DURATION AND SPECTRAL BALANCE OF INTERVOCALIC CONSONANTS: A CASE FOR EFFICIENT COMMUNICATION [*]

R.J.J.H. van Son
Chair of Phonetic Sciences/ACLC, University of Amsterdam,
Herengracht 338, 1016 CG Amsterdam, the Netherlands,
Tel: +31 205252183/Fax: +31 205252197
Email: R.J.J.H.vanSon@uva.nl

Jan P.H. van Santen[#]
Center for Spoken Language Understanding, Oregon Graduate Institute of Science and technology, Beaverton, Oregon
Email: vansanten@ece.ogi.edu

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

[#] This work was partly performed at Bell Labs, Lucent Technologies, Murray Hill NJ, USA

Abstract

The prosodic structure of speech and the redundancy of words can significantly strengthen or weaken segmental articulation. This paper investigates the acoustic effects of lexical stress, intra-word location, and predictability on sentence internal intervocalic consonants from accented words, using meaningful reading materials from 4157 sentences read by two American English speakers. Consonant duration and spectral balance in such reading materials show reduction in unstressed consonants and in consonants occurring later in the word (Initial vs. Medial vs. Final). Coronal consonants behaved distinctly, which was interpreted as a shift from full to flap or tap articulation in a subset of the phoneme realizations. This shift in articulation, and part of the consonant specific acoustic variation, could be linked to the frequency distribution of consonant classes over the investigated conditions. A higher frequency of occurrence of a consonant class in our corpus and a CELEX word-list was associated with shorter durations and differences in spectral balance that would increase the communicative efficiency of speech.

Keywords: Prosodic Structure; Acoustic Reduction; Redundancy

1. Introduction

The prosodic structure of utterances is reflected in the acoustic realization of the constituent phonemes. Quite a large body of research has shown that syllable stress and pitch accent strongly affect vowel duration and the level of vowel reduction (eg., Fourakis, 1991; Koopmans-Van Beinum, 1980; Lindblom, 1990; Van Bergem, 1995; Van Son and Pols, 1990, 1992; Wang, 1997) and consonant reduction (de Jong et al., 1993; de Jong, 1995; Farnetani, 1995; Van Son and Pols, 1996, 1999a). An archetypal example of reduction is that of vowel realizations becoming more like a schwa in unstressed syllables (eg., Koopmans-Van Beinum, 1980; Lindblom, 1990; Van Bergem, 1995). It has been shown that (lexical) syllable stress also affects the overall spectral balance of vowels (ie., spectral slope, Sluijter, 1995a, b; Sluijter and Van Heuven, 1996; Van Son and Pols, 1999a; cf., spectral tilt in Tabain, 2003). 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; Sluijter, 1995a, b; Sluijter et al., 1997). Furthermore, the presence of edges between prosodic domains such as intonational phrases, words, and syllables, is reflected in the duration and articulation of the surrounding phonetic segments (eg., Byrd, 1993; Byrd and Saltzman, 1998; Cooper, 1991; Tabain, 2003; Turk and Sawusch, 1997; Turk and Shattuck-Hufnagel, 2000; Wightman et al., 1992; see Fougeron and Keating, 1997 for a review).

Prosodic structure can seen as part of the global information structure of speech (eg., Van Son and Pols, 2003b). The information structure of utterances affects the acoustic realizations of segments as a mechanism to increase the communicative efficiency and robustness by emphasizing unpredictable items and de-emphasizing redundant items (eg., Aylett, 1999a, 1999b; Aylett and Turk, 2004; Van Son et al., 2004; Van Son and Pols, 1999b, 2002, 2003a, 2003b and references therein). For instance, within a class of mutually confusable phonemes, eg., the voiced plosives or voiceless fricatives, members with a high frequency of occurrence should be more susceptible to reduction than low frequency members (eg., Boersma, 1998; sections 9.5 and 10.5). This is an extension of the 'redundancy' based framework of acoustic reduction (Aylett, 1999a, 1999b; Aylett and Turk, 2004; Borsky et al., 1998; Lieberman, 1963; Van Son et al., 1998; Van Son and Pols, 1999b, 2003a, 2003b; Van Son et al., 2004; Vitevitch et al., 1997). These studies indicated that the redundancy, or predictability, of individual phonemes is directly related to acoustic reduction in a way that increases communicative efficiency, ie., less speech effort is used to produce redundant items. Prosodic structure already emphasizes unpredictable elements and de-emphasizes predictable ones to help comprehension (eg., Aylett and Turk, 2004). For instance, the most complex syllables in a word tend to be stressed (eg., Cutler and Carter, 1987), and rare and new words get prominence (eg., Aylett and Turk, 2004; Cutler et al., 1997). Furthermore, the boundaries (beginning and end) of constituents have different importance for comprehension. For instance, the word onset is more important to recognition than word offset (Cutler and Carter, 1987; Cutler et al., 1997; cf., Van Son and Pols, 2003a, 2003b).

This leads to two roads to achieving communicative efficiency from reducing redundant phonemes. The first is to distribute stress, prominence, and boundaries in such a way as to reflect the redundancy of the items, ie., the prosodic structure reflects the information structure (eg., Aylett and Turk, 2004; Van Son and Pols 2003b). The second is to reduce and emphasize the articulation of items according to their individual redundancy, in addition to the effects of the prosodic structure. The possibility of a direct effect of redundancy on the realization of individual phonemes was excluded from the 'Smooth signal redundancy hypothesis' of Aylett and Turk (2004) who concluded that prosodic structure might account for all redundancy effects. However, Aylett and Turk (2004) worked with syllable durations instead of phoneme realizations. These redundancy strategies are not mutually exclusive and can, in principle, both be used at the same time.

The challenge is to disentangle the effects of the prosodic structure and redundancy from each other and all the other factors that affect the acoustic realization of individual phonemes. To do this, it is necessary to build a quantitative model of the effects of prosodic structure after accounting for other structural factors (eg., speaker identity) and then study whether the predictability of the individual phonemes can explain variation in acoustic reduction that remains after applying the model.

In this paper we focus on two specific prosodic factors and their interactions: realized (lexical) syllable stress and intra-(lexical)-word location. This choice was partly motivated by practical considerations. In the prosodic hierarchy, lexical stress is the lowest level of emphasis and word boundaries, or intra-word locations, are the first prosodic boundary level above the syllable for consonants (Nespor and Vogel, 1986; cf., Byrd and Saltzman, 1998; de Jong et al., 1993; Fougeron and Keating, 1997; cf., references in Turk and Shattuck-Hufnagel, 2000). As a consequence, word boundaries are also more numerous than higher level boundaries and more data can be gathered for them from a given body of speech. 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; cf., 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 a content word and one or more monosyllabic function words and producing them as single, prosodic, words[1] (eg., Wightman et al, 1992). Ideally, this study should also have been done using the boundaries of the prosodic words instead of the lexical words. However, the difficulty of determining these prosodic word boundaries in large amounts of running speech precluded this approach.

We use two techniques to study the effects of the prosodic, and positional factors mentioned above on the acoustic realization of intervocalic consonants from read sentences. First, the problems of interactions between factors and confounding of factors in natural reading materials are addressed using statistical methods described elsewhere (a special case of the general linear model: *quasi-minimal pairs analysis*, and *corrected means*; Van Santen, 1992; see also section 2.3). Second, we needed an acoustic measure that can be used to link articulatory 'strength' (articulatory emphasis as used eg., by Fougeron and Keating, 1997) to the acoustics of speech over comprehensive phoneme inventories (cf., Chennoukh et al., 1997 for an alternative modeling effort; see also Jongman et al., 2000). The latter problem can be pragmatically solved by using the spectral slope as studied by Sluijter and others (Sluijter 1995 a, 1995b; Sluijter and Van Heuven, 1996; Sluijter et al. 1997; Hanson, 1997; Hanson and Chuang, 1999; Tabain, 2003), or the associated spectral balance, which is an indicator of the acoustic power relative to a normal speaking level (Boersma, 1998, section 4.3; cf., Hanson, 1997; Hanson and Chuang, 1999). Within any phone class, a flatter slope generally indicates perceptually louder speech (eg., Sluijter 1995 a, b; Sluijter and Van Heuven, 1996; Sluijter et al. 1997) and may indicate more effort on the part of the speaker, eg., higher lung pressure and muscle tension (see section 2.2). Note that it is *not* possible to define a universal notion of "speech effort" unambiguously (Boersma, 1998, section 4.3; Pouplier, 2003; see Discussion).

2. Material and methods

2.1 Consonant segments

Phonemically labeled and segmented speech of a professional male and female speaker of American English was used. All speech was recorded with a sampling frequency of 16 kHz and 16 bit resolution. The recordings were (isolated) meaningful read sentences from several corpora, eg., from newspaper articles and train table announcements ("The train leaves at 2 pm"). Segmentation was done by professional labelers (Van Santen, 1992; and references therein). Labels were computed from text with the pronunciation component of the Bell Labs text-to-speech system (Sproat, 1998; cf. references in Van Santen, 1992), and manually adjusted to reflect dialectic and allophonic variations (ie., pronunciation variants). The intervocalic consonants were segmented manually aided by the *waves* program, which displays the speech wave, spectrogram, and other acoustic representations of an utterance. The following conventions were used for the determination of consonant-vowel boundaries: In general, vowel onset was determined by the first zero crossing at which the formant structure characteristic for the vowel was visible; the consonantal aspiration, if present, was not included in the vowel. Vowel offset was determined similarly. Transitions from vowels to or from approximants required special attention. These transitions could typically be detected by a visible discontinuity in the spectrogram. The visible discontinuity would be in the form of a relatively sudden amplitude change in a broad frequency band. If no reliable discontinuity was found, the midpoint of the transition region was used; and when no transition region was detectable, fixed formant values were used (for example, for our male speaker, an F_2-value of 900 Hz for /w/, and an F3-value of 1750 for /r/). These values are determined by trial and error for each speaker to give a good separation of phones. The reliability of the measurements was checked using vowel durations, thirty-eight utterances were segmented independently by four phoneticians (the same as the labelers), generating 320 sets of four vowel durations. For each of these sets, *the median absolute devia*tion from *the me*dian (MAD) was computed to indicate segmentation reliability. Across the 320 groups, the MAD varied between 0.1 and 38 ms, with a median of 3 ms. Ninety percent of the MAD values were less than 10 ms, and seventy percent less than 5 ms (Van Santen, 1992).

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-function words and non-sentence final words were analyzed. This resulted in 4116 VCV segments for the male speaker and 8957 VCV segments for the female speaker[2], of which 1430 and 3464, respectively, were plosives. 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). To compensate for this reduced number of realizations, we also used CoG values (see section 2.2) of inter-vocalic consonants from sentence-final words of the female speaker (but *not* sentence-final consonants, of course). The spectral effects of sentence ends are generally limited to the last rhyme, which was excluded (Nord, 1987; see discussion of this point in Van Son and Pols, 1990, 1992).

For plosives, the start of the burst had also been indicated. Labels included the presence of syllable stress as indicated by the labelers ('realized' lexical stress, for non-clitics) and whether or not a word could be cliticized (eg., *am* in *I'm* or *not* in *don't*). Phrasal accent was also marked for the speech of the female speaker. Only consonants from accented words were used for her speech. Pitch 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 pitch accent for his speech and used all words that could carry a pitch accent (which excluded all words that could be cliticized). With respect to the acoustics of speech segments, lexical stress and pitch accent are two separate, but not necessarily *independent*, phenomena (Sluijter and Van Heuven, 1996; Sluijter, 1995b; Turk and Sawusch, 1997). Therefore, the inclusion of unaccented words from the male speaker could have influenced the relation between lexical stress and the other factors in our corrected means analysis. However, the averaged behavior of the male speaker's consonant realizations was, in all respects, identical to those of the female speaker both qualitatively and quantitatively. Therefore, we will not elaborate on the difference in word selection between the material of the speakers. However, we decided to keep the best available data and did not include possibly unaccented words from the female speaker. We also ignored all consonants from the last word of each sentence for duration measurements because these words are known to behave differently in many respects (Van Santen, 1992). Note that no sentence initial or final consonants were used, as isolated sentences do not contain intervocalic initial or final consonants. Also, no consonants around pauses inside the utterances were used. As the speech consists of read, short isolated sentences with standard punctuation, excluding pauses would exclude most, if not all, phrase boundaries.

For practical reasons, we excluded the glottal consonants, affricates, and /j/ from our analysis, 916 realizations in total. The former groups, glottal consonants (/h/, glottal stop) and affricates, were dropped because of the inherent ambiguities in the determination of their segmentation and the interpretation of their composite structure. The realizations of the latter group, /j/, were dropped because unequivocal identification of realizations was hampered by a labeling ambiguity[3]. We used all VCV realizations of the 20 remaining consonants /v f ð θ s ʒ ʃ m n ŋ b p d t g k w l r/. The duration and spectral features of the burst+aspiration part of a plosive might not react in the same way to stress and position in the word as the duration and spectral features of the stop-closures. Therefore, both parts of a plosive were treated as separate segments. That is, each plosive was split into an independent closure and a burst+aspiration part (burst+aspiration are indicated by italic, underlined IPA symbols, eg., /p/ for the stop-closure and /p/_for the burst+aspiration). Therefore, in total we used 26 consonant types.

