Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 21 (1997), 21-36

WORD-LEVEL PROSODICAL MARKING OF CONSONANT DURATION AND SPECTRAL BALANCE *

R.J.J.H. van Son and Jan P.H. van Santen [1]
[1] Bell Labs, Lucent Technologies, Murray Hill NJ, USA

Abstract

The prosodic structure of speech determines to a significant part the strengthening and weakening of segmental articulation. For prosodic boundaries, only durational and articulatory data are currently available from only a limited number of speaking styles and consonants. The durations and spectral balance of a comprehensive inventory of consonants from 4157 informative sentences read by two American English speakers, confirm published results on articulatory and durational changes from reiterant speech. New statistical methods were used to deconfound the many factors that influence consonant duration and spectral balance in read speech. It proves that the behavior of the spectral balance of consonants parallels that of their duration with respect to syllable stress and word-level structure and both could be combined into a single prosodic model. Both factors emphasized word boundaries and stressed syllables. Coronal consonants behaved irregular, which could be explained as the result of a shift from full to flap articulation in de-emphasized contexts.

1. Introduction

The prosodic structure of utterances is reflected in the segmental qualities of the constituent phoneme realizations. Quite a large body of research has shown that syllable and sentence accent strongly affect vowel duration and the level of vowel reduction (e.g., Koopmans-Van Beinum, 1980; Lindblom, 1990; Van Son and Pols, 1990, 1992; Fourakis, 1991; Van Bergem, 1995; Wang, 1997) and more recently consonant reduction (De Jong et al., 1993; De Jong, 1995; Farnetani, 1995; Van Son and Pols, 1996b). Recently it has been shown that syllable stress also affects the overall spectral balance of vowels (Sluyter, 1995 a, b; Sluyter and Van Heuven, 1996). It has also been shown that manipulating the spectral characteristics of vowels can induce the perception of syllable stress (Rietveld and Koopmans-Van Beinum, 1987; Sluyter, 1995 a, b; Sluyter et al., 1997).

It has also become clear that the presence of edges between prosodic domains is reflected in the duration and articulation of the surrounding phonetic segments (e.g., Turk and Sawush, 1997; Wightman et al., 1992; see Fougeron and Keating, 1997 for a review). However, much less is known about the influences of the presence of prosodic boundaries on the acoustics of neighboring phonemes. In their paper,

* Parts of this paper have been presented at Eurospeech'97 (Van Son and Van Santen, 1997).

Fougeron and Keating (1997) discuss two problems with the available evidence. First, the relevant studies used only a very limited inventory of phonemes (e.g., a single consonant and vowel) and quite unusual speaking styles (e.g., reiterated speech). Second, these studies are largely limited to measuring articulatory parameters for which the acoustical relevance is often opaque (e.g., linguopalatal contact, jaw or lip positions). Together, these problems make it very difficult to extrapolate the results of earlier studies to the acoustics of more common speaking styles (Fougeron and Keating, 1997).

In this paper we want to investigate two questions raised by Fougeron and Keating (1997). First, to what extent is the prosodic structure reflected in the acoustics of consonant segments of normal (read) speech? Second, is there a spectral correlate of prosodic structure that indicates a change in the articulation of consonants? The components of the prosodic structure we choose to investigate were realized syllable stress and lexical word boundaries. This choice was partly motivated by practical considerations. Both syllable stress and lexical word boundaries are the lowest level and most numerous components of prosodic structure (c.f., De Jong et al., 1993; Fougeron and Keating, 1997). In addition, the identification of syllable stress and word boundaries is important in word-recognition and, therefore, for the comprehension of speech (Cutler and Carter, 1987; c.f., Cutler et al., 1997 for a review). The boundaries between the constituent lexical items are not always reflected in the surface structure of connected speech. From a phonological viewpoint, larger units are often formed by combining several lexical items and producing them as single, prosodic, words (e.g., Wightman et al, 1992). The practical and methodological problems associated with determining prosodic word boundaries for a very large corpus of read sentences were prohibitive and prevented us from using *realized* word boundaries instead of *lexical* ones (c.f., Wightman et al, 1992).

The underlying cause of the problems identified by Fougeron and Keating (1997) can be traced to strong factor interactions (Van Santen, 1992) and a lack of general measures of spectral structure that can be used to link articulatory "effort" to the acoustics of speech over comprehensive phoneme inventories (c.f., Chennoukh et al., 1997 for a recent modeling effort). The latter problem can be pragmatically solved by using the spectral balance as studied by Sluyter and her colleagues. This spectral balance is related to speaking effort (Sluyter, 1995 a, b; Sluyter and Van Heuven, 1996; Sluyter et al., 1997). The spectral balance can be quantified by the "mean frequency", i.e., the Spectral Center of Gravity (Van Son and Pols, 1996; 1997). This measure is defined for all speech sounds. The problem of strong factor interactions, i.e., the inability to evaluate factors independent of each other, will be discussed with respect to the study of segmental duration where it has been discussed earlier.

1.1. Tackling factor confounding and strong interactions

Models of segmental duration generally use single factor independence. These models are based on the assumption that the effect of one factor on duration can be modeled without taking into account the values of the other factors (cf., Klatt, 1987; Fant and Kruckenberg, 1989; Van Santen, 1992; Wang, 1997). Interaction between factors, where the effect of one factor indeed depends on the values of other factors, is well known. For example, the effect of post-vocalic voicing on vowel duration is much larger (measured in ms or as a percentage) in pre-pausal syllables than in non-pre-pausal syllables (c.f., Klatt, 1973; Van Santen, 1992). Many other examples are described in the literature (e.g., Farnetani and Kori, 1986; Crystal and House, 1990; Eefting, 1991; Turk and Sawush, 1997). However, it is difficult to model such factor

interaction because of a lack of quantitative data (or even to be sure that these interactions really exist, c.f., Crystal and House, 1988, 1990). The problem lies in the large amount of speech data needed to resolve interactions between factors. If factors act independently, the number of "examples" needed to resolve them is roughly proportional to the sum of the value levels of these factors. In case they interact, and are not independent, the number of examples needed becomes proportional to the product of all factor levels. This product can reach quite large values. In practice it is nearly impossible to collect enough speech to cover all possible combinations of factor levels. Especially so because of factor confounding (Van Santen, 1992), the fact that some factor values have a low frequency in some contexts, e.g., in English, vowels occurring in word-initial syllables are much more likely to be stressed than word-final vowels (Cutler and Butterfield, 1990, 1991; Cutler and Carter, 1987; Van Kuijk, 1996); as a result, the former have a longer average duration than the latter. However, when properly analyzed, we find that word-final vowels are longer than word-initial vowels having the same stress level. Thus, the initial findings are deceptive.

