STRONG INTERACTION BETWEEN FACTORS INFLUENCING CONSONANT DURATION

R.J.J.H. van Son
Institute for Phonetic Sciences, University of Amsterdam, Herengracht 338,
NL-1016CG Amsterdam, The Netherlands, E-mail: rob@fon.let.uva.nl
Jan P.H. van Santen
Bell Labs, Lucent Technologies, Murray Hill NJ, USA, E-Mail: jphvs@research.bell-labs.com

ABSTRACT

Interactions between factors affecting consonant duration are well known. It has proved difficult to quantify these interactions. The difficulty lies in the enormous amount of speech necessary to resolve all factor combinations and their uneven distribution in speech, i.e., factor confounding. Assuming piecewise independence of factor combinations and an additive duration model, it is possible to reconstruct “balanced” mean durations from unbalanced data. Analysis of a corpus of read speech from two speakers allowed us to model the interaction between syllable stress, position in the word, and consonant identity. The strong interactions could be attributed to a “floor” in the shortest durations and irregular behavior of Coronal consonants. The distribution of durations of Coronal consonants is linked to a shift to ballistic articulation, i.e., flaps, in reducing circumstances.

1. INTRODUCTION

Models of segmental duration generally use single factor independence. These models are based on the assumption that the effects of one factor on duration can be modeled without taking into account the values of other factors (cf., discussions in [3],[10],[14],[17]). 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 ([9],[14]). Many other examples are described in the literature (e.g., [2],[4-8],[11-13]). However, it is difficult to model such factor interaction because of a lack of quantitative data ([1],[2]).

The problem is the large amount of speech data needed to resolve interactions between factors. For independent factors, the number of “examples” needed to resolve them scales as the *sum* of their value levels. For interacting factors, the number of examples needed scales as the *product* of the number of factor levels. In practice it is nearly impossible to collect enough speech to cover all possible combinations of factor levels. Especially so because of factor confounding, the fact that some factor values have a low frequency in some contexts [14]. For example, in English, vowels occurring in word-initial syllables are much more likely to be stressed than vowels in word-final syllables; 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 were 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" [14]. This means that we can divide the set of factors into non-overlapping sub-groups, 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.

There are two types of speech corpus. In one, a carefully designed ("balanced") set of sentences is recorded with the property that factor confounding is minimized. 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., [1],[2], [14]). 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 ([14-16]). This technique uses pairwise differences between "Quasi Minimal Pairs" to calculate "Corrected Means" that approximate the hypothetical balanced mean values, i.e., corrected with respect to the unbalanced distribution of realizations [14]. 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. MATERIALS AND METHODS

2.1. Consonant segments

Read aloud sentences of a male and a female speaker of American English were fully labeled and segmented by professional labelers. Both meaningful sentences and phonetically “rich” sentences were used for a total of 1206 sentences for the male speaker and 2951 for the female speaker. Only consonants from accented words were used from the female speaker. Word (or sentence) accent was not indicated reliably for the speech of the male speaker. Therefore, we ignored word accent for this speech and used both accented and unaccented words.

We used all VCV realizations of the 21 consonants /vfðθzsʒʃmnŋbpdtɡkwlrj/. For practical reasons, glottal consonants, affricates and velar fricatives were left out of the analysis. Each plosive was split in a closure and a burst+aspiration part, for a total of 27 “types”.

All intervocalic consonants (VCV, also crossing word boundaries) of non-clitics and non-sentence final words

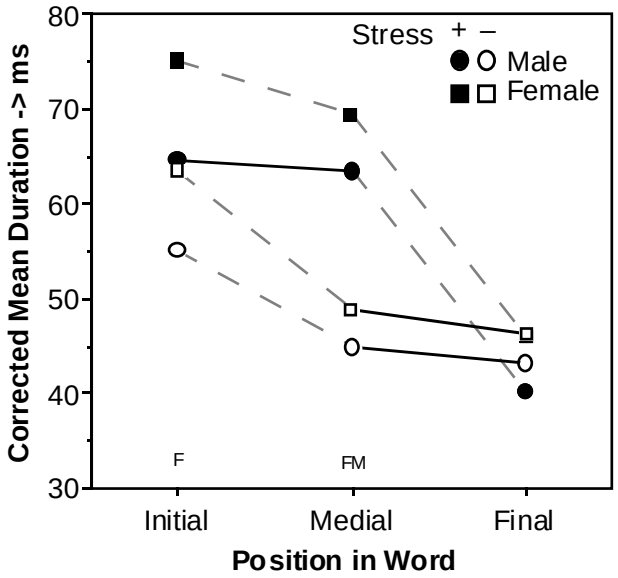

Figure 1. Corrected mean durations of consonants for both speakers. Syllable stress versus position in the word. Significant differences are indicated by dashed lines and 'F, M': $p \leq 0.001$, two tailed WMPSR test between word positions and syllable stress conditions respectively.

