Santen, J. P. H., "Contextual effects on vowel duration", *Speech Communication*, Vol. 11, 1992, pp. 513–546.
13. Crystal, T. H., and House, A. S., "Segmental durations in connected-speech signals: Current results", *Journal of the Acoustical Society of America*, Vol. 83, 1988a, pp. 1553–1573.
14. Crystal, T. H., and House, A. S., "Segmental durations in connected-speech signals: Syllabic stress", *Journal of the Acoustical Society of America*, Vol. 83, 1988b, pp. 1574–1585.
15. Crystal, T. H., and House, A. S., 1990. "Articulation rate and the duration of syllables and stress groups in connected speech", *Journal of the Acoustical Society of America*, Vol. 88, 1990, pp. 101–112.
16. Hastie, T. J., and Tibshirani, R. J., *Generalized Additive Models.* Chapman and Hall, London, 1990.
17. Sabourin, M., and Mitiche, A., "Optical character recognition by a neural network", *Neural Networks*, Vol. 5, 1992, pp. 843–852.
18. van Santen, J. P. H., "Deriving text-to-speech durations from natural speech", In G. Bailly and C. Benoit, editors, *Talking Machines: Theories, Models, and Designs*, pp. 275-285, Elsevier, Amsterdam, 1992.
19. Krantz, D. H., Luce, R. D., Suppes, P., and Tversky, A., *Foundations of Measurement, Vol. I*, Wiley, New York, 1971.
20. van Santen, J. P. H., "Perceptual experiments for diagnostic testing of text-to-speech systems", *Computer Speech and Language*, Vol. 7, 1993 (In press).
21. Campbell, W. N., "Syllable-based segmental duration", In G. Bailly, C. Benoit, and T.R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pp. 211–224, Elsevier, Amsterdam, 1992.
22. Collier, R., "A comment on the prediction of prosody", In G. Bailly, C. Benoit, and T.R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pp. 205–207, Elsevier, Amsterdam, 1992.
23. Gay, Th., "Effect of speaking rate on diphthong formant movements", *Journal of the Acoustical Society of America*, Vol. 44, 1968, pp. 1570–1573.
24. Hertz, S. R., "Streams, phones and transitions: toward a new phonological and phonetic model of formant timing", *Journal of Phonetics*, Vol. 19, 1991, pp. 91–109.
25. van Santen, J. P. H., Coleman, J. C., and Randolph, M. A., "Effects of postvocalic voicing on the time course of vowels and diphthongs", *J. Acoust. Soc. Am.*, Vol. 92(4, Pt. 2), 1992, pp. 2444.
26. Coleman, J.S., "Synthesis-by-rule" without segments of rewrite-rules", In G. Bailly, C. Benoit, and T.R. Sawallis, editors, *Talking Machines: Theories, Models, and Designs*, pp. 43–60, Elsevier, Amsterdam, 1992.
27. Stevens, K. N., and Bickley, C. A., "Constraints among parameters simplify control of Klatt formant synthesizer", *Journal of Phonetics*, Vol. 19, 1991, pp. 161–174.