It is to be expected that not all factors interact. The factors that affect segmental duration will be, in a first approximation, "piecewise independent" (Van Santen, 1992). This means that we can divide the set of factors into non-overlapping subgroups, such that interactions occur only between factors in a subgroup. This allows us to investigate segmental duration with less than complete coverage of all possible combinations of factor levels. Interacting factors can be spotted graphically by plotting measured values versus factor combinations. If factors act independently, all "lines" in the plot should be parallel, e.g., the differences between separate factor values like stress or speaker, should be independent of the position in the word. If the lines are not parallel, the effect of one factor depends on the values of the other factors.

There are two types of speech corpora. In one, a carefully designed ("balanced") set of sentences is recorded with the property that factor confounding is minimized (c.f., Fougeron and Keating, 1997; Wightman et al., 1992). However, this typically requires usage of repetitive carrier phrases, which may seriously undermine how naturally the text is read. In the other type, which we have used, naturally occurring meaningful sentences are used (c.f., Crystal and House, 1988, 1990). This has the advantage of a more natural reading style, but the disadvantage of creating confounding. However, under the assumption of piecewise independence, we can analyze such data without strong concerns about factor confounding.

We used a new statistical method developed at Bell-Labs (Van Santen, 1992, 1993b,c).This technique uses pairwise differences between "Quasi Minimal Pairs" to calculate "Corrected Means" that approximate the hypothetical balanced mean values, corrected with respect to the unbalanced distribution of realizations (Van Santen, 1992). Non-parametric tests can be performed on the "Quasi Minimal Pairs" to determine the statistical significance of any effects found. These corrected means are then used to model the interactions between the relevant factors with respect to consonant duration.

2. Material and methods

2.1. Consonant segments

Fully labeled and segmented speech of a professional male and female speaker was used. Segmentation had been done by professional labelers. For plosives, the start of

the burst was indicated. Labels included the presence of realized syllable stress (for non-clitics) and whether or not a word could be cliticized. Sentence accent was also marked for the speech of the female speaker. Only consonants from accented words were used for her speech. Sentence accent was not indicated reliably for the speech of the male speaker. Furthermore, much less material was available from the male speaker. Therefore, we ignored sentence accent for his speech and used all words that could carry sentence accent (which excluded all words that could be cliticized). However, it proved that the presence of unaccented words in the material of the male speaker did not alter the averaged behavior of his consonant realizations, so we will not elaborate on this difference between the material of the speakers. We also ignored all consonants from the last word of each sentence because these words are known to behave differently in many respects (Van Santen, 1992).

For practical reasons, we excluded the Glottal consonants, Affricates, and /j/ from analysis and used all VCV realizations of the 20 consonants /vfðθzsʒʃmnŋbpdtgkwlr/. Because the duration of the burst+aspiration part of a plosive varies rather independently of the closure, both parts were treated as separate segments. That is, each plosive was split into an independent closure and burst+aspiration part (indicated by italic IPA symbols). Therefore, in total we used 26 consonant types.

The speech consisted of read aloud sentences, both "normal" and phonetically "rich" meaningful sentences. In total 1206 sentences were available for the male speaker and 2951 sentences for the female speaker. Durations of all intervocalic consonants (VCV, also crossing word boundaries) of non-clitics and non-sentence final words were analyzed (only accented words for the female speaker). This resulted in 4116 VCV segments for the male speaker and 8957 VCV segments for the female speaker, of which 1430 and 3464, respectively, were plosives. All speech was recorded with a sampling frequency of 16 kHz and 16 bit resolution. Not all original sampled speech files of the female speaker were available at the time of our investigations. This left 4432 intervocalic segments of her speech for the determination of spectral characteristics (1722 plosive realizations). Five factors were selected for investigation: Consonant identity, Syllable stress (Stressed or Unstressed), position in the word (Initial, Medial, and Final), word length (in syllables: 1, 2, 3, and more), and the frontedness of the syllabic vowel (as measured by F_2: High, Middle, and Low F_2, and Diphthongs). The last two, word length and vowel frontedness, proved to have minimal or no effects in our corpus. We still kept them to define more homogeneous "cells" (see Table 1). The stress value of a consonant was defined to be the stress of the following vowel in word initial and word medial position, and to be the stress on the preceding vowel for word final consonants. This implements a maximal onset definition of the syllable. In English, there is a strong confounding between stress, position in the word, and word length (Van Santen, 1992; Cutler and Carter, 1987; Cutler et al., 1997; Van Kuijk, 1996). The use of more elaborate schemes to apply attribute stress would have required more data than was available and was, therefore, not attempted.

The *spectral slope* of a speech segment is determined by the underlying phoneme and by the syllable stress and is itself a perceptual cue for syllable stress (Sluyter, 1995 a, b; Sluyter and Van Heuven, 1996; Sluyter et al., 1997). However, it is difficult to measure the spectral slope in a uniform manner for different types of phonemes. It is better to use a measure of spectral balance which is strongly related to the spectral slope. In the present study, the *Spectral Center of Gravity* was chosen as a measure of the spectral balance of the realizations and thus as a measure of the spectral slope. The center of gravity of a spectrum is proportional to the air-speed/area in the constriction of obstruents for turbulent noise and is related to the speed of the

vocal folds at closure for sonorants. In both cases, an increase in the center of gravity of the spectrum indicates more effort (muscle tension) of the speaker.

The Center of Gravity of a spectrum (CoG) is in a sense, the "mean" frequency. It is calculated by dividing $\int f \cdot E(f) \cdot df$ by $\int E(f) \cdot df$ in which E(f) is the power spectrum. Used in this way, the CoG correlates with both consonant reduction due to stress and speaking style (Van Son and Pols, 1996) and consonant intelligibility (Van Son and Pols, 1997). For the current study, E(f) was calculated via FFT from the waveform using a Gaussian window with an effective length of 25 ms. The window was centered at the "nucleus" of the realizations as indicated by the labellers. For plosives, the analysis window was centered 12.5 ms before the release burst (stops) and 5 ms after the start of the release burst (burst+aspiration).

The CoG frequencies and variances vary widely, from around 200 Hz for nasals to over 5000 Hz for labial fricatives. This would severely distort any comparison between consonants because the variance of the larger values (e.g., fricatives) would completely dominate that of the smaller values (e.g., nasals). Therefore, we decided to express the CoG frequencies in semitones, i.e., $12 \cdot \mathrm{Log}_2(\mathrm{CoG})$, which equalizes the variances over the range.

The CoG measures for plosive stops and bursts caused problems. Plosive bursts are frequently shorter than the 25 ms window used (therefore the offset of only 5 ms in the burst). As a consequence, the power spectrum will sample part of the plosive stop and part of the aspiration and following vowel (the latter two have similar spectra). If the acoustic energy in the burst is small, the vowel spectrum will reduce the CoG. The ideal voiceless plosive stop has *no* acoustic power at all. In this ideal case of complete silence, the power spectrum will be flat and the CoG will be half the Nyquist frequency, i.e., 4 kHz (~ 144 semitones). But, any overlap with the preceding vowel will cause the vowel spectrum to determine the CoG. However, in both cases the results are that short, "weak" plosives get lower CoG values than long, "strong" plosives. As this is in line with the effects expected for the other consonants, we think there will be no problems with the interpretation of the results.