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 frequency: separating vowels with High, Middle, and Low F_2 frequency, and Diphthongs). Vowel height was excluded as it interacted perfectly with frontedness and only one factor could be used at the same time. The last two, word length and vowel frontedness, proved to have minimal or no effects on consonant duration or on CoG in our corpus. We still kept them to define more homogeneous subsamples (see section 2.3). The stress value of a consonant was defined as follows. For word-initial and word-medial consonants it was defined as the stress of the following vowel, whereas for word-final consonants stress was defined as the stress of the preceding vowel. In a preliminary analysis, we found this maximum-onset stress-assignment rule to give consistent results regarding the effects of stress on both duration and CoG values.[4]

2.2 The Spectral Center of Gravity (CoG)

Phoneme duration is correlated to reduction and emphasis, but does not describe articulation. In the context of this study, we would like to be able to evaluate the "effort" invested in the articulation of the consonants more directly. For vowels it is well known that formant values correlate quite strongly with reduction, emphasis, and speaking effort in general, because they are linked to the relative positions and movements of the articulators themselves (for references, see eg., Koopmans-van Beinum, 1980; Van Bergem, 1995; Van Son and Pols, 1990, 1992, 1999a). For consonants, no such generally accepted spectral measures exist. However, earlier work by Sluijter and colleges showed that the *spectral slope* of a speech segment can be related to its stress and loudness (Sluijter, 1995a, b; Sluijter and Van Heuven, 1996; Sluijter et al., 1997). The relation to emphasis and loudness in both production and perception makes the spectral slope a candidate for a measure of effort.

The *spectral slope* of a speech segment is determined by the underlying articulation and by the syllable stress and is itself a perceptual cue for syllable stress (Sluijter, 1995a, b; Sluijter and Van Heuven, 1996; Sluijter et al., 1997; cf., Hanson, 1997; Hanson and Chuang, 1999; Tabain, 2003). 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. Sluijter and her colleagues used spectral slope in stressed, unaccented vowels. An evaluation of her work, and the remarks on it in Boersma, (1998, section 4.3), convinced us that the spectral slope measure she used for vowels was actually equivalent to the spectral balance (first spectral moment), which is a rough indicator of both the relative produced power and the perceived loudness.

The spectral balance can be described quite well by its "balancing" point: the Center of Gravity. For the CoG it is known how it reacts to relevant acoustic parameters, like air-flow and turbulence, which are linked to "articulation strength" (cf., Boersma, 1998 section 4.3; cf., Hanson, 1997; Hanson and Chuang, 1999). 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 'surrogate' 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 the time of closure for sonorants (actually, the steepness of the air-flow profile at closure, see Hanson, 1997; Hanson and Chuang, 1999 for a detailed articulatory modeling of the spectral slope of the glottal source). That is, to increase the CoG of a fricative, ie., turbulence caused by a small opening, a speaker must either increase the lung pressure to increase the air-speed or decrease the size of the opening against the air pressure. To increase the CoG of a sonorant, the speaker must increase the closing speed of the vocal folds by increasing the amplitude of the vocal fold movements at a fixed frequency. In both cases, an increase in the center of gravity of the spectrum of a specific phoneme requires the speaker to exert more effort in terms of muscle tension or lung pressure (cf., Boersma, 1998, sections 4.3 and 7.1).

For a large inventory of consonants, the spectral balance is a much more practical measure than the spectral slope. On the perceptual level, the CoG is, rather indirectly, related to the perceived loudness and the effective spectral bandwidth of speech sounds. A lower CoG means less high frequency power and therefore, a smaller effective frequency range for recognition (Van Son and Pols, 1999a). The CoG is also defined for almost all phonemes *and* correlates well with consonant reduction, ie., reduction predictably changes the CoG (Van Son and Pols, 1999a). All this induced us to choose the CoG to represent spectral balance.[5]

The Center of Gravity of a spectrum (CoG) is, in a sense, the 'mean' frequency. It is calculated as the first spectral moment by dividing $\int f \cdot E(f) \cdot df$ by $\int E(f) \cdot df$ in which $E(f)$ is the power spectrum and f the frequency. Used in this way, the CoG correlates with both consonant reduction due to stress and speaking style (Van Son and Pols, 1996, 1999a, cf. Jongman et al., 2000) 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 labelers (ie., away from the vowel boundaries). 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 window length of 25 ms was chosen as a trade off between spectral and time resolution. It is of the same order as the temporal integration window of the human ear (often set to 20 ms, see O'Shaughnessy, 1987, p159).

The CoG frequencies vary widely, from around 200 Hz for nasals to over 5000 Hz for labial fricatives. As a consequence the variance of the CoG frequencies for consonants is of the same order as the average (see Jongman et al., 2000). This would severely distort any comparison between consonants because the variance of the larger values (eg., fricatives) would completely dominate that of the smaller values (eg., nasals). Therefore, we decided to express the CoG frequencies in semitones with respect to 1 Hz, ie., $12 \cdot \mathrm{Log}_2(\mathrm{CoG})$, which equalizes the variances over the range.

The CoG is a spectral measure and has to be calculated over a time-window. This can cause problems if a consonant is shorter than the effective time-window (here 25 ms). In our study, the effects of realizations shorter than the effective time-window are limited. Only 69 realizations of non-plosive consonants were shorter than 25 ms (0.9%). However, the problem was marked for the plosives, 3027 out of 4854 bursts (62%) and 1040 out of 4924 closures (21%) were shorter than 25 ms. Closures shorter than 25 ms were almost completely confined to realizations of /t d/ (974 for /t d/ versus 66 for the others, ie., 94% vs. 6%). For bursts, the distribution was more even (1674 were from /t d/ and 1353 from other plosives, ie., 55% vs. 45%). However, given the fact that bursts are preceded by a closure, and given the placement of the window (centered at 5 ms into the burst), we should use 15 ms as the relevant cutoff duration for the CoG window on bursts. This translates into 2038 bursts shorter than 15 ms (42%), of which 1252 are from /t d/ (61%). This leads to the conclusion that our 25 ms window size catches more than 99% of all non-plosive consonant realizations. It catches almost 80% of all plosive closures. The special position of the bursts (preceded by 'silence') gives a minimal burst-size of ~15 ms. Our CoG measurements are 'relatively safe' for almost 60% of the burst realizations. What is a problem is that the very short plosive fragments are concentrated in the /t d/ realizations in a systematic way.

The CoG of plosive bursts is determined by the pressure buildup before release, the (effective) area of the release opening, and the filtering of the speech tract. As such, there is not much difference with the fricatives. Due to their short duration, the power spectrum will sometimes sample part of the plosive closure and part of the aspiration and following vowel (the latter two have similar spectra). If the acoustic energy in the burst is small, the overlap with the vowel sound can reduce the CoG markedly. Van Son and Pols (1999a) indeed found that CoG frequencies of plosive bursts differed between realizations from stressed and unstressed syllables (however, they did not find such a difference for speaking style, note that Dutch plosives are unaspirated).

The ideal voiceless plosive closure has *no* acoustic power at all, only quantization noise (the white noise resulting from digitalization) and other noise originating in the recording equipment. In this ideal case of recording noise only, the power spectrum will be flat (white) and the CoG measured will be half the Nyquist frequency by definition, ie., 4 kHz (~ 144 semitones). But in real speech, there will almost always be some speech noise inside the stop or overlap with part of the preceding vowel and this will cause the non-silent *speech* spectrum to determine the CoG.

2.3. 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 and Olive, 1990; 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, 1993c).

```
###################################################################
#                                                                 #
#                      PLACE TABLE 1 HERE                         #
#                                                                 #
###################################################################
```

The corrected means analysis model is a special case of the general linear model. Corrected means analysis eliminates a higher percentage of the data than standard methods, eg., ANOVA or MANOVA, but is more robust (cf., Dodge, 1981). In corrected means analysis, we use the *incidence matrix* as the starting point (see Table 1). The rows of the incidence matrix contain all the combinations of factor levels we are interested in, eg., *lexical stress* and *position in the word* in Table 1. All other, nuisance, factor levels that affect the measure of interest but are outside the scope of (and a nuisance to) the current analysis, eg., length in syllables and speaker identity in Table 1, are collected in the columns. From this incidence matrix, we obtain for each pair of rows (ie., pairs of combinations of levels on the factors of interest) a list of cell pairs such that each cell pair is obtained from the same column (ie., it is a quasi-minimal pair differing only in the row factor values but not in the nuisance factors) and is not empty. For each such list, we can compute the mean difference between the means in the pairs. From this, we can construct a square matrix whose order is equal to the number of rows in the incidence matrix. This matrix contains the means of the within-quasi-minimal pair differences. The square means matrix can then be fitted with a linear model using standard least-squares methods.

To calculate the means of the within-quasi-minimal pair differences, ie., the mean differences between rows, we needed some way to account for the varying sample sizes that underlay each table cell mean. These varying sample sizes determined the variances of the within-quasi-minimal pair differences, ie., table cell differences. The weighting of each difference should reflect that differences based on smaller samples had a higher variance, ie., error. Under the assumptions of equal variance (σ^2) for the individual measurements, the variance of the differences between two table cell means ($C_{i,k}$, $C_{j,k}$) scales as

$$var(C_{i,k} - C_{j,k}) = \sigma^2 \cdot (1/N_{i,k} + 1/N_{j,k})$$

With $N_{i,k}$ and $N_{j,k}$ the sample sizes in cells *(i,k)* and *(j,k)*. To calculate the mean difference between rows (i-j) from the individual cell differences ($C_{i,k}$ - $C_{j,k}$) we choose as the weighting factor of the difference ($w_{i-j,k}$):

$$w_{i-j,k} = 1/\sqrt{(1/ N_{i,k} + 1/N_{j,k})}$$

which corresponded to the scaling of the reciprocal of the standard error due to the sample sizes ($N_{i,k}$,$N_{j,k,}$ are the number of samples in each cell). It must be noted that the choice of weighting factors had only a small effect on the corrected mean values, as long as larger samples had larger weights.

The reason that this method is more robust than conventional methods (eg., Dodge's R-method, Dodge, 1981) is that it only uses columns where at least two cells are filled. If the combination of nuisance factor levels corresponding to this column produces an unusually large duration, this does not affect the difference score associated with the row pairs that do have filled cells in that column because both values in the pair are unusually large, nor does it affect the difference scores of row pairs that do not have filled cells in that column because the column plays no role in the overall difference score of this pair. By contrast, in the standard method, many columns where only one cell is filled are used in the estimation process, and many of these cells have large standard errors because of data sparsity, which then creates unreliable estimates for the corresponding row parameters.

The results are the relative *Corrected Means* of the rows, eg., 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 and CoG values in an additive manner. However, if all durations and CoG values are replaced by their logarithm, the resulting model will be multiplicative. No further changes are necessary to cover a multiplicative model. Preliminary analysis showed that the results for a multiplicative model were more extreme with larger differences between factors than those for the additive model so we decided to report results from the more conservative additive model.

2.4. 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. Distribution-free tests are generally considered to be less sensitive than tests based on the Normal distribution like the Student-t test. However, consonant durations and CoG values are not normally distributed and we wanted 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.

We performed our analysis in the form of planned comparisons, using the conservative significance level of $p \leq 0.002$ as our cut-off to account for repeated tests when appropriate (Bonferroni correction).

3. Results

The presentation of the results is divided into 5 sections. In sections 3.1-3.3 the basic results are presented regarding the relation between duration and CoG on one hand and speaker, stress, position in the word, and consonant identity on the other hand. In section 3.4 these results are integrated into a single quantitative model of relative acoustic reduction as a function of stress, position in the word, and consonant identity. In section 3.5 important aspects of the model of section 3.4 are linked to segmental redundancy in terms of frequency of occurrence.

3.1. Syllable stress and position in the word

For each speaker 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 were plotted in Figure 1.a with absolute offsets (ie., as differences from the overall mean values). The results for the spectral Center of Gravity were plotted in Figure 1.b. The overall corrected mean difference in duration and CoG 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). These corrected mean differences are used to normalize the measurements before pooling the results of both speakers.

```
###################################################################
#                                                                 #
#                    PLACE FIGURE 1.a & b HERE                    #
#                                                                 #
###################################################################
```

For both speakers we saw that the stressed word-initial and word-medial consonants had similar durations, the difference was 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 saw the opposite. Unstressed consonants from a medial and final position in the word had similar durations (2 ms difference) and both differed markedly from unstressed consonants from a word-initial position, which were around 13 ms longer ($p \leq 0.001$, two-tailed WMPSR test). Moreover, in word-final position there was no difference in duration between stressed and unstressed consonants (1 ms difference, $p > 0.002$, two-tailed WMPSR test).

The CoG data largely mirrored the duration data (Figure 1.b). However, due to smaller samples and a higher variance, we could not ascertain that all the differences were statistically significant. The differences between stressed word-*initial* and word-*final* consonants was 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 was 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 were statistically significant ($p \leq 0.001$, two-tailed WMPSR test). The corrected mean difference between word-initial and word-final consonants was around 10 semitones for both stressed and unstressed consonants (both speakers pooled).

3.2. Syllable stress and primary articulator

To investigate the influence of consonant identity on the corrected mean values we focused on the corrected mean durations (see section 2.3). For each position in the word, ie., 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 for individual phonemes were plotted in Figures 2.a and 2.b for both speakers separately. It was clear that the behavior of both speakers was quite similar. For analysis, we divided the consonants into three articulatory defined classes: Labials, Coronals, and Post-Coronals. 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 primary articulator, but not included in our restricted definition of Coronals, were pooled as Post-Coronals. There were too few realizations of phonemes with other primary articulators than Coronal or Labial to allow us to distinguish them in our analysis. However, the phonemes grouped here as Post-Coronals have in common that they all involve the heavier (slower) parts of the tongue in articulation. That was also the reason to include the retroflex American /r/ with the Post-Coronals instead of with the Coronals[7].

```
##################################################################
#                                                                #
#                  PLACE FIGURE 2.a&b HERE                       #
#                                                                #
##################################################################
```

In Figures 2.a and 2.b 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 pairwise differences in the size of the effect of stress on duration, as displayed in Figures 2.a and 2.b, between word-initial, -medial, and -final position were all statistically significant (p ≤ 0.001, two-tailed WMPSR test on the numbers in Figures 2.a and 2.b, 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*td*nl/[6] (word-medial versus word-initial, p ≤ 0.001, two-tailed, WMPSR test on the numbers in Figure 2, n=14).