were isolated and analyzed. This resulted in 4380 VCV segments for the male speaker and 9606 VCV segments for the female speaker. All speech was recorded with a sampling frequency of 16 kHz and 16 bit resolution. 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).

2.2. Calculating corrected means and statistics

The "Corrected Means" are calculated from the mean values of homogeneous subsets of realizations, i.e., sets for which the values of all five factors are equal. A table is constructed with the factor values for which the average is to be calculated as the row headings and all combinations of values of the other factors as column headings. Each cell contains the mean value of the "homogeneous" set of realizations that conform to the row and column factor values, e.g., there is a cell with the mean duration of all stressed, word-initial /n/ realizations from the male speaker which are followed by a High-F_2 vowel in a three syllable word. For the data in our study, the table contains $27 \cdot 2 \cdot 3 \cdot 4 \cdot 4 = 2592$ cells, 5184 if we pool the values of the two speakers. Less than one-third of these cells contains more than a single realization. Due to this extreme sparsity, standard statistical techniques (e.g., Factor Analysis, ANOVA, or MANOVA) will give results of only limited value (c.f., [1],[2]).

To handle this sparsity, we model 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 [14]. That is, the influence of the row-factors is independent of the influence of the column

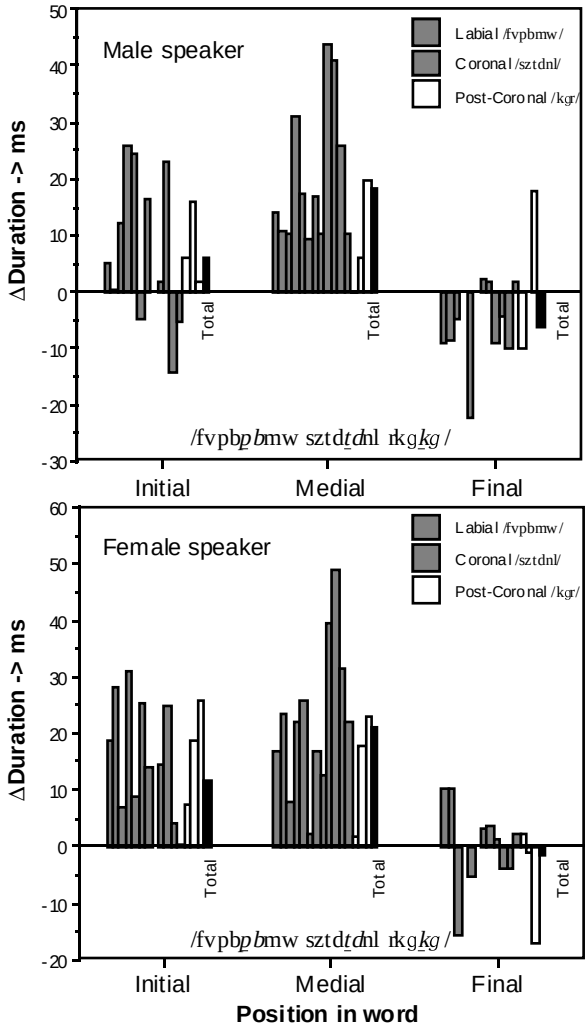

Figure 2. Differences in corrected mean duration between consonants from stressed and unstressed syllables. Unless a consonant did occur in both stressed AND unstressed syllables, no difference was assigned. The order of the consonants is given in the string of phonetic symbols below the graphs. /pbtdkg/: closure durations, /*pbtdkg*/: burst + aspiration durations. Total values include /θðʒʃŋj/ which were left out of the graph.

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.