2.2. Calculating corrected means

As our speech material was not balanced, we were faced with widely varying numbers of realizations for each of the consonants with respect to all the other relevant factors. This means that "raw" means of duration or CoG cannot be compared between conditions (cf. discussion of this topic in Van Santen, 1992). The large undersampling of possible combinations of factor values and the variability in sample sizes precludes the use of normal ANOVA and MANOVA statistics. To solve this problem we used a method developed by Van Santen (1993b, see also Van Santen, 1992) based on theoretical work described in Van Santen (1993c).

Essentially, this method estimates the mean values from pairwise differences between "Quasi Minimal Pairs" (Van Santen, 1992). This way the hypothetical "balanced" mean values are approximated by "Corrected Means", corrected with respect to the unbalanced distribution of realizations. The method starts with selecting a set of relevant factors. We used six factors: Speaker (Male or Female), Consonant Identity (26), Syllable Stress (present, absent), Position in the word (Initial, Medial, or Final), Word length (1, 2, 3, and 4 or more syllables), Vowel frontedness (in F_2: low, mid, high, and diphthongs). A table is constructed with the factor values for which the average is to be calculated as the row headings, e.g., all 6 combinations of syllable stress and position in the word, and all combinations of values of the other factors as column headings, e.g., a column for /ŋ/ realizations combined with a high-F_2 vowel from two-syllabic words of the female speaker, and another column for /ð/

realizations combined with a mid-F_2 vowel from 3-syllablic words of the male speaker. Each cell contains the mean value of all realizations that conform to the row and column factor values. That is, each cell contains the mean of a set of realizations that is completely "homogeneous", at least as far as the factors used are concerned.

For the data in our study, the table contained 2·26·2·3·4·4 = 4992 cells, (i.e., 2496 for each speaker). Some columns in this table, i.e., combinations of factor values, will always be empty due to phonotactic constraints. Still, even after ignoring all empty columns, over 55% of the remaining cells were empty for our unbalanced corpus. An additional 5-10% of the cells contained only a single consonant realization. Another measure of "unbalance" is the perplexity of the rows in the table. This perplexity is a measure of the "effective" number of filled columns per row (Bahl and Jelinek, 1990). In our corpus this perplexity is typically less than one-third of the actual number of columns present. In other words, only about one-third of the table is reasonably filled with measurements, the rest is either empty or nearly empty.

Due to this extreme sparsity, standard statistical techniques (e.g., Factor Analysis, ANOVA, or MANOVA) will give results of only limited value. This is even true if realizations are pooled for individual factors. Still, by making some assumptions it is possible to unravel the influences of single or combined factors without unduly pooling data.

In this paper we assume that the factors that affect segmental duration and CoG are piecewise independent (Van Santen, 1992). This means that we model, e.g., segmental duration as: DUR(all factors) = A(row-factors) + B(column-factors), i.e., the duration as a function of all relevant factors is the sum of the effects of the row factors and the effects of the column factors. That is, the influence of the row-factors is independent of the influence of the column factors. Under this assumption, the average, pair-wise, difference between corresponding cells in any two rows should only depend on the values of the row-factors involved, and not on the values of the column factors themselves. For example, if the rows are composed of the position in the word (I, M, F) and syllable stress (+, -), it is assumed that the effect of these on duration is independent of consonant identity or word length, or any of the other factors (see Table 1). However, the effects of column factors themselves might indeed

Table 1. Example of a table used to calculate the corrected mean values of all six combinations of syllable stress (+ or –) and position in the word (Initial, Medial, and Final). There are 15 mean cell-by-cell row differences for 6 rows, e.g., (+&M) – (–&F) or (+&I) – (–&I), and therefore 15 linear equations for the 6 hypothetical mean values. Solving these 15 equations with singular value decomposition (SVD) gives the "best estimates", in a least RMS error sense, for the mean duration of each row. Note that, in general, less than 50% of the cells are filled, e.g., /ŋ/ cannot be Word-Initial, monosyllabic words have *no* Word-Medial consonants, and /ð/ did not occur in Word-Final position in our data.

	Female, /ŋ/, high-F_2, 2-syllables	Male, /ð/, mid-F_2, 3-syllables	Male, /f/, low-F_2, 1-syllable	..... 829 further columns
+ & I	–	*mean*	*mean*	.....
– & I	–	*mean*	*mean*	.....
+ & M	*mean*	*mean*	–	.....
– & M	*mean*	*mean*	–	.....
+ & F	*mean*	–	*mean*	.....
– & F	*mean*	–	*mean*	.....

be inter-dependent. The choice of the factors combined in the rows or columns depends on the expected levels of interaction. So in Table 1, it is assumed that the effects of the prosodic factors, i.e., stress and position with respect to the word boundaries, are largely independent from the other factors.

This way it is possible to calculate the average pair-wise cell differences between all pairs of rows, using only pairs of cells from the same column for which there are realizations in both rows. These average differences are strictly based on equivalent realizations and should be largely independent of the way the realizations are distributed over the columns.

Just averaging the differences between corresponding table cells for each pair of rows to calculate the mean difference between the rows, unduly increases the variance in the mean difference. The differences should be weighted to account the variation in the number of realizations in each cell. A weight of $w = 1/\sqrt{(1/\#Cell1 + 1/\#Cell2)}$ will weight each difference according to its standard error (better, its inverse). However, the exact form of the weighting function had little effect on the outcome, as long as the weights were related to the number of realizations in the cells.

The set of average differences between all possible pairs of rows constitutes a set of linear equations of the mean row values that can be solved using standard techniques (i.e., minimizing RMS-error with a Singular Value Decomposition, SVD, Press et al., 1988). The results are the relative *Corrected Means* of the rows, e.g., the corrected mean durations of the combinations of the position in the word and stress levels. For any fully balanced set of realizations, the result of this procedure would be identical to the raw means. Therefore, the corrected mean values can be interpreted as a least RMS-error approximation of "balanced" means with an unbalanced data set. Because the corrected means are calculated from differences only, they need an absolute offset value to get "real" means. We choose as the offset the overall mean duration of all realizations used for the calculation of the corrected means.

The above description was based on the assumption that the factors affected the segmental duration in an additive manner. However, if all durations are replaced by their logarithm, the resulting model of segmental duration will be multiplicative. No further changes are necessary to cover a multiplicative model of segmental duration. It showed that the results for a multiplicative model were more extreme than those for the additive model so we decided to use the more conservative additive model.