Both for Labial and Post-Coronal type consonants in Figure 2, /fvpb*pb*mw/ and /kg*kg*r/, respectively, there was no significant difference between consonant durations in Word-Initial and Word-Medial position (p > 0.05, n=16 and n=10, respectively). The sizes of the stress related differences in duration for word-medial Coronal consonants were larger than those for Labial consonants. This could be shown when the bar-sizes of consonants with identical manner of articulation but different articulators were compared in Figures 2.a and b (eg., /p/ versus /t/, /m/ versus /n/, /f/ versus /s/, etc.). The stress related differences between pairs with identical manner but different articulators were statistically significant for the Word-Medial position (p ≤ 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=14).

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: Primary articulator (Labial, Coronal, Post-Coronal), Manner of Articulation (Fricative, Plosive Closure, 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.

```
###########################################################################
#                                                                         #
#                       PLACE FIGURE 3.a&b HERE                           #
#                                                                         #
###########################################################################
```

From the corrected mean values, it was 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 quite differently in other situations (unstressed word-medial and all word-final Coronals). This different behavior was concentrated in the stops and nasals (ie., /tdn/). We suspect that these very short realizations of the Coronals can be interpreted as the result of a switch to reduced allophones of /t/, /d/, and here /n/, eg., flaps or taps (cf., Clark and Yallop, 1990; O'Shaughnessy, 1987). Note that our speech data have been transcribed with the full phoneme labels, ie., without marking flaps. Also, the start of a release was indicated in the label file for every plosive (including inaudible release bursts).

When inspected in detail, the picture became more complicated for the corrected mean CoG values (Figure 3.b). Stressed consonant realizations generally had higher CoG frequencies than comparable unstressed realizations but this was only statistically significant for word-medial 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 for CoG values, 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).

```
####################################################################
#                                                                  #
#                      PLACE FIGURE 4 HERE                         #
#                                                                  #
####################################################################
```

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 duration and CoG values *could* be determined. This correlation was plotted in Figure 4 ($R = 0.581$, $p \leq 0.05$). Clearly, a significant portion of the remaining, unexplained, variance could be attributed to the overall effects of the primary articulator. After removing the overall effects of primary articulator, the correlation strength could be improved to $R = 0.829$ ($p \leq 0.001$, Figure 4), ie., the correlation explained 69% of the variance (Testing significance of R with Student $t = 5.12$, $\nu=13$, $p \leq 0.001$; note the reduction in degrees of freedom ν). This strong correlation illustrates that both duration and CoG can be linked to reduction.

3.3. Manner of Articulation and Voicing

```
####################################################################
#                                                                  #
#                    PLACE FIGURE 5.a&b HERE                       #
#                                                                  #
####################################################################
```

The picture would not be complete without including the effects of manner of articulation and voicing on corrected mean duration and CoG (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 that the effects on duration can be concisely summarized as follows. 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 closure durations could be due to the fact that the other factors (stress, position in the word and primary articulator) had already absorbed any effects. The corrected mean durations of the voiced consonants (both sonorants and obstruents) were fairly independent of the manner of articulation, ie., between 60 and 70 ms (the total duration of the plosives). There was a possibility that the results could have been distorted due to unreleased plosive bursts in certain situations, eg., in word-final position. However, the overall picture stayed the same when the analysis was repeated with the plosive burst+aspiration durations removed (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).

3.4. A combined picture of the effects of stress, position, and primary articulator

The results presented in the previous sections are rather intricate. There are at least four main factors that influence the duration and spectral balance of our consonants: Stress, Position in the Word, Primary Articulator, and a systematic shift in Allophone used. The results can be understood better if the effects of these factors are combined into a single quantitative model of consonant duration and CoG.

The complicated picture shown in Figures 3.a and 3.b can be simplified by first removing the global effects of the primary articulator. The points in Figures 3.a and 3.b are shifted vertically by different amounts corresponding to the three primary articulators and the two independent variables stress and position in the word. We do this as follows. First, we note that there was a strong correlation between the corrected mean durations and corrected mean CoG values (R = 0.829 after correction, Figure 4). This correlation is comparable to that between vowel duration and formant reduction (cf., Lindblom, 1990; Van Son and Pols, 1990, 1992, 1999a). After accounting for the overall effect of the primary articulator, ie., modeling the effect of the primary articulator as an overall shift in the corrected mean duration or CoG, 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 primary articulator is straightforward for the Labials and Post-Coronals. The Coronals in word-initial and stressed word-medial position also had reasonable values which could be considered 'regular' in the sense that they followed the pattern of the Post-Coronals. But the unstressed word-medial and word-final Coronals had more or less 'irregular' values that are different from the other consonant classes (ie., they were replaced by flaps/taps). Just compensating for the corrected mean value of all Coronals pooled would have hidden this difference due to position in the word and stress. Using the correction in duration 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 for the Coronals could be determined iteratively to maximize the correlation strength between duration and CoG frequency (to R = 0.858, ie., explaining 74% of the variance). The values after correction are plotted in Figures 6.a and b on corresponding scales.

```
####################################################################
#                                                                  #
#                      PLACE FIGURE 6.a&b HERE                     #
#                                                                  #
####################################################################
```

We can now combine the data in Figures 6.a and b in a single (descriptive) picture which abstracts the effects of stress and position in the word from consonant identity. The corrected mean values are grouped in three tiers according to our acoustic estimate of 'strength': strong (stressed), weak (unstressed), and extra-weak, ie., flapped/tapped or reduced allophone (word-final and unstressed word-medial Coronals). The third, extra-weak tier was added to account for the strong reduction found in 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 strong realizations. The corrected mean CoG values of the word-final Post-Coronals were considered to be outliers and we ignored them in our picture.

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 transforming the CoG values to the corresponding durations, using the optimized correlation with the corrected mean duration according to the formula: CoG = 0.3896•duration [ms] + 90.86 [semitones]. All six slopes were forced to be equal, effectively calculating a single regression for *position in the word* for all conditions included with 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). The regression analysis indicated that the average difference between word positions was 4.64 ms and 1.81 semitones and between strong and weak realizations (+/- stress) 8.6 ms and 3.5 semitones. This means that if we compare a stressed word-initial consonant with an unstressed word-final consonant, the durational difference is 17.7 ms and the CoG difference 7.1 semitones. However, if the consonants are Coronals, or more precisely, Coronal stops or nasals, the differences are doubled to 34.4 ms and 15.2 semitones due to the presence of flaps or taps (see Figures 6.a and 6.b). Note that these values are based on an idealized averaged model in which all other factors are already accounted for. Tabain (2003) did not find an effect of boundary strength in French consonants. This contrasts with our results for word boundaries. However, Tabain's (2003) study looked only at word initial obstruents after prosodic boundaries of varying strength. Our consonants vary between word-initial to word-final, with no obvious other prosodic boundaries preceding or following them. This might explain the differences between her results and ours.

```
####################################################################
#                                                                  #
#                      PLACE FIGURE 7.a&b HERE                     #
#                                                                  #
####################################################################
```

The introduction of a third, 'extra-weak' or reduced, tier in this model (Figure 6.a and b) to account for the different behavior of the Coronals needs more justification than just a better fit. If there is indeed a large scale shift in articulation between 'strong' (word-initial and stressed word-medial) and 'extra-weak' (unstressed word-medial and word-final) Coronal realizations, this should be visible in the distribution of durations (mostly /t d n/). As an illustration, Figure 7 compares the durational histograms of realizations of a Coronal (/n/, Figure 7.a) and the corresponding Labial (/m/, Figure 7.b) for both speakers pooled (after discounting the difference of 8.44 ms between speakers). It is clear that the overlap between 'strong' and 'weak' realizations of the /n/ (≈ 16%) is much smaller than the corresponding overlap for the /m/ (≈ 32%) to the extent that the combined distribution for the durations of the /n/ is actually bimodal. Moreover, the 'weak' /n/ realizations are quite short indeed (median 47 ms versus 66 ms for /m/). The differences visible between Figures 7.a and b again illustrate that the differences between strong and weak Coronals could indeed be more than a gradual shift and might involve a key change in articulation. The differences between weak versus strong /n/ versus /m/ are all statistical significant ($p \leq 0.001$, two-tailed Median test). The difference in duration between weak and strong /n/ realizations are larger than those between the corresponding /m/ realizations ($p \leq 0.001$, two-tailed Student-t test on average difference).

3.5. The influence of frequency of occurrence

```
##################################################################
#                                                                #
#                      PLACE TABLE 2 HERE                        #
#                                                                #
##################################################################
```

In the introduction, we predicted stronger reduction for the more frequent phonemes (cf., Boersma 1998, sections 9.5 and 10.5; Aylett, 1999a, 1999b; Aylett and Turk, 2004; Van Son et al., 1998; Van Son and Pols, 1999b, 2003a, 2003b; Van Son et al., 2004). The large reduction of the Coronals found in 'weak' positions (word-final and unstressed word-medial) confirms this expectation. In these 'weak' positions, Coronals make up 55% of all realizations versus 41% in 'strong' positions (percentages excluding affricates and /h ? j/). For comparison, the Labials constitute only 23% of all realizations in 'weak' positions versus 38% in strong positions (see Table 2). These numbers are comparable to those found in a list of 88051 word forms from CELEX (not weighted for word frequency, see Table 2). Information content indicates the level of unpredictability. In terms of information content (Information(segment) = $-\log_2$(Prob(segment)), the feature values 'Coronal' and 'Labial' need roughly equal amounts of information in 'strong' positions, but the feature 'Labial' needs 2.6 times as much (acoustic) information as 'Coronal' in 'weak' positions to reach the same predictability (Table 2). In theory, equalizing the unpredictability of redundant phonemes would allow for a proportional increase in the reduction of duration and loudness of Coronals.

This can be illustrated with the differences between the distribution of /t d/ realizations versus those of /p b/ realizations (see Table 2). It was found that coronals from word-medial positions preceding unstressed vowels were shortened (after correction for other factors) with respect to stressed realizations in this position. An explanation might be found in the fact that in the current corpus, coronals are more prevalent in unstressed word-medial position than in stressed word-medial position (for /t d/: 18% vs. 9%, ie. *2.51* vs *3.43* bits, Table 2). In the CELEX word list, the "stressed" word-medial /t d/ realizations make up 13% of all intervocalic consonants in this position (ie., stressed word-medial realizations). Unstressed word-medial /t d/ realizations make up 19% of all intervocalic word-medial consonants in the CELEX word list (ie. 2.90 vs 2.38 bits, Table 2). If all occurrences are used (including clusters), the numbers become 16% for /t d/ in stressed word-medial position and 20% in unstressed word-medial position (ie. *2.63* vs. *2.32* bits, Table 2). It seems that word-medial /t d/ realizations were somewhat under-represented in the corpus used here. However, the pattern seems to be the same. If we compare these distributions with those of /p b/, a different pattern appears. In our corpus, the corresponding numbers are: 14% and 8% (stressed and unstressed word-medial, ie. *2.87* vs. *3.62* bits). In CELEX these frequencies are, respectively 10% and 8% (single consonants, *3.28* vs. *3.57* bits, Table 2), and 10% and 8% (consonant clusters, *3.28* vs. *3.61* bits, Table 2). This is almost a mirror image of the distribution of /t d/.

```
#####################################################################
#                                                                   #
#                    PLACE FIGURE 8.a&b HERE                        #
#                                                                   #
#####################################################################
```

Figures 8.a and b present data for all realizations used in this study. A linear relation between phone duration, CoG, and information content is expected from considerations of communicative efficiency. In Figure 8.a, the relative corrected mean duration of each 'class' of realizations from Figure 3.a with respect to the corrected mean duration of the stress and word-position alone (cf., Figure 1.a, but now for both speakers combined) is plotted against the information contained in the primary articulator label, ie., the negative logarithm of the frequency of occurrence relative to the total number of realizations with the specified stress and position in the word in bits (ie., logarithm to base 2). The correlation between relative duration and the content is obvious (Figure 8.a, Spearman Rank Correlation Coefficient, R = 0.672, p ≤ 0.01, N=18). However, there is a troublesome outlier that prevents a reliable linear regression analysis. The unstressed word-final realizations of the Labials (indicated in gray in Figure 8.a) were exceedingly rare in our corpus: only 71 realizations were present. After removing this outlier, the regression line through the 17 remaining points explained more than 60% of the variance (R=0.798). Using the unexplained variance, the outlier lies more than 4 standard deviations from the corresponding regression line.

The CoG values showed larger consonant intrinsic differences, CoG values are often consonant specific (cf., Jongman et al., 2000), and a straightforward correlation between (relative) CoG values and information was not informative (p>0.05). The CoG values have to be normalized to remove the strong phoneme-specific effects. Therefore, we used a somewhat different procedure. Instead of straight corrected means CoG values, we used the observed difference due to stress and position within a class (ΔO = CoG - Mean(Class)) and the expected difference due to stress and position for this position (ΔE = Mean(Stress&Position) - Mean(All)). The relative observed differences with respect to the expected differences, ie., [ΔO- ΔE]/ ΔE, are plotted in Figure 8.b. For easy comparison with Figure 8.a, we reversed the vertical axis in Figure 8.b. In our analysis of Figure 4, and the construction of our model (Figure 6) we noted that the value of the corrected mean CoG of all the Coronals was not a good value for the correction of the intrinsic CoG value of the Coronals. Therefore, in line with our model, we replaced the Mean(Coronal) with the Mean (PostCoronal) in our calculation of the relative ΔCoG. Note, that the corrected mean values used to determine ΔO and ΔE are insensitive to the number of realizations underlying the individual values.