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. The differences are weighted to account the variation in the number of realizations in each cell, the weight being $w=1/\sqrt{(1/\#Cell_1 + 1/\#Cell_2)}$. However, the exact form of the weighting function has little effect on the outcome, as long as the weights are related to the number of realizations in the cells.

The set of average differences between all pairs of rows constitutes a set of linear equations on the mean row values that can be solved using standard techniques (i.e., minimizing RMS-error with a Singular Value Decomposition, SVD). The results are the Corrected Mean durations of the rows, *relative* to the overall mean duration. For any fully *balanced* set of realizations, the

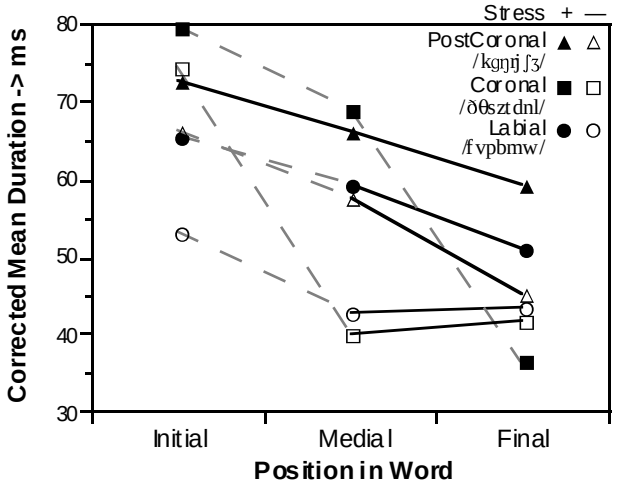

Figure 3. Corrected mean durations of consonants split on Prime Articulator (i.e., Labial, Coronal and Post-Coronal). Syllable stress versus position in the word. Speech for both speakers combined. Dashed lines: p≤0.001, two tailed WMPSR test between word positions.

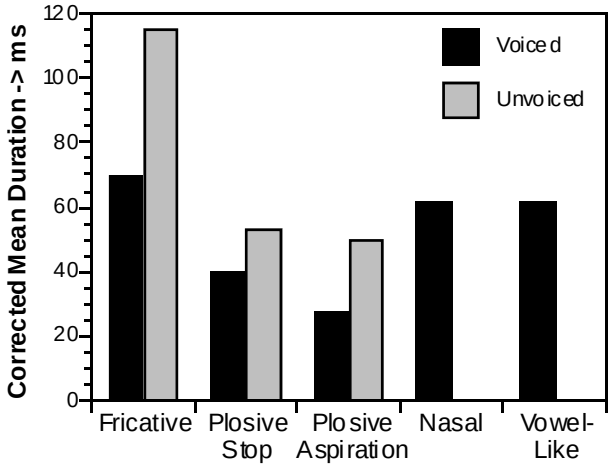

Figure 4. Corrected mean durations of consonants split on Manner of articulation. Speech for both speakers combined. Differences between voiced and unvoiced realizations are all significant (p≤0.001, two tailed WMPSR test). Fricatives: /vfðθzsʒʃ/, Plosives: /pbtdkg/, Nasals: /mnŋ/, Vowel-Like: /wlrj/.

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. The overall mean duration of all realizations from which the corrected means are calculated is used to transform the *relative* durations to *absolute* durations.

The original mean row differences are calculated from pair-wise cell differences. The non-parametric Wilcoxon Matched-Pairs Signed-Ranks test (WMPSR) is used to test the statistical significance of the differences. Each pair of table cells is used as a single matched pair in the analysis, i.e., we do not look "inside" the table cells.

3 RESULTS

3.1. Stress and position in the word

For both speakers we calculated the corrected mean durations 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. The overall corrected mean difference between the speakers amounted to 8.44 ms. For both speakers we see that the stressed word-initial and word-medial consonants have similar durations and both are longer than stressed consonants from a word final position (p≤0.001, two-tailed WMPSR test). For consonants from unstressed syllables we see a different pattern. Unstressed consonants from a medial and final position in the word have similar durations and both differ markedly from unstressed consonants from a word-initial position. Moreover, in word-final position there is no difference in duration between stressed and unstressed consonants (p≥0.001, two-tailed WMPSR test).