2.3. Statistical analysis

The original mean row differences are calculated from pair-wise cell differences. The statistical significance of the size of the difference between each two rows can be tested on the collection of cell-pairs used to determine this difference. Because of the unbalanced distribution of realizations over the table cells, we decided to limit the statistical analysis to a distribution-free test, the Wilcoxon Matched-Pairs Signed-Ranks test (WMPSR). Each pair of table cells was used as a single matched pair in the analysis, i.e., we did not look "inside" the table cells. Distribution-free test are generally considered to be less sensitive than tests based on the Normal distribution, e.g., the asymptotic relative efficiency of the WMPSR test with respect to the Student-t test is 0.95 when we assume a Normal distribution. However, consonant durations are not distributed Normal and we want to check the differences independent of the details of the chosen model and weighting function. Both facts together give the Wilcoxon Matched-Pairs Signed-Ranks test an advantage over the Student-t test. Using the WMPSR test on the set of differences between a pair of rows is completely independent of the weighting function used to calculate the mean difference between the rows.

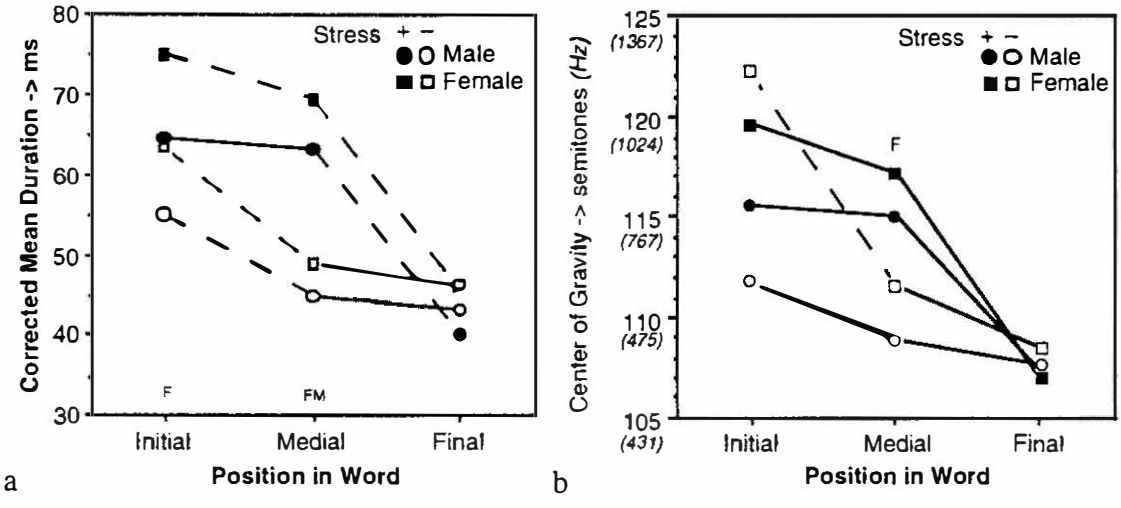

Figure 1. Corrected mean durations (a) and Spectral Center of Gravity (b) of consonants for both speakers separately. Syllable stress versus position in the word. Dashed lines and the symbols 'M' (male), and 'F' (female) indicate statistical significance: $p \leq 0.001$, two tailed WMPSR test between word positions and syllable stress conditions respectively (see text). The CoG differences between stressed word-initial and word-final consonants were statistically significant for both speakers ($p \leq 0.001$, two tailed WMPSR, Figure b).

3. Results

3.1. Syllable stress and position in the word

For both speakers we calculated the corrected mean duration of the consonants for each of the six combinations of syllable stress (stressed and unstressed) and position in the word (initial, medial, and final). The results are plotted in Figure 1.a. The results for the spectral Center of Gravity are plotted in Figure 1.b. The overall corrected mean difference between the two speakers amounted to 8.44 ms and 3.45 semitones (both differences are statistically significant, $p \leq 0.001$, two-tailed MWPSR test, not shown).

For both speakers we see that the stressed word-initial and word-medial consonants had similar durations, the difference is only 4 ms overall, although the difference was significant for one speaker (female speaker, 6 ms, $p \leq 0.001$, two-tailed WMPSR test). Stressed consonants from both Initial and Medial positions were longer than stressed consonants from a word final position (28 ms and 23 ms longer, respectively, $p \leq 0.001$, two-tailed WMPSR test). For consonants from unstressed syllables we see the opposite. Unstressed consonants from a medial and final position in the word have similar durations (2 ms difference) and both differ markedly from unstressed consonants from a word-initial position, which are around 13 ms longer. Moreover, in word-final position there is no difference in duration between stressed and unstressed consonants (1 ms difference, $p \geq 0.001$, two-tailed WMPSR test).

The CoG data largely mirror the duration data (Figure 1.b). However, we cannot ascertain that all the differences are statistically significant. The differences between stressed word-*initial* and word-*final* consonants is statistically significant for both speakers ($p \leq 0.001$, two-tailed WMPSR test). The difference between unstressed word-initial and word-medial consonants of the female speaker is also statistically significant ($p \leq 0.001$, two-tailed WMPSR test). Finally, the differences between stressed and unstressed word-medial consonants of the female speaker are statistically significant ($p \leq 0.001$, two-tailed WMPSR test). The corrected mean difference

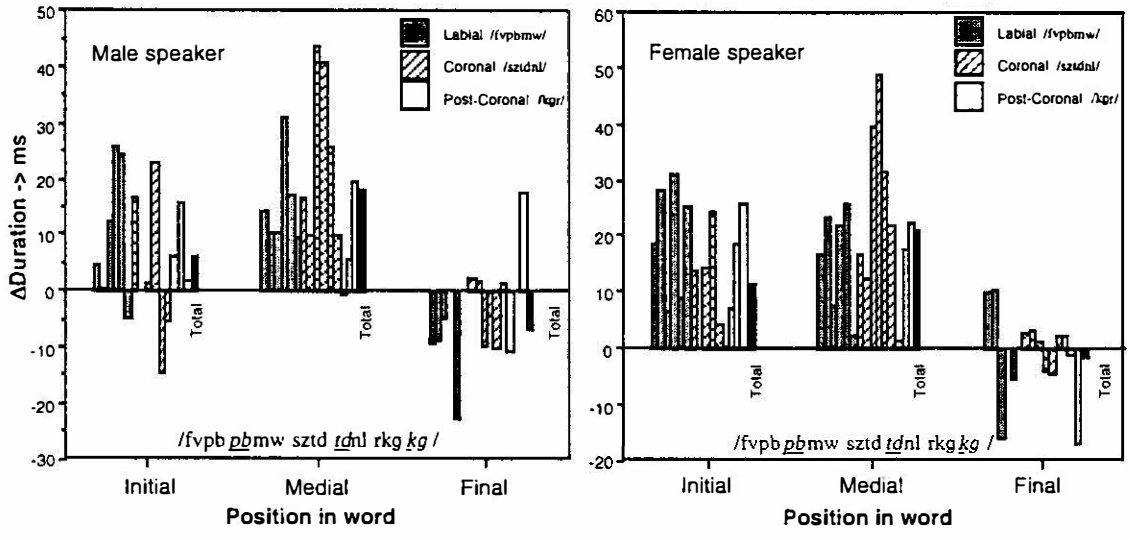

Figure 2. Differences in corrected mean duration between consonants from stressed and unstressed syllables. Because of their rarity in our corpus, /θðʒʃŋ/ are not included (however, they are used to calculate the total values). Unless a consonant did occur in both stressed AND unstressed syllables, no difference was assigned. The order of the consonants is given in the string of phonetic symbols below the graphs. /pbtdkg/: plosive closure durations, /pbtdkg/: plosive burst+aspiration durations.

between word initial and word final consonants is around 10 semitones for both stressed and unstressed consonants (both speakers pooled).