The regression line in Figure 8.b explains more than 40% of the variation (R=0.650) with the outlier removed. Using the unexplained variance, the designated outlier lies 3 standard deviations from the regression line. Overall, the rank correlation over all 18 points is statistically significant ($p \leq 0.01$, R=-0.645, Spearman Rank Correlation Test, N=18). A positive value of the relative ΔCoG indicates an observed effect that exaggerates the expected effect, whereas a negative value indicates an observed effect that diminishes the expected effect. The data plotted in Figure 8.b suggest that the more frequent a phoneme (class) is found in a specific condition, the more its spectral balance will conform to the trends of the prosodic structure.

From Figure 8 it seems that part of the variation in duration due to phoneme identity, and some of the peculiarities of the spectral balance, can be explained from the relative prevalence of the phonemes in the specific conditions of stress and position in the word. These correlations with frequency of occurrence are even somewhat stronger than their uncorrected mutual correlation (figure 4). The excessive reduction found in the Coronals in 'weak' positions is now in line with the reduction found for the other consonants. The reduction is more extreme because these phonemes are so much more predictable. The results were only marginally different when we used the CELEX derived frequencies of occurrence instead of those from our corpus (regression line fit: duration R=0.843, CoG R=-0.608). The low frequency of occurrence of the labials in unstressed word-final position was even somewhat lower in CELEX than in our corpus.

4. Discussion

Before discussing the results of this study, we would like to reiterate that our study was based on read isolated sentences using all accented words from the female speaker and all words that could be accented (ie., not cliticized) for the male speaker.

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, see section 3.4 and Figure 3.b). 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 also real (Figure 4).

There remains a problem in explaining the consistent behavior of the CoG. It can be defended that the CoG frequencies of the realizations of a particular phoneme are indeed related to speaking effort (see section 2.1). However, both Boersma (1998) and

Pouplier (2003) convincingly argue that there is no global measure of effort that can be used across different articulators (attempts to do this end in circularities). Therefore, there is no a-priori reason to assume that changes in CoG frequencies can be compared between the relevant articulators, ie., sub-glottal pressure and vocal folds for sonorants versus lips and tongue tip, body, and root for obstruents. Still, both Van Son and Pols (1996, 1999a) and the current study point towards a consistent effect for most consonant classes (and the former also for vowels). Part of this consistency can be explained by the fact that the differences all point in the same direction: less effort means lower CoG for all phonemes. But there is a circularity here, as effort is measured as emphasis, which itself is linked to duration and CoG. This still does not explain why the CoG values (in semi-tones) should strongly correlate with duration (in ms, $R^2 \approx 0.70$), as was demonstrated in Figures 4 and 6. This correlation is so strong that we are able to link the duration and CoG in a single model that can be optimized to explain almost half of the variance in the relevant parts of the data sets (Figures 6.a and 6.b duration and CoG versus position in the word, note that we excluded two outliers).

We think that an important additional reason for the consistent behavior of the CoG and its link to duration lies in the implied connection between the CoG and perceived loudness (Boersma, 1998; Sluijter, 1995a, b; Sluijter et al., 1997; Sluijter and Van Heuven, 1996). Subjective loudness is strongly influenced by the spectral balance (eg., Boersma, 1998). English stress perception might be partially based on loudness (as suggested by Sluijter, 1995a, b; Sluijter et al., 1997; Sluijter and Van Heuven, 1996). It is quite conceivable that perception of the other elements of prosody, boundaries, is also at least partially based on loudness. Additionally, the CoG measures how much acoustic power there is in the high frequencies. The lower the CoG, the less information will be available for identification (Van Son and Pols, 1997). The expression of the CoG frequency in semitones instead of Hz also makes more sense when it primarily affects perception. All this would mean that the CoG does not passively follow changes in speaking effort per se, but could be actively manipulated by the speaker to influence loudness and possibly information content, eg., by using more lung pressure. In this case, the CoG value would be an additional marker, both in production and perception, of the prosodic structure of utterances (cf., Tabain, 2003).

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, ie., a strong interaction between these factors was found (Figures 1 and 2). Especially, the identity of the primary 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 (Figures 3.a and 3.b, respectively).

The model as presented in Figure 6.b shows some peculiarities. First, it seems that the unstressed word-initial Coronals behaved like stressed ones. With respect to the CoG of most Coronals, it seemed as if stress was only apparent as a difference between full and flap/tap 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 values in our model, but we have no explanation for this behavior. Finally, after rescaling, the difference between the CoG values of full and reduced, ie., mostly flapped or tapped, Coronals is about 20% larger than the corresponding difference between durations.

After correcting for the overall effects of articulator, we are able to combine our results into a single 'model' (Figures 6.a and b). This model captures a few of the known interactions between main articulator, lexical stress, prosodic (word) boundaries, and allophonic replacement (Zue and Laferriere 1979, Turk and Sawusch 1997, Turk 1992, 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 on the one hand, and word-final and unstressed word-medial realizations on the other hand was extremely large for both duration and CoG (cf., Figure 7.a and b). We strongly suspect that this was the result of a shift in articulation towards reduced allophones. The former were primarily 'fully' articulated, the latter to a large extent reduced, ie., flaps or taps. If we allowed for a third 'extra-weak' level for flaps and taps, all durational and CoG values behaved 'regular' in the sense of fitting a uniform model, as was shown in Figures 6.a, and b. The existence of this flap/tap level is illustrated by the bimodal distribution of the durations of the Coronal /n/ compared to the unimodal distribution of the Labial /m/ (Figures 7.a and 7.b). The fact that we did not separate the flaps from the 'full' consonants might explain some of the differences between our findings and those of Byrd (1993). Using TIMIT, she found less dramatic differences between /n/ and /m/ than we did. However, the large differences in the consonant selection and statistical methods between her study and ours precludes a detailed comparison of results.

Our results support the conclusions of studies on lexical stress and the articulatory emphasizing of prosodic boundaries (de Jong et al., 1993; de Jong, 1995; Fougeron and Keating, 1997; Tabain, 2003; Turk and Sawusch, 1997; also cf., the shortening of word-final schwa in Byrd and Saltzman, 1998). The articulatory measurements presented in these studies are generally consistent with the idea that the emphasizing of syllables at prosodic boundaries was accompanied by an increase in effort in the form of more extensive articulation (eg., Fougeron and Keating, 1997; but see the warning of Pouplier, 2003). Our results suggest that this increase in effort might also be visible as a corresponding shift in the spectral balance of the consonants, which is known to reflect subjective loudness and possibly speech effort (Sluijter, 1995 a, b; Sluijter and Van Heuven, 1996; Sluijter et al., 1997; Van Son and Pols, 1996; 1997, 1999a). This shift in the spectral Center of Gravity and, therefore, possibly effort, 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, eg., 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 (note: their word-final /n/ realizations are word-medial by our definition).

We expected stronger reduction for the more frequent phonemes (see Introduction, cf., Aylett, 1999a, 1999b; Aylett and Turk, 2004; Boersma 1998; Borsky et al., 1998; Lieberman, 1963; Van Son et al., 1998; Van Son and Pols, 1999b, 2003a, 2003b; Van Son et al., 2004; Vitevitch et al., 1997). Phone duration and Center of Gravity are linearly related to the “bandwidth” of the speech communication channel (eg., Van Son and Pols, 1999a). Therefore, after accounting for other phonetic influences, some kind of linear relation between these acoustic measures and information content is theoretically the most efficient. Our results show that redundancy is indeed a powerful, and linear, predictor of the level of reduction (Figures 8.a and 8.b). The large 'irregular' reduction of the Coronals in 'weak' positions can to a large extent (eg., up to 60%) be explained as the result of a very high relative frequency of occurrence in these conditions. The correlations shown in Figures 8.a and 8.b without the weak coronals would actually predict the reductions in duration and spectral CoG seen in flaps/taps given their redundancy.

The correlations found in figure 8.a and b support the interpretation that the level of reduction in speech is to some part a reflection of the predictability of the item, even if the item is a single phoneme. This contrasts with the study of Aylett and Turk (2004) who concluded that prosodic structure could account for almost all redundancy effects in their syllable data. A purely prosodic explanation of the results in our figures 8.a and b would be rather complex. In contrast, a straightforward explanation of our data is possible based on predictability. The more predictable an item is, the less information a listener needs to identify it. A shorter duration, lower loudness, and a smaller spectral 'bandwidth' (ie., lower CoG) all reduce the amount of information available to the listener while at the same time reducing the articulatory 'effort' of the speaker. So these three markers of reduction might be mutually correlated because they are all linked to the amount of 'acoustic' information required by the listener. Of course, this has as a side effect that spoken communication becomes more efficient.

Note that the current study does not identify *how* speakers achieve this efficiency. It is certainly more complex than that “reduction directly follows local information content”. The tongue-tip used to articulate Coronals is the most agile articulator. Coronal consonants are more easy to articulate, and can be shorter, than corresponding consonants pronounced with other articulators. According to Boersma (1998) this would tend to increase the frequency of the Coronals in languages and consequently, would make it economical to reduce them. Likewise, stops and nasals are “easy” manners of articulation which would tend to increase their frequency. So it might not be accidental that the easy to articulate Coronal stops and nasals have high frequencies and at the same time have a strongly reduced allophone, the flap or tap. However, these phonological questions are outside the scope of the current study.

Our study also showed that factors affecting psycholinguistic processes, eg., the effects of frequency of occurrence and position in the word, can help to understand prosodic structure, eg., variation in duration, and that large amounts of speech data are needed to 'deconfound' the interacting factors in fluent speech, and even then it is often difficult to obtain statistically convincing evidence.

5. Conclusions

From a large corpus of read meaningful sentences from two speakers of American English we were able to model the quantitative effects of stress and position in the word on duration and spectral balance of intervocalic consonants. Special statistical methods were used that could 'deconfound' the unbalanced data and estimate the 'balanced' mean values. It proved that the articulatory strengthening and durational lengthening for stressed syllables and around word boundaries reported in other studies was reflected in the spectral balance of our consonantal data. A simple three-tiered representation captured both our durational and spectral results. A large modeling 'irregularity' in the behavior of Coronal consonants suggested a switch in articulation between *full* and *reduced* (ie., flaps or taps) in de-emphasizing circumstances. This switch, together with a large part of the consonant specific variation, could be correlated to the stress and position specific frequency of occurrence, or redundancy, of the individual phoneme classes. This leads us to conclude that acoustic reduction in intervocalic consonants is affected by both prosodic structure and phoneme redundancy. The effects of these factors on duration and spectral balance are such that the communicative efficiency of speech is increased.

6. Acknowledgments

This research was made possible by grants 300-173-029, 355-75-001, and 276-75-002 of the Netherlands Organization for Scientific Research (NWO). NWO also supplied the additional funding which made this collaboration possible. This work was partly performed at Bell Labs, Lucent Technologies, Murray Hill NJ, USA. We would like to thank the anonymous reviewers for their many detailed and helpful comments.

7. References

Aylett, M., 1**999a**. 'Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and syllabic duration', *Proceedings of ICPhS'99*, San Francisco, 289-292.

Aylett, M., 1999b. *Stochastic suprasegmentals: Relationships between redundancy, prosodic structure and care of articulation in spontaneous speech*, PhD thesis, University of Edinburgh, 190 pp.

Aylett, M., Turk, A., 2**004**) 'The Smooth Signal Redundancy Hypothesis: A Functional Explanation for Relationships between Redundancy, Prosodic Prominence and Duration in Spontaneous Speech', *Language and Speech* **47**, 31-56

Boersma, P.P.G., **1998**. *Functional phonology: formalizing the interactions between articulatory and perceptual drives*, IFOTT/:LOT **11**, PhD. thesis, University of Amsterdam, p 493., parts are available at Rutgers Optimality Archive: http://roa.rutgers.edu/ and http://www.fon.hum.uva.nl/paul/papers/ contains the complete thesis)

Borsky, S., Tuller, B. and Shapiro, L.P., **1998**. "How to milk a coat:' The effects of semantic and acoustic information on phoneme categorization'. *Journal of the Acoustical Society of America* **103**, 2670-2676.

Byrd, D., **1993**. '54,000 American stops'. *UCLA Working Papers in Phonetics* **83**, 97-115.

Byrd, D. and Saltzman, E., **1998**. 'Intragestural dynamics of multiple prosodic boundaries', *Journal of Phonetics* **26**, 173-199.

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.

Clark, J. and Yallop, C., **1990**. *An introduction to phonetics and phonology,* Basil Blackwell, Oxford Cambridge.

Cooper, A.M., **1991**. 'Laryngeal and oral gestures in English /p, t, k/', *Proceedings of ICPhS'91*, Aix-en-Provence, 50-53.

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 Speec*h **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.

Dodge, Y., **1981**. *Analysis of experiments with missing data*, Wiley, New York NY.

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

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.

Hanson, H.M., **1997.** 'Glottal characteristics of female speakers: Acoustic correlates' *Journal of the Acoustical Society of America* **101**, 466-481.

Hanson, H.M., and Chuang E.S., **1999**. 'Glottal characteristics of male speakers: Acoustic correlates and comparison with female data', *Journal of the Acoustical Society of America* **106**, 1064-1077.