For realizations from each position in the word, i.e., word-initial, word-medial, and word-final, we determined the corrected mean difference between stressed and unstressed realizations of each consonant. The values are plotted in figure 2. It can be seen that the behavior found for all consonants pooled is representative of the behavior of the individual consonants. Differences between stressed and unstressed consonants are large in initial and medial position and erratic in final position. The differences in the *size* of the effect of stress on the corrected mean duration for each consonant between initial, final and medial position are all statistically significant (p≤0.002, two-tailed WMPSR test on the values of figure 2, both speakers combined). However, it is also evident that the large influence of syllable stress on consonants in word-medial position can be attributed to the behavior of Coronal consonants, /sztdnl/ (word-medial versus word-initial, p≤0.001, two-tailed, WMPSR test, n=12). Both for Labial and Post-Coronal consonants (i.e., Dorsal, Body, and Root articulation combined), there is no real difference between consonant durations in word-initial and word-medial position (p>0.05, n=16 and n=10). The differences in duration between Coronal and Labial consonants are statistically significant for the word-medial position (p≤0.001, two-tailed WMPSR test on the values in figures 2, both speakers combined, n=16), but not for the word-initial position, (p>0.05, n=12).

3.2. The prime articulator

The differences due to the effect of the primary articulator are investigated by describing each phoneme by three values: Prime articulator (Labial, Coronal, Post-Coronal), Manner of Articulation (Fricative, Plosive stop, Plosive burst+aspiration, Nasal, and Vowel-Like), and voicing (for non-sonorants) and calculating the corrected means. The results for the primary articulator are summarized in figure 3.

There seem to be three “tiers” of duration: Long, Middle, and Short. The duration in each tier reduces from word initial to medial to final position. The differences between these three distinct durational tiers are statistically significant in word-initial and word-medial position. That is, there is a statistically significant difference between at least one member on one tier and one member on another tier in the same word position (p≤0.001, two-tailed WMPSR test) but no more than three tiers are found this way. Very weak evidence for two distinct durations can be found at the word-final

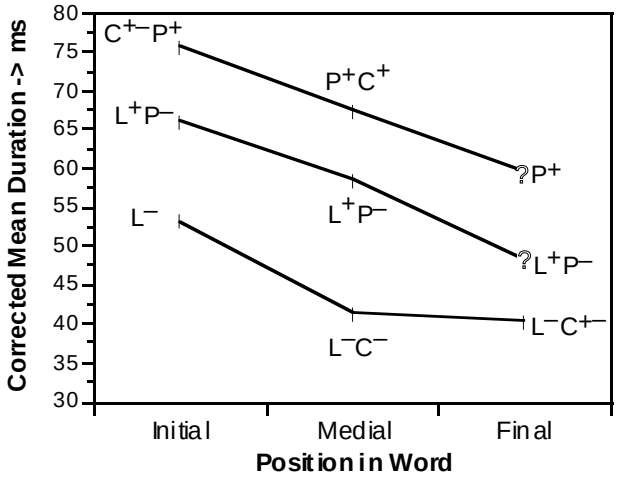

Figure 5. A simplified three tiered model of consonant duration based on figure 3. Plotted is the model duration of consonants as a function of the position in the word. The RMS error, with respect to the data in figure 3, is 2 ms (<4%). L: Labials, C: Coronals, P: Post-Coronals. +: Stressed, –: Unstressed. In the final position, no conclusive evidence for tiers could be found (ʔ-mark, see text).

position (i.e., $p < 0.005$ for only a single pair: Stressed Post-Coronals versus Stressed Coronals, two-tailed WMPSR test). However, all word-final durations might as well collapse into only a single value. This lack of resolution is most likely caused by a lack of data.

For completeness, we included the effect of manner of articulation and voicing which we found to be fairly independent of position in the word and stress. Figure 4 shows a quite simple behavior. All voiced consonants have comparable corrected mean durations (60-70 ms, combine the plosive stop durations and the burst + aspiration durations). Unvoiced consonants are about 40 ms longer than voiced consonants.