3.2. Syllable stress and prime articulator

To investigate the influence of consonant identity on the mean corrected values we focused on the corrected mean durations (see section 2.2). For realizations from each position in the word, i.e., word-initial, word-medial, and word-final, we determined the corrected mean difference of the durations between stressed and unstressed realizations of each consonant. The values are plotted in figure 2.a and b for both speakers separately. It was clear that the behavior of both speakers was quite similar.

In Figure 2 it can be seen that the behavior found for all consonants pooled was representative of the behavior of individual consonants. Differences between stressed and unstressed consonants were large in word-initial and -medial position and erratic in final position. The differences in the size of the effect of stress on duration, as displayed in figure 2, between word-initial, -medial, and -final position were all statistically significant ($p \leq 0.002$, two-tailed WMPSR test, both speakers combined). However, it is also evident that the large influence of syllable stress on consonants in word-medial position could be attributed to the behavior of the Coronal consonants, /sztd*t*dnl/ (word-medial versus word-initial, $p \leq 0.001$, two-tailed, WMPSR test on the numbers in figure 2, n=12). Here we restrict the Coronals to include only consonants with dental, alveolar and post-alveolar places of articulation. All other consonants using the tongue as prime articulator, but not included in our restricted definition of Coronals, will be pooled as Post-Coronals. Both for Labial and Post-Coronal type consonants /fvpb*p*b*b*mw/ and /kg*k*g*g*ɪ/ respectively, there was no real difference between consonant durations in Word-Initial and Word-Medial position ($p > 0.05$, n=16 and n=10, respectively). The stress related differences in duration between Coronal and Labial consonants were statistically significant for the Word-Medial position

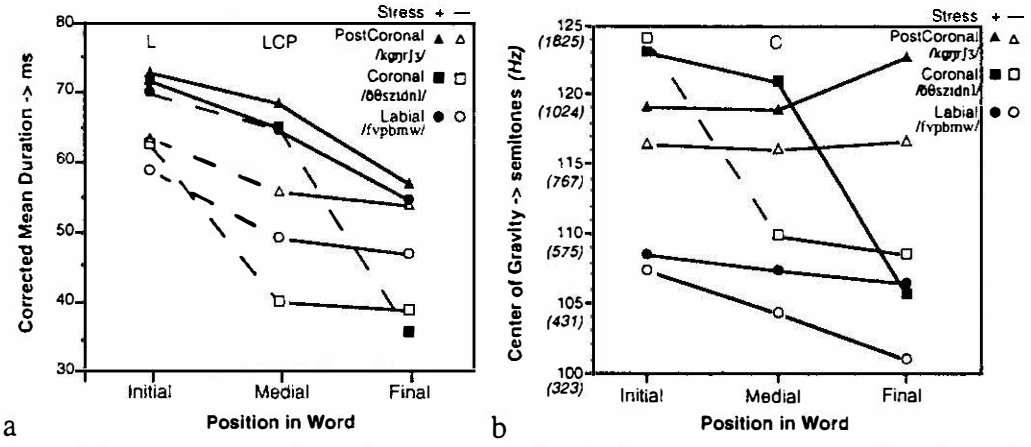

Figure 3 Corrected mean values of consonants split on Prime Articulator (i.e., Labial, Coronal and Post-Coronal, see text). Syllable stress versus position in the word for both speakers combined. Dashed lines: $p \leq 0.001$, two tailed WMPSR test between word positions. LCP: $p \leq 0.001$, two tailed WMPSR test between stressed and unstressed realizations for Labials, Coronals and Post-Coronals, respectively.

a. Duration

b. Spectral Center of Gravity. The differences between both stressed and unstressed word-initial and word-final Coronals were significant ($p \leq 0.001$).

($p \leq 0.001$, two-tailed WMPSR test on the numbers in figures 2, both speakers combined, n=16), but not for the word-initial position, ($p > 0.05$, n=12).

The differences due to the effect of the primary articulator (Labial, Coronal, or Post-Coronal consonants) were investigated by replacing the identity of each phoneme by three values: Prime articulator (Labial, Coronal, Post-Coronal), Manner of Articulation (Fricative, Plosive Stop, Plosive Burst+Aspiration, Nasal, and Vowel-Like consonants), and voicing (only for obstruents) and calculating the corrected means. The results for the primary articulator, from both speakers pooled, were summarized in figures 3.a and b.

From the corrected mean values, it is obvious that, overall, consonant duration became shorter and CoG values lower towards the end of the word and in unstressed syllables. The Coronal consonants behaved like the Post-Coronal consonants in some situations (all word-initial and stressed word-medial Coronals) but completely different in other situations (unstressed word-medial and all word-final Coronals). When inspected in detail, the picture became more complicated for the corrected mean CoG values. Overall, stressed consonant realizations seemed to have higher CoG frequencies than comparable unstressed realizations (only statistically significant for Coronals, $p \leq 0.001$, two-tailed WMPSR test). The overall corrected mean CoG value of Labial consonants was lower than that of the Coronal and Post-Coronal consonants (by 8.87 and 10.60 semitones, respectively, $p \leq 0.001$, two-tailed WMPSR test), whereas the overall corrected mean difference between Coronals and Post-Coronals was small (1.73 semitones) and only statistically significant between unstressed word-medial and word-final realizations ($p \leq 0.001$, two-tailed WMPSR test). For the Coronals, there seemed to be only two levels, a high CoG for word-initial and stressed word-medial realizations, and a low CoG for unstressed word-medial and all word-final realizations. This was more extreme than the picture found for the durations (Figure 3.a).

Figures 3.a and b suggested that the corrected mean CoG values and durations were correlated. A direct correlation between individual durations and CoG values was not informative due to the unbalanced nature of our corpus. However, the correlation between the corrected mean values *could* be determined. This correlation