Jongman, A., Wayland, R., and Wong, S., 2**000**. "Acoustic characteristics of English fricatives", *Journal of the Acoustical Society of America* 106, 1252-1263.

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, p 163.

Lieberman, P., **1963**. 'Some effects of semantic and grammatical context on the production and perception of speech', *Language and Speech* **6**, 172-187.

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.

Nespor, M. and Vogel, N., 1986. *Prosodic phonology*, Foris publications, Holland, in series: *Studies in generative grammar*, eds.: Jan Koster and Henk van Riemsdijk, p 327.

Nord, L., **1987**. 'Vowel reduction in Swedish', *Papers from the Swedish Phonetics Conference*, edited by O. Engstrand, 16-21.

O'Shaughnessy, D., **1987**. *Speech Communication, Human and Machine*, in *Addison-Wesley Series in Electrical Engineering: Digital Signal Processing*, Addison-Wesley), p 568.

Pouplier, M., 2003. 'The dynamics of error', *Proceedings of ICPhS 2003*, Barcelona, 2245-2248.

Rietveld, A.C.M. and 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, p 188.

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.

Sproat, R., **1998**. *Multilingual text-to-speech synthesis: The Bell-labs approach*, R. Sproat, ed.), Kluwer academic publishers, Dordrecht, p 300.

Tabain, M., 2**003**. Effects of prosodic boundary on /aC/ sequences: Acoustic results, *Journal of the Acoustical Society of America* **113**, 516-531.

Turk, A.E., **1992**. 'The American English flapping rule and the effect of stress on stop consonant durations', *Working Papers of the Cornell Phonetics Laboratory*, Vol. 7, 103-133.

Turk, A.E. and Sawusch, J.S., **1997**. 'The domain of accentual lengthening in American English', *Journal of Phonetics* **25**, 25-41.

Turk, A.E., and Shattuck-Hufnagel, S., 2**000**. "Word-boundary-related duration patterns in English", *Journal of Phonetics* **28**, 397-444.

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* IFOTT **16**. Ph.D. Thesis, University of Amsterdam, p 195.

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*, Berlin, 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., Bolotova, O., Lennes, M., and Pols, L.C.W., 2**004**. 'Frequency Effects on Vowel Reduction in Three Typologically Different Languages, (Dutch, Finnish, Russian)', *Proceedings of INTERSPEECH 2004*, Jeju Island, South Korea, 1277-1280.

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, 2135-2138.

Van Son R.J.J.H. and Pols, L.C.W., **1999a**. 'An acoustic description of consonant reduction', *Speech Communication* **28**, 125-140.

Van Son R.J.J.H. and Pols, L.C.W., **1999b**. 'Effects of Stress and Lexical Structure on Speech Efficiency', *Proceedings of Eurospeech'99*, Budapest, 439-442.

Van Son, R.J.J.H. and Pols, L.C.W., 2**002**. 'Evidence for Efficiency in vowel production', *Proceedings of ICSLP2002*, Denver, USA, Vol I, 37-40.

Van Son, R.J.J.H. and Pols, L.C.W., 2**003a**. 'An Acoustic Model of Communicative Efficiency in Consonants and Vowels taking into Account Context Distinctiveness', *Proceedings of ICPhS 2003*, Barcelona, Spain, 2141-2144.

Van Son, R.J.J.H. and Pols, L.C.W., 2**003b**. 'Information Structure and Efficiency in Speech Production', *Proceedings of EUROSPEECH2003*, Geneva, Switzerland, 769-772.

Van Son, R.J.J.H., Koopmans-van Beinum, F.J. and Pols, L.C.W., **1998**. 'Efficiency as an organizing principle of natural speech', *Proceedings ICSLP'98*, Sidney, Australia, 2395-2398.

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

Vitevitch, M.S., Luce, P.A., Charles-Luce, J., and Kemmerer, D., **1997**. 'Phonotactics and syllable stress: Implications for the processing of spoken nonsense words', *Language and Speech* **50**, 47-62.

Wang, X., **1997**. *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, in *Studies in Language and Language Use* IFOTT **29** Ph.D. Thesis, University of Amsterdam, p 190.

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.

Zue, V.W., and Laferriere, M., **1979**. 'Acoustic study of medial /t,d/ in American English', *Journal of the Acoustical Society of America* **66**, 1039-1050.

Footnotes

[1] 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 (cf., Wightman et al, 1992).

[2] In Van Son and Van Santen (1997) we accidentally reported the total number of intervocalic consonants, ie., including 683 realizations of the affricates and /h/ which are not used in the current study (see text).

[3] 233 presumed realizations of /j/ were included in Van Son and Van Santen (1997). The complete corpus was labeled in the context of several projects. Over the course of these efforts, different labels had been used for /j/ and this was not well documented. We were unable to ensure that all realizations of the /j/ would be extracted and, as a result, there would be a danger of ending up with a biased sample. Therefore, we decided to exclude all /j/ realizations.

[4] No consistent effects were found using the alternative rule of using the stress of the preceding vowel on word-initial or word-medial consonants, or the stress of the following vowel on word-final consonants. However, there is a very strong patterning of stress, eg., stress alternation and the avoidance of stress clashes. This patterning introduces very strong correlations between the occurrences of stressed and unstressed syllables. For instance, a stressed syllable will *always* be surrounded by unstressed syllables in the same word (ignoring secondary stress or emphatic speech), and can only rarely be found next to a stressed syllable from another word. These correlations prevented us from pursuing the problem of stress assignment in intervocalic consonants in depth with our limited corpus.

Another possibility is that word-medial intervocalic consonants are ambisyllabic and acquire (some of) the stress of a preceding stressed vowel. An unweighted, ie, dictionary, count of 88051 English CELEX word-forms showed that of all word-medial intervocalic consonant realizations, more are *preceded* than *followed* by a stressed vowel (24.8% vs. 20.6%, 54.6% were surrounded by unstressed vowels). From this distribution and using our maximum-onset rule of stress assignment; we would predict that the effects of lexical stress in word-medial position would be smaller than those in word-initial or word-final position. The effects would be smaller because stressed realizations would be mixed with unstressed ones in ambiguous medial position, whereas there are no ambiguities about stress assignment in initial and final positions. Moreover because of the prevalence of word-initial stress, if intervocalic consonants were all ambisyllabic, we would expect that assigning the stress of the preceding vowel to the consonant would show equal or greater effects of syllable stress on duration and CoG of this consonant, than using the stress of the following vowel as was done. Neither of these propositions was supported by our preliminary analysis.

A weaker proposition is that the ambisyllabicity of word-medial intervocalic consonants only affects consonants that originated from a coda-position. A count of 88051 English word-forms in CELEX (not weighted for word frequency), showed that only 8.2% of the intervocalic consonants are both word-medial and in the coda of a syllable (1.5% following and 3.6% preceding a stressed syllable, the rest were surrounded by unstressed vowels). These numbers are much too low to allow us to investigate this matter using our corpus. Therefore, we decided to limit our research to stress assignment according to the maximum-onset criterion. We leave the question of ambisyllabicity to studies using a more suitable corpus.

5: The supra-glottal filtering of the voiced sounds could alter the CoG independent of the source. However, it is not clear why supra-glottal filtering should lead to a uniform decrease in CoG with reduction (as was shown in Van Son and Pols, 1999). Even in nasals, where there is severe damping, the CoG is lower in reduced speech (Van Son and Pols, 1999a).

6: Throughout the paper we indicate plosive closures with the plain IPA label, and the plosive burst & aspiration with the italic underlined IPA label.

7: This definition of Coronals is somewhat arbitrary. For example, the primary articulator of the /l/, tongue tip or blade, depends on its position in the word, and there will often be a strong secondary blade articulation. Looking at Figure 2, we might rather have grouped the /l/ with the Post-Coronals. However, this would have been rather post-hoc. Furthermore, the primary articulator of the /r/ is the tongue tip, but as a retroflex, it has a different place of articulation as the other coronals. As the articulation of the /r/ also depends strongly on movements of the other parts of the tongue and has its constriction further to the back than all the other coronals, we decided not to include it with the other coronals in our analysis.

Table 1. Example of an incidence matrix used to calculate the corrected mean durations of all six combinations of syllable stress (+ or -) and position in the word (Initial, Medial, and Final, eg, 'I+' - stressed word initial). Rows contain combinations of the factor values of interest, eg., stress and position in the word. Columns contain combinations of the other factors that do affect the measured values, but are not of interest, so called *nuisance* factor levels. There are 15 mean cell-by-cell row differences (ie., $\Sigma_k(C_{i,k.}-C_{j,k})$) for 6 rows, eg., (M+) - (F-) or (I+) - (I-), and therefore 15 linear ('normal') equations for the 6 hypothetical mean values. Solving these 15 equations gives the 'best estimates', in a least squares error sense, for the mean duration of each row. Note that, in general, across the whole incidence matrix less than 50% of the cells will be filled, eg., /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*	$C_{1,4.}$....
I -	-	*mean*	*mean*	$C_{2,4.}$.....
M +	*mean*	*mean*	-	$C_{3,4.}$.....
M -	*mean*	*mean*	-	$C_{4,4.}$.....
F +	*mean*	-	*mean*	$C_{5,4.}$.....
F -	*mean*	-	*mean*	$C_{6,4.}$.....

Table 2. Frequency of occurrence and $-\log_2$ thereof (bits, italic) of intervocalic consonants in our corpus and of single and clustered consonants in an unweighted list of 88051 English word forms from CELEX (percentage of relevant consonant realizations combined from a dictionary count). Strong: all Word-Initial and stressed Word-Medial consonants. Weak: all Word-Final and unstressed Word-Medial consonants. Corpus: consonants used in this study. Word C_1: counts of non-cluster (single consonant) word-initial, intervocalic medial, and final consonants in CELEX word forms. Word C_N: as Word C_1 but pooling consonants from multi-consonant clusters. The CELEX counts includes all consonants, ie., also the palatals, glottals and affricates which were excluded from our corpus. Percentages might not add to 100% due to rounding. A specific example is given in the lower part for /p b/ versus /t d/ realizations in Word Medial position.

	Strong			Weak		
	Corpus	Word C_1	Word C_N	Corpus	Word C_1	Word C_N
Labials	38% *1.41*	33% *1.59*	28% *1.83*	23% *2.15*	17% *2.17*	16% *2.65*
Coronals	41% *1.29*	34% *1.52*	40%*1.32*	55% *0.85*	54% *0.89*	60% *0.74*
PostCor	22% *2.21*	32% *1.65*	32% *1.66*	22% *2.18*	29% *1.79*	24% *2.05*
	Stressed Word Medial			Unstressed Word Medial		
/p b/	14% *2.87*	10% *3.28*	10% *3.28*	8% *3.62*	8% *3.57*	8% *3.61*
/t d/	9% *3.43*	13% *2.90*	16% *2.63*	18% *2.51*	19% *2.38*	20% *2.32*

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 indicate statistically significant differences between adjacent word positions; the symbols 'FM' (female, male) indicate statistically significant differences between stress conditions, within a word position: $p \leq 0.002$, 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).

Figure 2. Differences in corrected mean duration between consonants from stressed and unstressed syllables. Because of their rarity in our corpus, /θ ð ʒ ʃ ŋ/ were not included (however, they were used to calculate the total values). Unless a consonant occurred 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. /gkdtbp/: plosive closure durations, / gkdtbp/: plosive burst+aspiration durations.

a. Male speaker

b. Female speaker

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

a. Corrected mean Duration

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

Figure 4. Correlation between the corrected mean frequencies of the corrected mean Spectral Center of Gravity (semitones) and the corrected mean durations. The data-points from Figures 3.a and 3.b are shown. 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 primary articulator (see text): R = 0.829 ($p \leq 0.001$, Student $t = 5.12$, $\nu = 13$, two-tailed, testing significance of R).

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

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

Figure 6. The data of Figure 3.a and 3.b, but now corrected for the overall effects of the primary articulator (indicated by the numbers inside the plot). Both plots were drawn to the same scale. The lines correspond to the least RMS error fits with forced equal slopes for strong, weak, and extra-weak tiers (data of Figures a and b pooled, R = -0.702, see text).

a. Corrected mean Duration. The overall effect of the Post-Coronals was used to adapt the mean corrected duration of the Coronals (see section 3.4).

b. Corrected mean 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 section 3.4).

Figure 7. Histograms of durations for realizations of a. /n/ (Coronal) and b. /m/ (Labial). Durations are pooled for all word-initial (I^{+-}) and Stressed word-medial (M^{+}) realizations (dark bars, right scale) and for Unstressed word-medial (M^{-}) and all word-final (F^{+-}) realizations (light bars, left scale). Realizations of both speakers are pooled, but 8.44 ms were added to the durations of the male speaker to account for the speaker differences (cf., Figure 1). The median durations (after speaker correction) are 47 and 77 ms for the /n/ and 66 and 88 ms for the /m/. The differences between Weak versus Strong /n/ versus /m/ are all statistical significant (p ≤ 0.001, two-tailed Median test). The difference in duration between Weak and Strong /n/ realizations are larger than those between the corresponding /m/ realizations (p ≤ 0.001, two-tailed Student-t test on average difference). Scale ranges are adapted to the total number of realizations. Note the difference in overlap between light and dark bars for /n/ and /m/ (see text).