4. DISCUSSION AND CONCLUSIONS

Our results show that the effects of syllable stress on consonant duration depend strongly on the position in the word and consonant identity, i.e., Labials, Coronals and Post-Coronals all behave differently. We have summarized this interaction in a simplified model that tries to localize the dependencies (figure 5). This model tries to capture a few of the known interactions ([12],[13]). It consists of three tiers: for Long, Middle, and Short duration. The extrapolation of these tiers into the word-final position is not backed by statistical evidence, it is inferred from the "regular" behavior at other positions in the word and minimizes the dependencies. Interaction between the durational *tiers* and the *position in the word* is limited to a "floor" in the Short tier (at a duration of 40 ms).

Consonants occupy the three tiers according to their prime articulator and syllable stress. The Labial and Post-Coronal consonants behave regularly. The stressed realizations occupy the higher tier, the unstressed realizations the lower (Labials /fvpbmw/ on the lower two tiers, Post-Coronals /kgŋrʃʒj/ on the upper two).

It is the Coronal consonants (/ðθsztdnl/) that behave irregularly. All word-initial and stressed word-medial Coronals occupy the Long tier like the stressed Post-Coronals. The other Coronals occupy the short tier like the unstressed Labials. This can be explained as a shift to *ballistic* articulation. That is., the "reduced" Coronals are uttered ballistically as very short *flaps*.

This strong interaction of factors might explain why the effects of Place-of-Articulation on plosive "hold" durations reported by Crystal and House [1] were so much smaller than ours. They used a representative sample of realizations, thereby "averaging out" most of the effect of the articulator.

We can conclude that it is possible to quantify and localize the interactions between factors affecting segmental durations using a normal, unbalanced speech corpus. It shows that the strongest dependencies exist with regard to word boundaries (word-initial versus final) and discontinuous changes in the articulation of Coronals.

5. REFERENCES

[1] T.H. Crystal and A.S. House. "Segmental durations in connected-speech signals: Current results", *J. Acoust. Soc. Am.* 83, 1553-1573, 1988.
[2] T.H. Crystal and A.S. House. "Articulation rate and the duration of syllables and stress groups in connected speech", *J.Acoust.Soc.Am.* 88, 101-112, 1990.
[3] G. Fant and A. Kruckenberg. "Preliminaries to the study of Swedish prose reading and reading style", STL-QPSR 2, 1-83, 1989.
[4] E. Farnetani and S. Kori. "Effects of syllable and word structure on segmental durations in spoken Italian", *Speech Communication* 5, 17-34, 1986.
[5] J. Fokes and Z.S. Bond. "The elusive/illusive syllable", *Phonetica* 50, 102-123, 1993.
[6] H.S. Gopal and A.K. Syrdal. "Effects of speaking rate on temporal and spectral characteristics of American English vowels", *Sp. Comm. Group. Working Papers* VI, Research Laboratories of Electronics, MIT, 162-180, 1988.
[7] M.S. Harris and N. Umeda. "Effect of speaking mode on temporal factors in speech: vowel duration", *J.Acoust.Soc.Am.* 56, 1016-1018, 1974.
[8] I. Hertrich and H. Ackermann. "Coarticulation in slow speech: Durational and spectral analysis", *Language and Speech* 38, 159-187, 1995.
[9] D.H. Klatt. "Interaction between two factors that influence vowel duration", J.Acoust.Soc.Am. 55, 1102-1104, 1973.
[10] D.H. Klatt. "Review of text-to-speech conversion for English", *J.Acoust.Soc.Am.* 82, 737-793, 1987.
[11] D.K. Oller. "The effect of position in utterance on speech segment duration in English", *J.Acoust.Soc.Am.* 54, 1235-1247, 1973.
[12] N. Umeda. "Vowel duration in English", *J.Acoust.Soc.Am.* 58, 434-445, 1975.
[13] J.P.H., Van Santen and J.P. Olive. "The analysis of contextual effects on segmental duration", *Computer Speech and Language* 4, 359-390, 1990.
[14] J.P.H. Van Santen. "Contextual effects on vowel duration", *Speech Communication* 11, 513-546, 1992.
[15] J.P.H. Van Santen. *Statistical package for constructing Text-to-Speech synthesis duration rules: A user's manual*, Bell Labs Technical Memorandum 930805-10-TM, 1993.
[16] J.P.H. Van Santen. "Exploring N-way tables with sums-of-products models", *Journal of Mathematical Psychology*, 37, 327-371, 1993.
[17] X. Wang. *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, 1997.

This research was made possible by grant 300-173-029 of the Netherlands Research Organization (NWO)