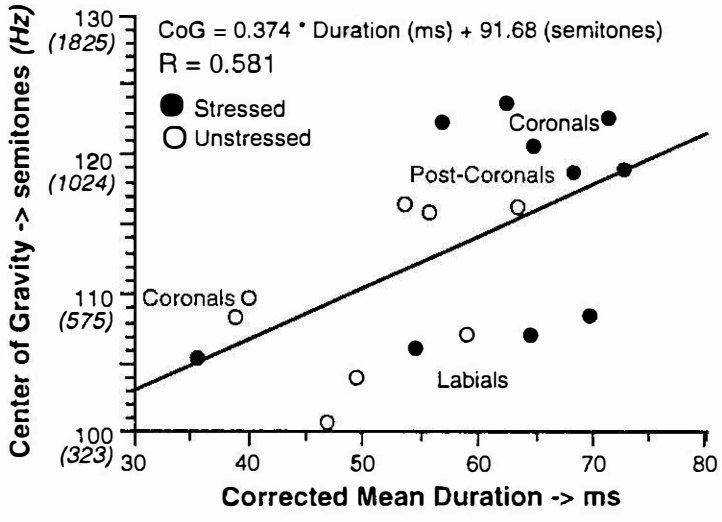

Figure 4. Correlation between the corrected mean frequencies of the Spectral Center of Gravity (semitones) and the corrected mean durations. Plotted were the data-points from figures 3a and 3b. The correlation was statistically significant ($p \leq 0.05$, R=0.581, $\nu = 16$, two-tailed). After correcting the durations and CoG values for the overall effect of the prime articulator (see text): R = 0.829 ($p \leq 0.001$, Student $t = 5.12$, $\nu = 13$, two-tailed).

was plotted in figure 4. It was clear that a lot of the remaining, unexplained, variance could be attributed to the overall effects of the prime articulator. After removing these overall effects, the correlation strength could be improved to R = 0.829, i.e., the correlation explained 69% of the variance (Student $t = 5.12$, $\nu=13$, $p \leq 0.001$).

3.3. Manner of Articulation and Voicing

The picture would not be complete without including the effects of manner of articulation and voicing (Figures 5.a and b). Figures 2.a and b already showed that the effects of these factors on duration were fairly independent of position in the word and stress. Figure 5.a shows a quite simple behavior for the duration. All voiced consonants had comparable corrected mean durations (60-70 ms, sum the plosive stop durations with the combined durations of the burst and aspiration). Unvoiced fricatives were about 60 ms longer than voiced ones. Unvoiced plosives were around 30 ms longer than voiced ones, mostly due to aspiration. The lack of an effect of voicing on plosive stop durations could be due to the fact that the other factors (stress, position in the word and prime articulator) already had absorbed any effects. The corrected mean durations of the voiced consonants (both sonorants and obstruents) were fairly independent of the manner of articulation, i.e., between 60 and 70 ms (use total durations of plosives). There was a possibility that the results could have been distorted due to the deletion of plosive bursts+aspirations in certain situations, e.g., in word-final position. However, the overall picture, particularly Figure 3.a, stayed the same when the analysis was repeated with only the plosive stop durations (not shown).

The CoG values give a less uniform picture (see Figure 5.b). Voiced consonants clearly have lower CoG values than unvoiced ones due to the presence of low frequency power from the glottal pulses. Obstruents have higher CoG values than sonorants due to a strong noise component (see also section 2.1).

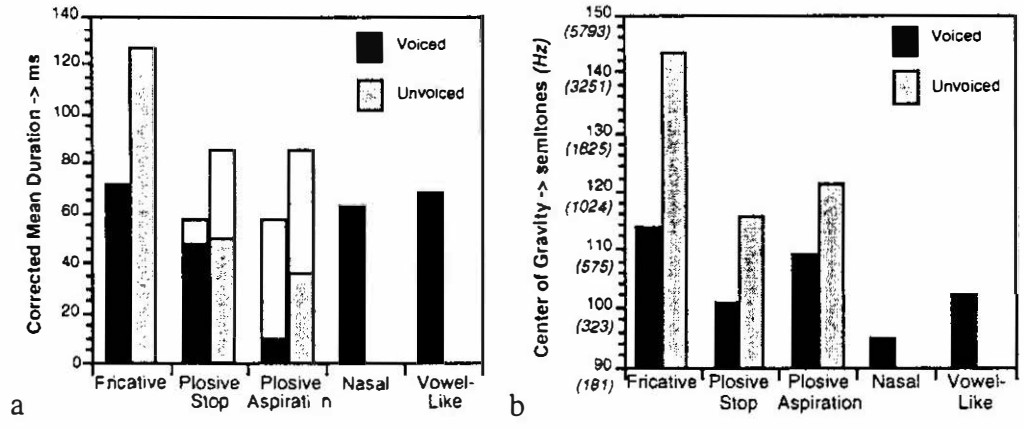

Figure 5 Corrected mean values for consonants split on Manner of articulation. Speech for both speakers combined. Fricatives: /vfðθzsʒʃ/, Plosives: /pbtdkg/, Nasals: /mnŋ/, Vowel-Like: /wlr/.

a. Durations. All differences are statistically significant ($p \leq 0.001$, two tailed WMPSR test), except between nasals and both voiced fricatives and vowel-like consonants and between voiced and unvoiced plosive stops. For the plosives, the open bars indicate the sum of the corrected mean duration of the plosive stop and the plosive burst plus aspiration.

b. Spectral Center of Gravity. All differences are significant ($p \leq 0.001$, two tailed WMPSR test) except those between voiced plosive stops and vowel-like consonants and between voiced fricatives and unvoiced plosive bursts.

3.4. A model of word position effects

There was a strong correlation between the corrected mean durations and corrected mean CoG values (Figure 4). After accounting for the overall effect of the prime articulator, the correlation explained almost 70% of the variance. This high correlation allowed us to combine duration and CoG values in a single description. Removing the overall corrected mean effect of the prime articulator is straightforward for the Labials and Post-Coronals. The word-initial and stressed word-medial Coronals had reasonable values which could be considered "regular". But the unstressed word-medial and word-final Coronals had more or less "irregular" values. Just compensating for the corrected mean value of Coronals would have hidden this difference. By using the correction calculated for the durations of the Post-Coronals for both the Post-Coronals and the Coronals removes most of the offset of the "regular" Coronal durations. The single CoG correction factor can be determined iteratively to maximize the correlation strength (to R = 0.858, i.e., explaining 74% of the variance). The values after correction are plotted in Figures 6.a and b on corresponding scales.

We can now combine the data in Figures 6.a and b in a single model of word-position marking. The corrected mean values are grouped in three tiers according to a kind of "pseudo-stress": stressed, unstressed, and ballistic articulation, i.e., flapped word-final and unstressed word-medial Coronals. For the CoG values we adapted the grouping somewhat. The unstressed word-initial Coronals were indistinguishable from the stressed Coronals and were, therefore, considered to be stressed. The corrected mean CoG values of the word-final Post-Coronals were considered to be outliers and we ignored them in our model.