Figure 8.a. Relative corrected mean durations versus the information carried by the consonant class, ie., the negative logarithm of the relative frequency of occurrence in bits. The relative durations are the corrected mean durations of Figure 3.a divided by the expected durations, ie., the corresponding corrected mean durations of only stress and position (cf., Figure 1.a, but for both speakers combined). The Spearman Rank Correlation coefficient for all 18 duration points is $R_{SP} = 0.688$, $p \leq 0.01$. The linear regression line was drawn after excluding the outlier and based on a Pearson's correlation coefficient $R_P=0.798$ (outlier >4 sd from regression line, shaded circle, unstressed word-final Labial, n = 17, see text, for CELEX frequencies R=0.843). The values along the horizontal axis represent the information contained in the consonant class and are defined as $\#Bits = -\log_2(n_{ART}/N_{POS})$ in which n_{ART} is the number of realizations of the specified item in Figure 3.a (ie., identical articulator, position in the word, and stress), and N_{POS} is the sum of all three n_{ART} values with the same position in the word and stress but different primary articulators.

Figure 8.b. As 8.a., but now the relative corrected mean CoG differences versus the information carried by the consonant class (in bits). Plotted is the difference between the observed effect of stress and position for the consonant class (ΔO = CoG - Mean(Class), Class is Labial, Coronal, or PostCoronal) and the expected effect (ΔE = Mean(Stress&Position) - Mean(All)) divided by the expected effect, ie., $[\Delta O- \Delta E]/ \Delta E$. In line with the analysis of Figure 4, the Mean(Coronal) was set equal to Mean(Post-Coronal), see text. The vertical axis has been reversed for easy comparison with Figure 8.a. The Spearman Rank Correlation coefficient for all 18 ΔCoG points is $R_{SP} = -0.645$, $p \leq 0.01$. The linear regression line was drawn after excluding the outlier (>3 sd from regression line, shaded circle) and based on a Pearson's correlation coefficient $R_P=-0.650$. Note that the same information values were used as in Figure 8.a (see text, for CELEX frequencies R=-0.608).

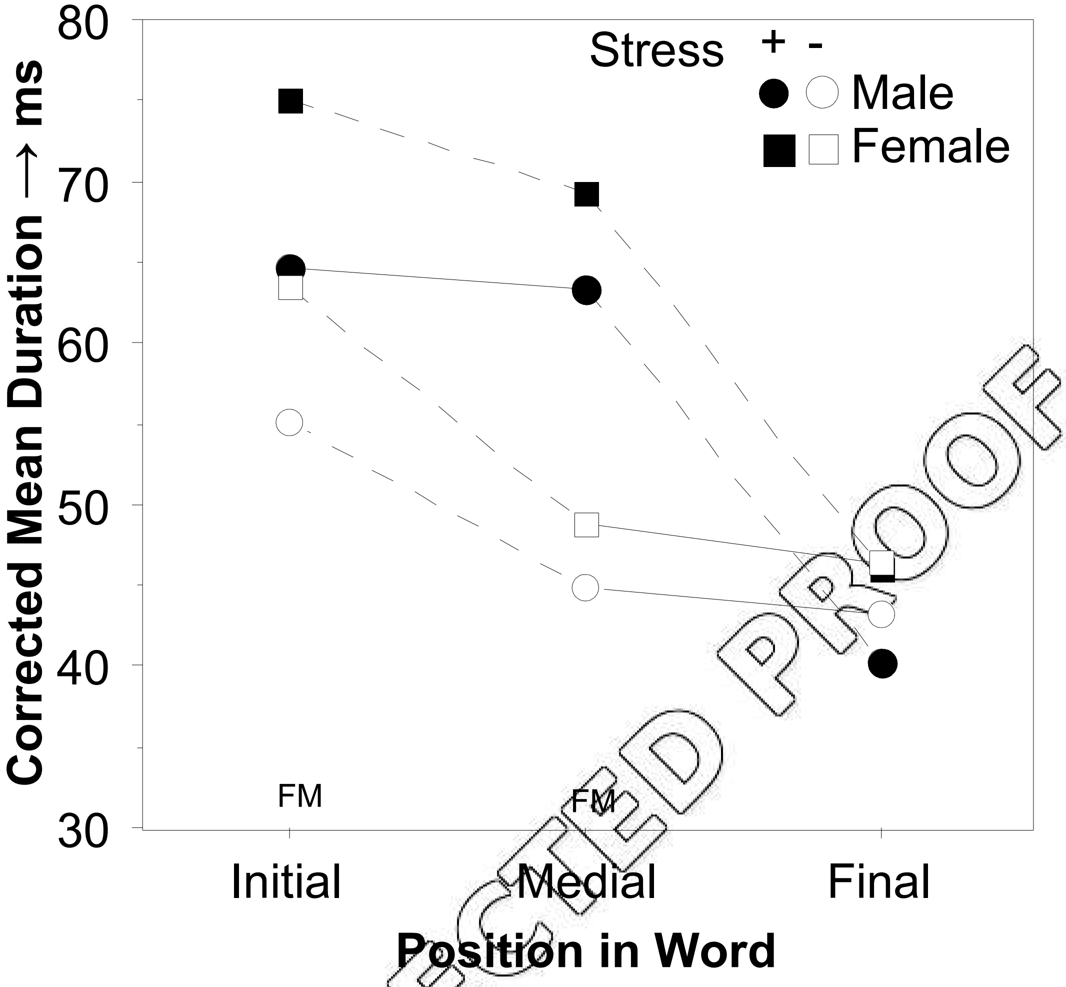
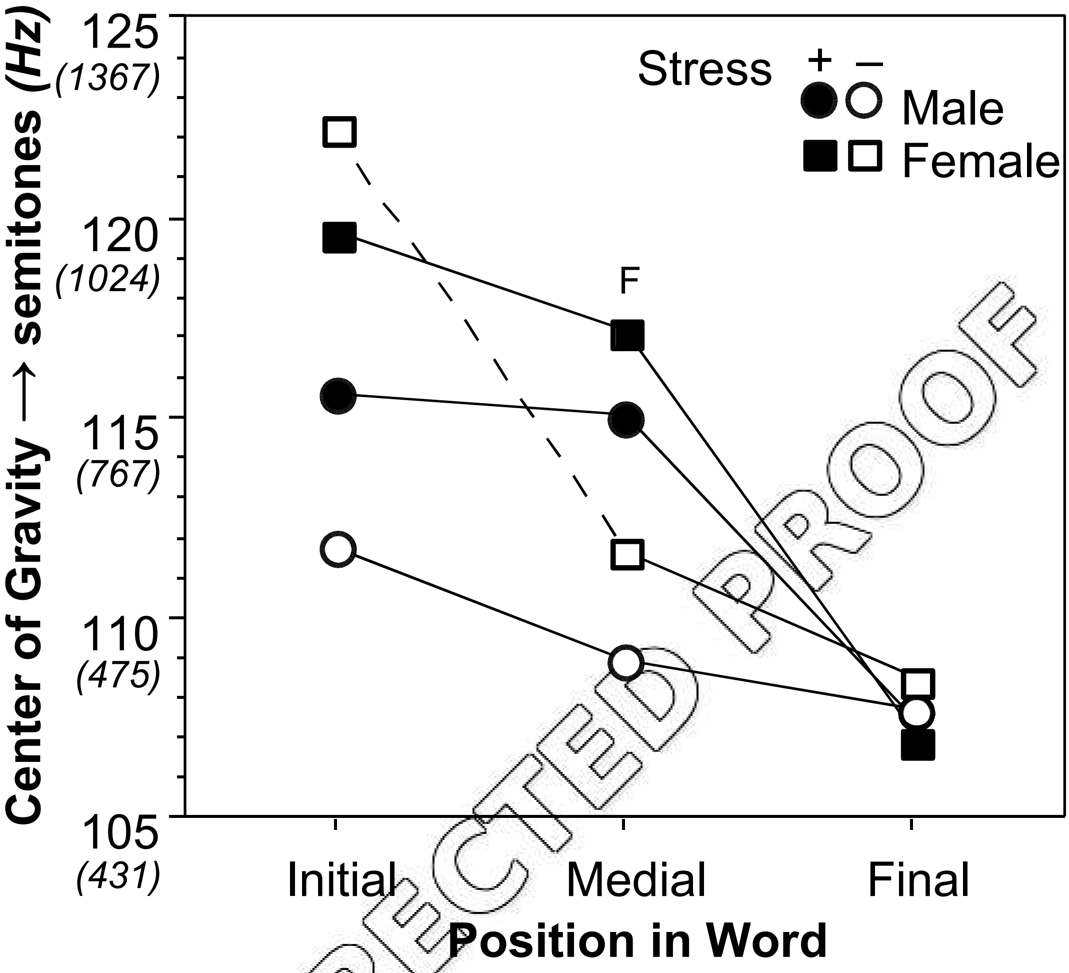

80
70
60
50
40
30
Corrected Mean Duration → ms
Stress + -
● ○ Male
■ □ Female
FM
FM
Initial
Medial
Final
Position in Word

Figure 1.a.


Center of Gravity → semitones (Hz)
125
(1367)
120
(1024)
115
(767)
110
(475)
105
(431)
Stress + –
Male
Female
F
Initial
Medial
Final
Position in Word

5 Figure 1.b.

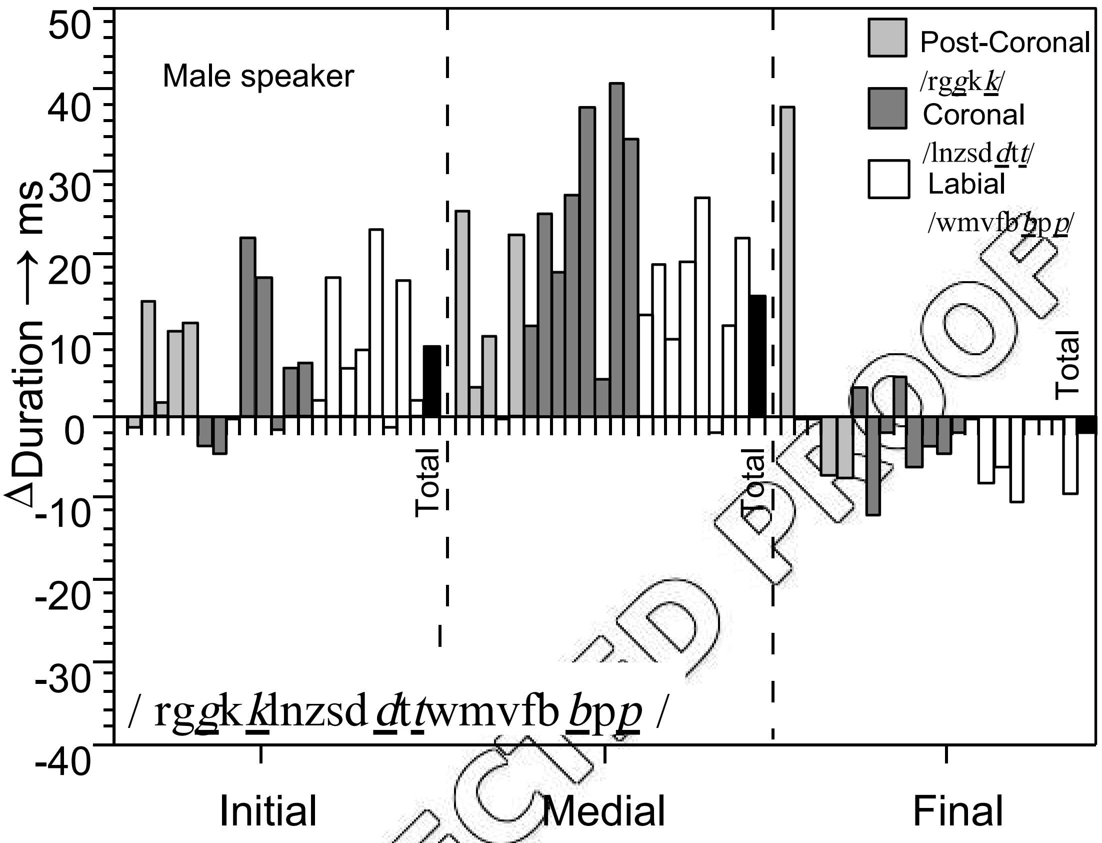
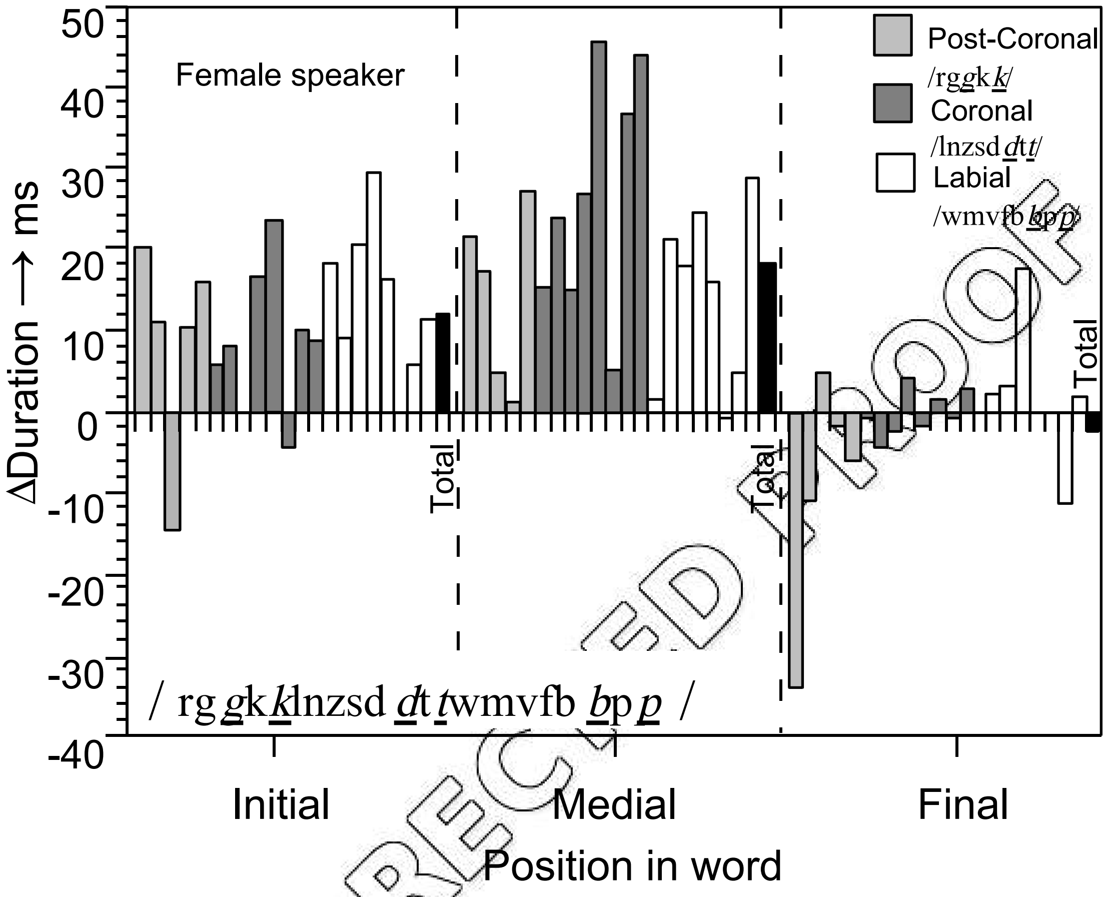

50
40
30
20
10
0
-10
-20
-30
-40
ΔDuration ⟶ ms
Male speaker
Post-Coronal
/rggkk/
Coronal
/lnzsddtt/
Labial
/wmvfbbpp/
Total
Total
Total
/ rggkklnzsd ddttwmvfb bpp /
Initial
Medial
Final

Figure 2.a.


50
40
30
20
10
0
-10
-20
-30
-40
ΔDuration → ms
Female speaker
Post-Coronal
/rg*g*k*k*/
Coronal
/lnzsd*d*t*t*/
Labial
/wmvfb*b*p*p*/
Total
Total
Total
/ rg *g*k *k*lnzsd *d*t *t*wmvfb *b*p *p* /
Initial
Medial
Final
Position in word

5 Figure 2.b.

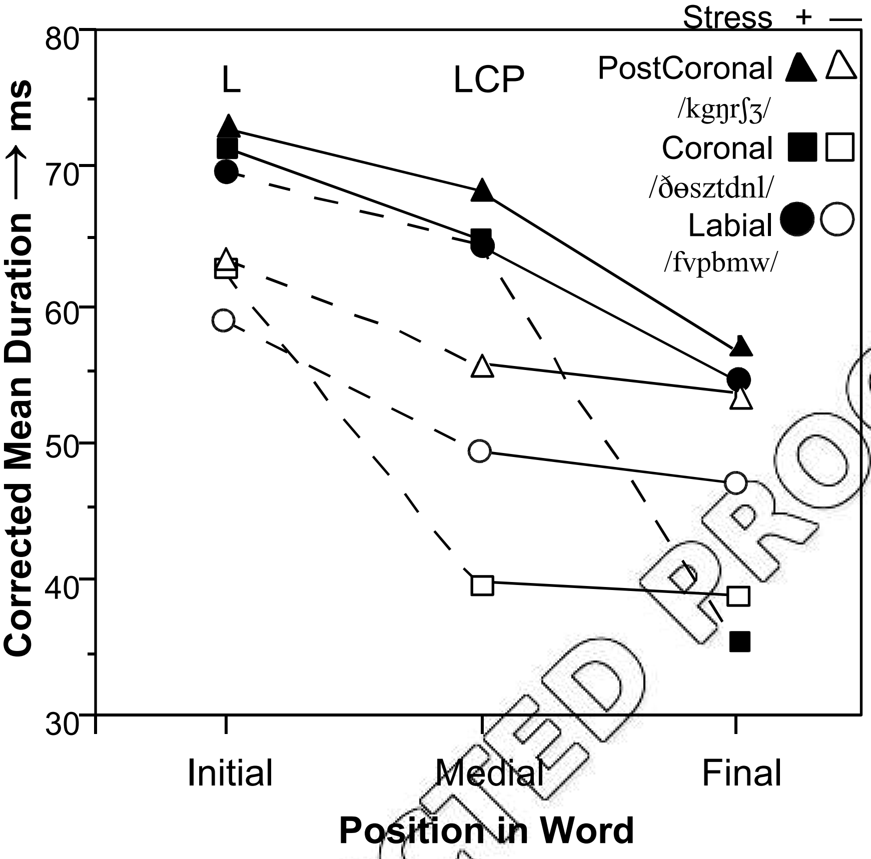

Stress + —
PostCoronal ▲ △
/kgŋrʃʒ/
Coronal ■ □
/ðɵsztdnl/
Labial ● ○
/fvpbmw/
L
LCP
80
70
60
50
40
30
Corrected Mean Duration ⟶ ms
Initial
Medial
Final
Position in Word

Figure 3a.

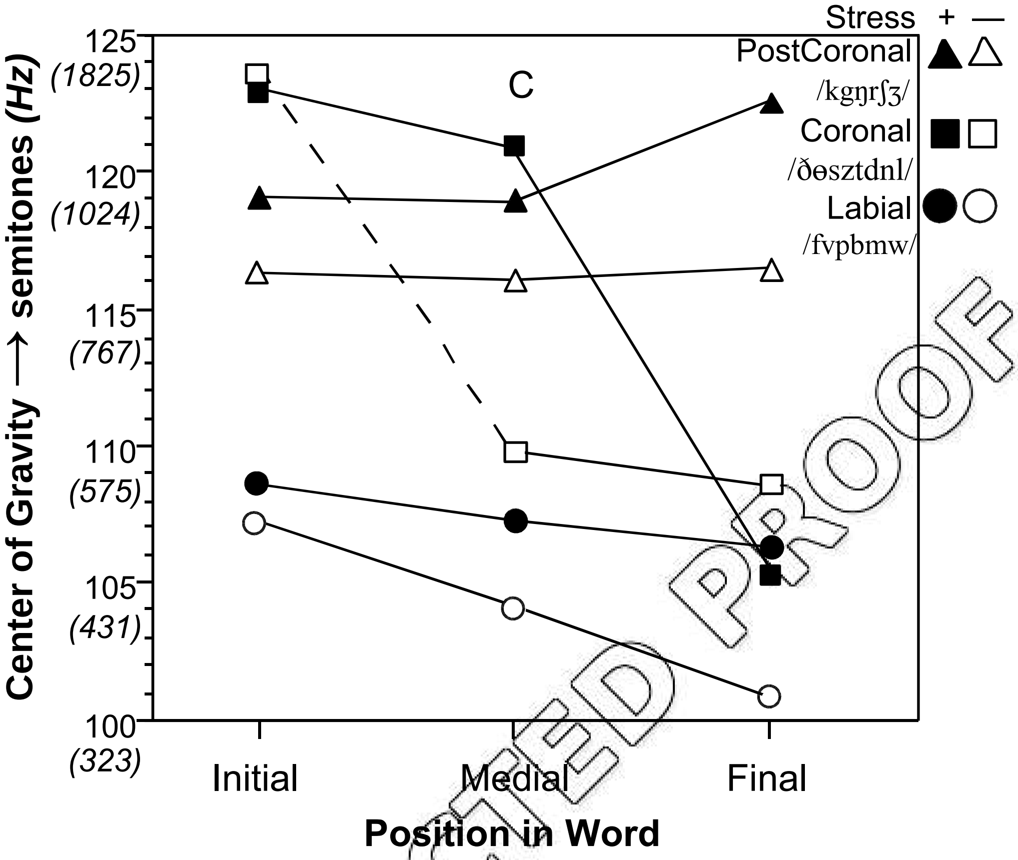

Stress + —
PostCoronal ▲ △
/kgŋrʃʒ/
Coronal ■ □
/ðθsztdnl/
Labial ● ○
/fvpbmw/
C
Center of Gravity ⟶ semitones (Hz)
125
(1825)
120
(1024)
115
(767)
110
(575)
105
(431)
100
(323)
Initial
Medial
Final
Position in Word

Figure 3. b.

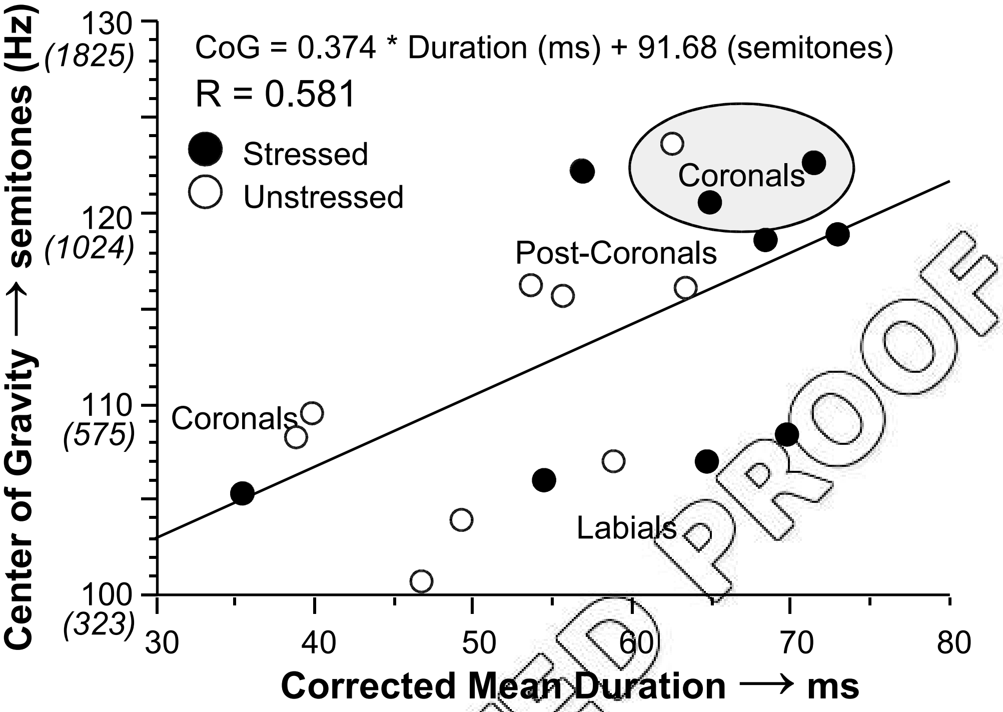

CoG = 0.374 * Duration (ms) + 91.68 (semitones)
R = 0.581
● Stressed
○ Unstressed
Coronals
Post-Coronals
Coronals
Labials
Center of Gravity → semitones (Hz)
130
(1825)
120
(1024)
110
(575)
100
(323)
30
40
50
60
70
80
Corrected Mean Duration → ms

Figure 4.

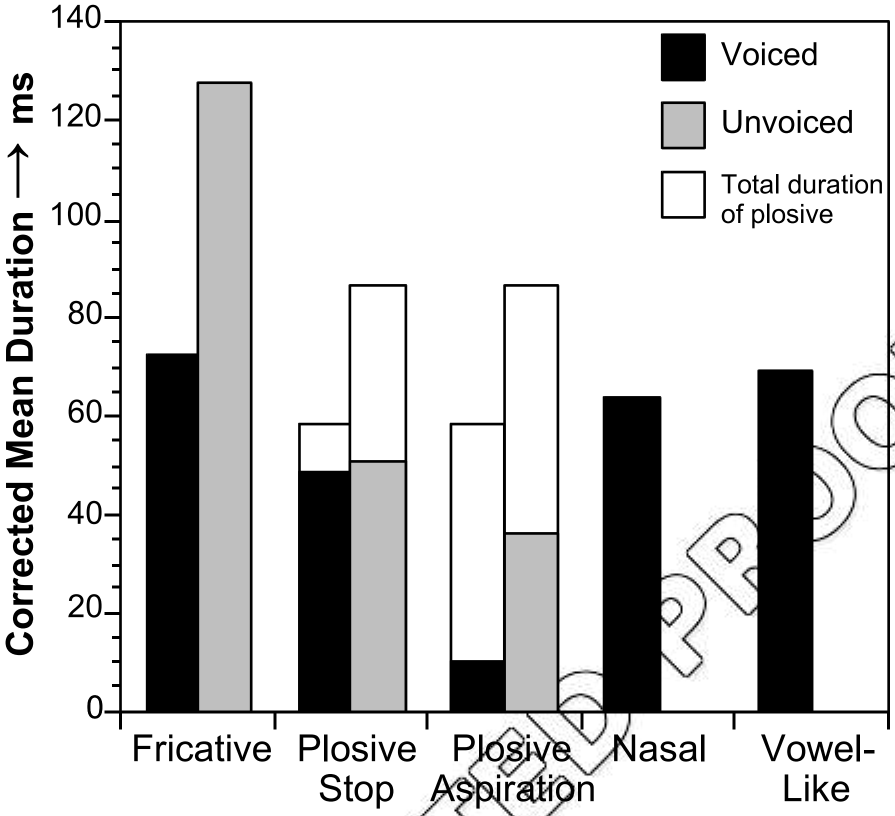

Corrected Mean Duration ⟶ ms
0
20
40
60
80
100
120
140
Voiced
Unvoiced
Total duration of plosive
Fricative
Plosive Stop
Plosive Aspiration
Nasal
Vowel-Like

Figure 5.a.