The six tiers in Figures 6.a and b are fitted with six regression lines: duration and CoG versus position in the word. These regression lines were calculated by recalculating the CoG values to corresponding durations, using the optimized correlation with the corrected mean duration (CoG = 0.3896*duration + 90.86). All six slopes were forced to be equal, effectively calculating a single regression with

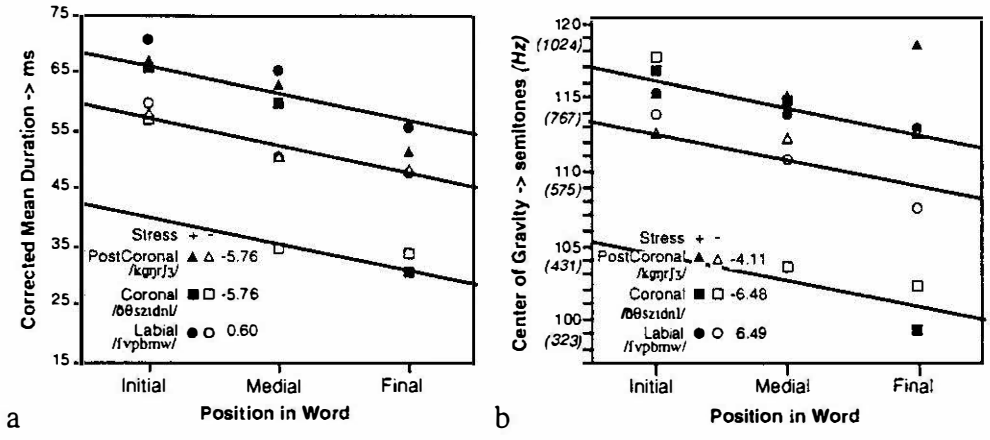

Figure 6. The data of figure 3., but now corrected for the overall effects of the prime articulator (indicated by the numbers inside the plot). Both plots are drawn to the same scale. The lines correspond to the least RMS error fits with forced equal slopes (data of figures a and b pooled, R = –0.702, see text).
a. Duration The overall effect of the Post-Coronals was used to adapt the mean corrected duration of the Coronals (see text).
b. Center of Gravity. The overall effect of the Coronals was estimated by an iterative procedure to maximize the correlation between duration and CoG (to R = 0.858, see text).

reduced degrees of freedom. The resulting regression lines explained 49% of the variance of the six groups combined (R = 0.702, excluding the CoG values of the word-final Post-Coronals) and were plotted in Figures 6.a and b. The regression analysis indicated that the average difference between word positions was 4.64 ms and 1.81 semitones. This meant that in the most common situation, a stressed word-initial consonant preceded by an unstressed word-final consonant, the durational difference over a word boundary is 17.7 ms and the CoG difference 7.1 semitones. However, if the consonants are Coronals, the differences are doubled to 34.4 ms and 15.2 semitones due to the presence of flaps. Note that these values were based on an averaged model in which all other factors are already accounted for.

4. Discussion

In general, the spectral Center of Gravity values (CoG) paralleled the duration data (Figures 3.a and b). For both duration and CoG, word-final consonants had lower corrected mean values than word-initial consonants (except for the CoG of Post-Coronals). Stressed consonants had longer durations and higher CoG values than unstressed consonants. The inherently noisy nature of the CoG values together with the reduced amount of speech available for analysis made the statistical support for the CoG effects less strong than for the durational values. Still, the strong correlation between corrected mean CoG values and durations convinced us that these CoG effects were real too (Figure 4).

Our results showed that the effect of syllable stress on consonant duration and spectral Center of Gravity values depended strongly on the position in the word and consonant identity, i.e., a strong interaction between these factors was found (Figures 1 and 2). Especially, the identity of the prime articulator mattered. Labials, Coronals and Post-Coronals all behaved differently. For all three groups of consonants we saw a difference in duration between word-initial and word final realizations, and less clear differences for CoG values (Figure 3.a and b, respectively). After correcting for the overall effects of articulator, we were able to combine our results into a single

"model" (Figures 6.a and b). This model captured a few of the known interactions (Umeda, 1975; Van Santen and Olive, 1990). It showed that we could localize the largest interaction to the Coronal consonants, which behaved differently from the Labials and Post-Coronals. The difference between word-initial and stressed word-medial realizations of Coronals at the one hand, and word-final and unstressed word-medial realizations at the other hand was extremely large for both duration and CoG. We strongly suspect that this was the result of a shift in the articulation. The former were "fully" articulated, the latter ballistically, i.e., flaps. If we allowed for a third "pseudo-stress" level, flaps, all durational and CoG values behaved "regular", as was shown in Figures 6.a, and b.

The model as presented in Figure 6.b showed some peculiarities. First, it seems that the unstressed word-initial Coronals behaved like stressed ones. With respect to the CoG of Coronals, it seemed as if stress was only apparent as a difference between full and flap articulation. Furthermore, the word-final Post-Coronals had CoG values that were much higher than would be expected from the corresponding durations. We ignored these value in our model, but we have no explanation for this behavior. Finally, the difference between the CoG values of full and flapped Coronals is about 20% larger than the corresponding difference between durations, after rescaling.

Our results support the conclusions of some recent studies on lexical stress and the articulatory emphasizing of prosodic boundaries (De Jong et al., 1992; De Jong, 1995; Fougeron and Keating, 1997; Turk and Sawush, 1997). The articulatory measurements presented in these studies indicated that the emphasizing of syllables and prosodic boundaries was accompanied by an increase in articulatory effort. Our results showed that this increase in effort was also visible as a corresponding shift in the spectral balance of the consonants, which is known to reflect speech effort (Sluyter, 1995 a, b; Sluyter and Van Heuven, 1996; Sluyter et al., 1997; Van Son and Pols, 1996; 1997). This shift, measured as the spectral Center of Gravity, completely paralleled the changes in the duration of the consonants. An analysis of the importance of the *actual* position of the word-medial consonants in the word, e.g., second, third, or later syllable, did not reveal any differences for our speech (not shown). This confirmed the results of Fougeron and Keating (1997) who concluded that there were no position related differences for word-medial consonants. Our results showed clearly that word-final lengthening did not include the word-final consonant. On the contrary, the word-final consonants in our read speech were considerably shortened. This corroborates some of the results of Turk and Sawush (1997).

We can now answer the two questions raised in Fougeron and Keating (1997). The durational and articulatory effects they found with reiterated speech and a single consonant were also present in a comprehensive inventory of consonants from fluent (read) speech and indeed affected the spectral balance of the consonants in the expected direction. Our study also showed that large amounts of speech data were needed to "deconfound" the interacting factors in fluent speech, and even then it was often difficult to get statistically convincing evidence.

5. Conclusions

From a large corpus of read informative sentences from two speakers of American English we were able to model the quantitative effects of stress and position in the word on consonant duration and spectral balance. New statistical methods were used that could "deconfound" the unbalanced data and estimate the "balanced" mean values. It proved that the durational and articulatory emphasizing of stressed syllables

and across word boundaries reported in the literature for reiterant speech and limited inventories of consonants was also found for fluent, read, speech and a comprehensive inventory. Moreover, the articulatory strengthening reported in other studies was reflected in the spectral balance of our consonantal data. A single model with three tiers could explain both our durational and spectral results. A large irregularity in the behavior of Coronal consonants could be explained as a switch between *full* and *ballistic* articulation (flaps) in de-emphasizing circumstances.

6. Acknowledgments

This research was made possible by grant 300-173-029 of the Dutch Organization of Research (NWO). NWO also supplied the additional funding which made this collaboration possible.

7. References

Bahl, L.R., and Jelinek, F.J. (1990). "A maximum likelihood approach to continuous speech recognition", in *Readings in speech recognition,* A. Waibel and K.F. Lee (ed.) Morgan Kaufmann publishers, San Mateo CA., USA, 308-319.

Chennoukh, S., Carré, R., and Lindblom, B. (1997). "Locus equations in the light of articulatory modeling", *Journal of the Acoustical Society of America* **102**, 2380-2389.

Crystal, T.H., and House, A.S. (1988). "Segmental durations in connected-speech signals: Current results", *Journal of the Acoustical Society of America* **83**, 1553-1573.

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* **88**, 101-112.

Cutler, A. and Butterfield, S. (1990). "Durational cues to word boundaries in clear speech", *Speech Communication* **9**, 485-495.

Cutler, A. and Butterfield, S. (1991). "Word boundary cues in clear speech: A supplementary report", *Speech Communication* **10**, 485-495.

Cutler, A. and Carter, D.M. (1987). "The predominance of strong initial syllables in English vocabulary", *Computer Speech and Language* **2**, 133-142.

Cutler, A., Dahan, D., and Van Donselaar, W. (1997). "Prosody in the comprehension of spoken language: A literature review", *Language and Speech* **40**, 141-201.

De Jong, K., Beckman, M.E., and Edwards, J. (1993). "The interplay between prosodic structure and coarticulation", *Language and Speech* **36**, 197-212.

De Jong, K. (1995). "The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation", *Journal of the Acoustical Society of America* **97**, 491-504.

Eefting, W. (1991). "The effect of 'information value' and 'accentuation' on the duration of Dutch words, syllables, and segments", *Journal of the Acoustical Society of America* **89**, 412-424.

Fant, G., and Kruckenberg, A. (1989). "Preliminaries to the study of Swedish prose reading and reading style", *STL-QPSR* **2**, 1-83.

Farnetani, E., and Kori, S. (1986). "Effects of syllable and word structure on segmental durations in spoken Italian", *Speech Communication* **5**, 17-34.

Farnetani, E. (1995). "The spatial and the temporal dimensions of consonant reduction in conversational Italian", *Proceedings of Eurospeech 95*, Madrid, 2255-2258.

Fokes, J., and Bond, Z.S. (1993). "The elusive/illusive syllable", *Phonetica* **50**, 102-123.

Fougeron, C. and Keating, P.A. (1997). "Articulatory strengthening at edges of prosodic domains", *Journal of the Acoustical Society of America* **101**, 3728-3740.

Fourakis, M (1991). "Tempo stress and vowel reduction in American English", *Journal of the Acoustical Society of America* **90**, 1816-1827.

Klatt, D.H. (1973). "Interaction between two factors that influence vowel duration", *Journal of the Acoustical Society of America* **55**, 1102-1104.

Klatt, D.H. (1987). "Review of text-to-speech conversion for English", *Journal of the Acoustical Society of America* **82**, 737-793.

Koopmans-Van Beinum, F.J. (1980). *Vowel contrast reduction, an acoustic and perceptual study of Dutch vowels in various speech conditions*, Ph.D. thesis of the University of Amsterdam, 163 p.

Lindblom, B. (1990). "Explaining phonetic variation: A sketch of the H&H theory", in *Speech production and speech modeling*, edited by W. Hardcastle and A. Marchal (Kluwer, Dordrecht), 403-439.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1988). *Numerical recipes in C*, Cambridge University Press, Cambridge MA, second edition 1992, 632-635.

Rietveld, A.C.M. en Koopmans-van Beinum, F.J. (1987). 'Vowel reduction and stress', *Speech Communication* **6**, 217-229.

Sluijter, A.M.C. (1995a). "Intensity and vocal effort as cues in the perception of stress", *Proceedings of Eurospeech 95*, Madrid, 941-944.

Sluijter, A.M.C. (1995b). *Phonetic correlates of stress and accent*, HIL dissertations **15**, Ph.D. Thesis, University of Leiden.

Sluijter, A.M.C. and Van Heuven, V.J. (1996). "Spectral balance as an acoustic correlate of linguistic stress", *Journal of the Acoustical Society of America* **100**, 2471-2485.

Sluijter, A.M.C., Van Heuven, V.J., and Pacilly, J.J.A. (1997). "Spectral balance as a cue in the perception of linguistic stress", *Journal of the Acoustical Society of America* **101**, 503-513.

Turk, A.E. and Sawush, J.S. (1997). "The domain of accentual lengthening in American English", *Journal of Phonetics* **25**, 25-41.

Umeda, N. (1975). "Vowel duration in English", Journal of the Acoustical Society of America **58**, 434-445.

Van Bergem, D. (1995). *Acoustic and lexical vowel reduction*, in *Studies in Language and Language Use 16*. Ph.D. Thesis, University of Amsterdam.

Van Kuijk, D. (1996). "Does lexical stress or metrical stress better predict word boundaries in Dutch", *Proceedings of the ICSLP'96*, 1585-1588.

Van Santen, J.P.H., and Olive, J.P. (1990). "The analysis of contextual effects on segmental duration", *Computer Speech and Language* **4**, 359-390.

Van Santen, J.P.H. (1992). "Contextual effects on vowel duration", *Speech Communication* **11**, 513-546.

Van Santen, J.P.H. (1993a). "Timing in Text-To-Speech systems", *Proceedings of Eurospeech '93*, 1397-1404.

Van Santen, J.P.H. (1993b). "Statistical package for constructing Text-to-Speech synthesis duration rules: A user's manual", *Bell Labs Technical Memorandum* 930805-10-TM.

Van Santen, J.P.H. (1993c). "Exploring N-way tables with sums-of-products models", *Journal of Mathematical Psychology* **37**, 327-371.

Van Son, R.J.J.H., and Pols, L.C.W. (1990). "Formant frequencies of Dutch vowels in a text, read at normal and fast rate", *Journal of the Acoustical Society of America* **88**, 1683-1693.

Van Son, R.J.J.H., and Pols, L.C.W. (1992). "Formant movements of Dutch vowels in a text, read at normal and fast rate", *Journal of the Acoustical Society of America* **92**, 121-127.

Van Son R.J.J.H. and Pols, L.C.W. (1996). "An acoustic profile of consonant reduction", *Proceedings of ICSLP'96*, Philadelphia, 1529-1532.

Van Son, R.J.J.H. and Pols, L.C.W. (1997). "The correlation between consonant identification and the amount of acoustic consonant reduction", *Proceedings of Eurospeech'97*, Rhodes, .

Van Son, R.J.J.H. and Van Santen, J.P.H. (1997). "Strong interaction between factors influencing consonant duration", *Proceedings Eurospeech'97*, Rhodes, 319-322.

Wang, X. (1997). *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, in *Studies in Language and Language Use* **29** Ph.D. Thesis, University of Amsterdam.

Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M., and Price, P.J. (1992). "Segmental durations in the vicinity of prosodic phrase boundaries", *Journal of the Acoustical Society of America* **91**, 1707-1717.