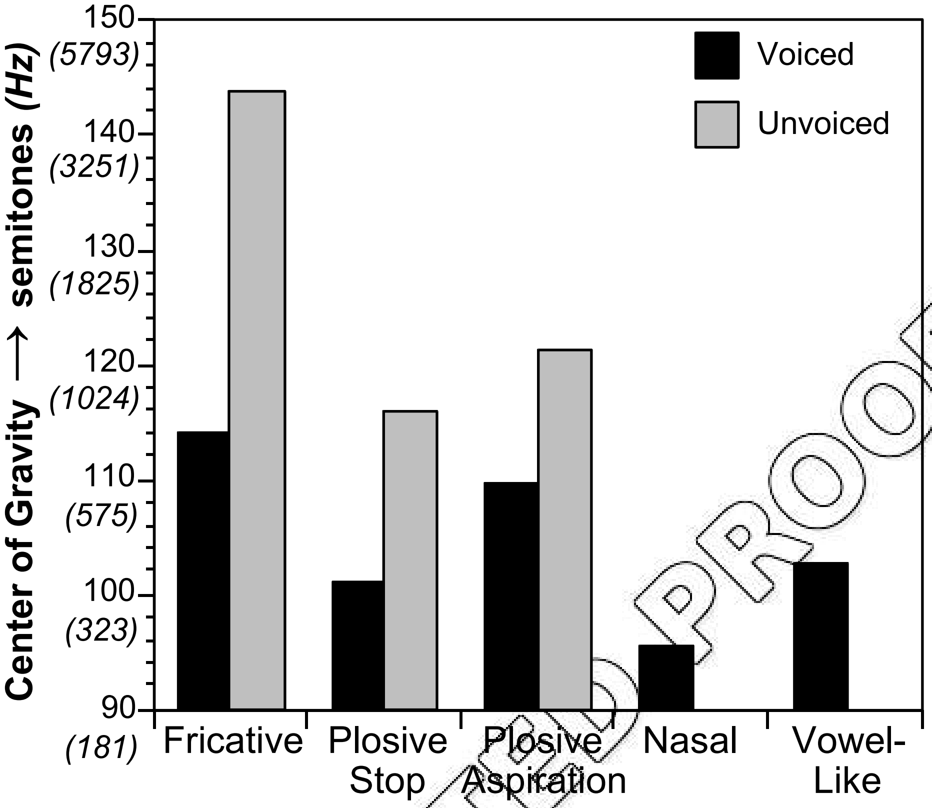

150
(5793)
140
(3251)
130
(1825)
120
(1024)
110
(575)
100
(323)
90
(181)
Center of Gravity → semitones (Hz)
Voiced
Unvoiced
Fricative
Plosive Stop
Plosive Aspiration
Nasal
Vowel-Like

Figure 5.b.

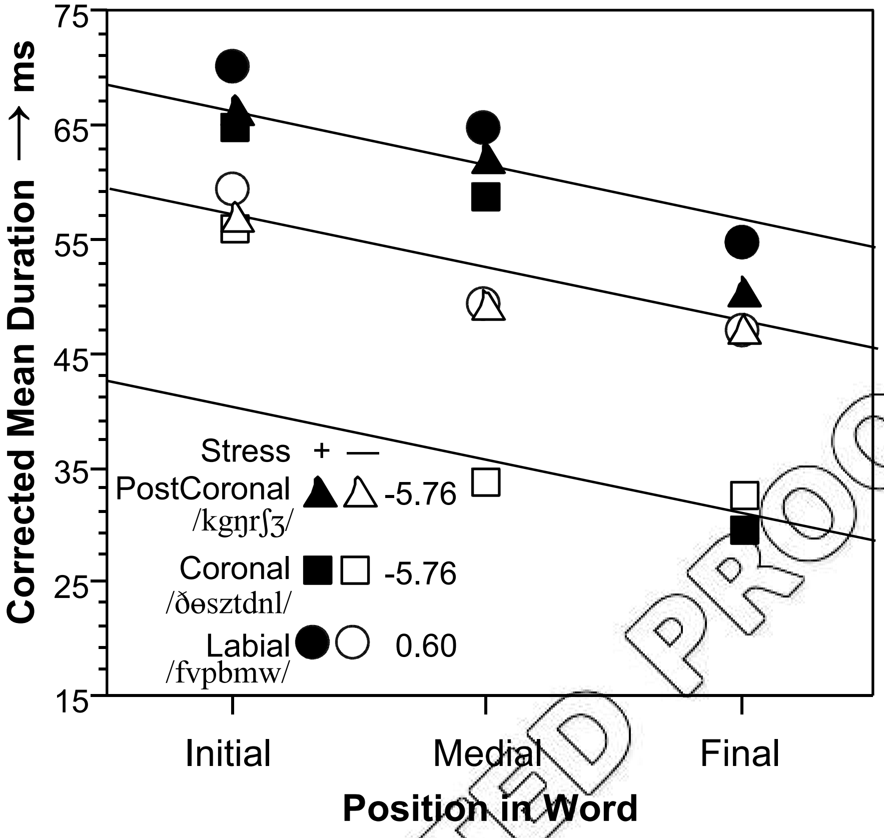

75
65
55
45
35
25
15
Corrected Mean Duration ⟶ ms
Stress + —
PostCoronal ▲ △ -5.76
/kgŋrʃʒ/
Coronal ■ □ -5.76
/ðɵsztdnl/
Labial ● ○ 0.60
/fvpbmw/
Initial
Medial
Final
Position in Word

Figure 6.a.

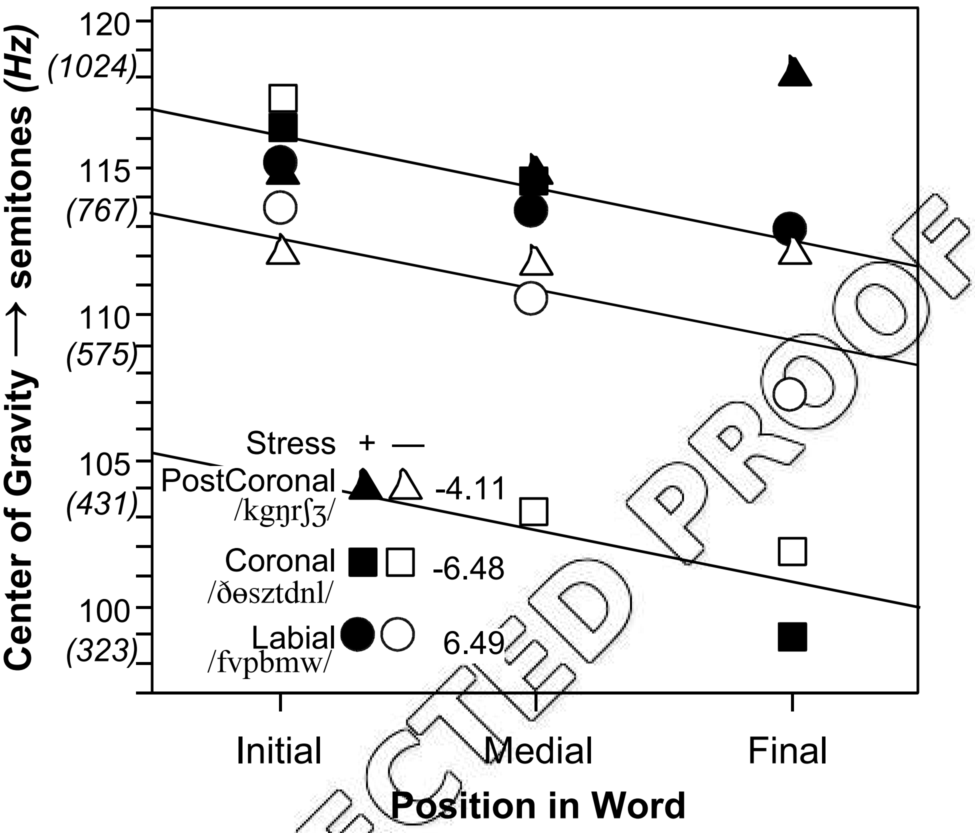
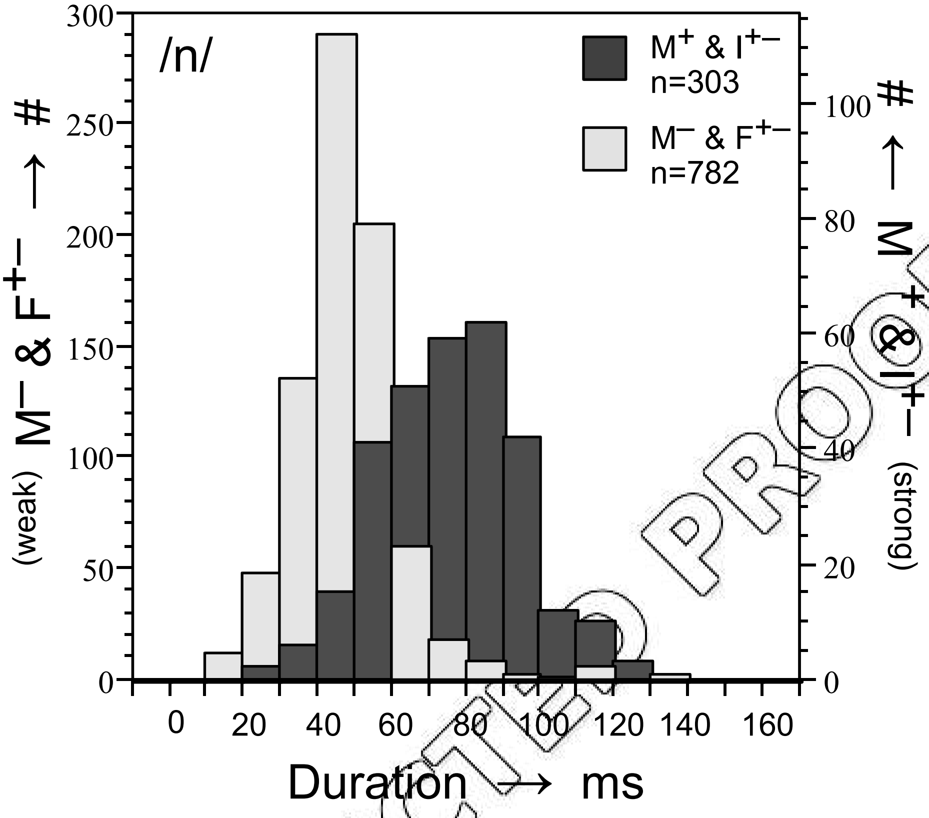
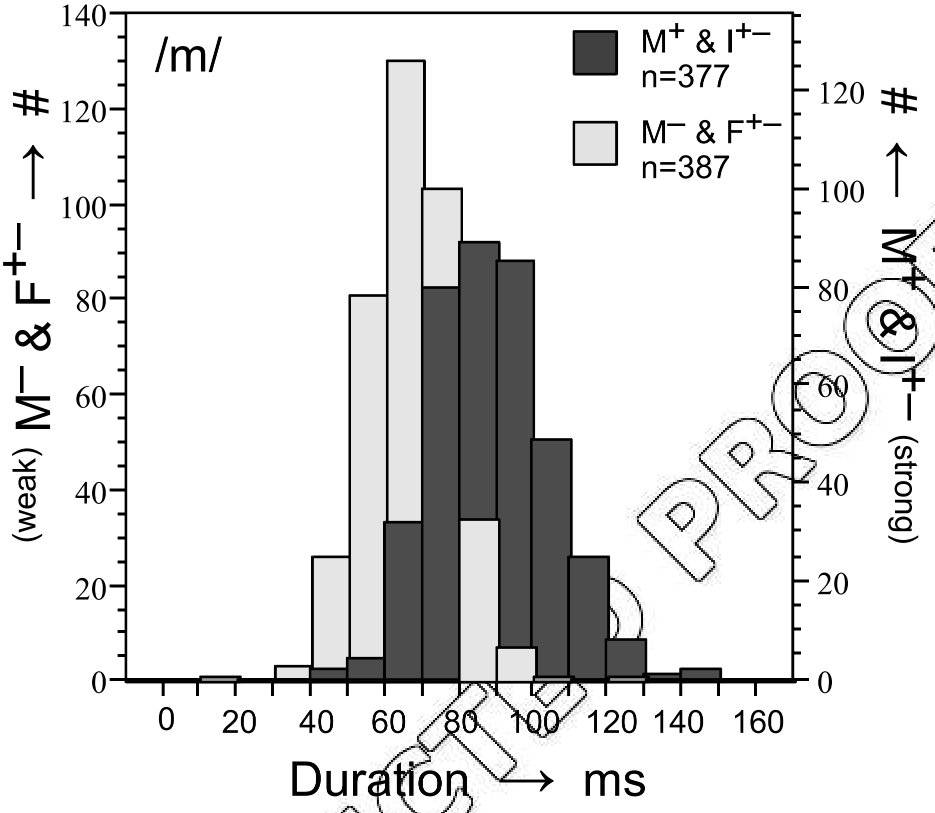

120
(1024)
115
(767)
110
(575)
105
(431)
100
(323)
Center of Gravity ⟶ semitones (Hz)
Stress + —
PostCoronal ▲△ -4.11
/kgŋrʃʒ/
Coronal ■□ -6.48
/ðɵsztdnl/
Labial ●○ 6.49
/fvpbmw/
Initial
Medial
Final
Position in Word

Figure 6.b.


/n/
M⁺ & I⁺⁻
n=303
M⁻ & F⁺⁻
n=782
300
250
200
150
100
50
0
(weak) M⁻ & F⁺⁻ → #
100
80
60
40
20
0
# → M⁺ & I⁺⁻ (strong)
0
20
40
60
80
100
120
140
160
Duration → ms

5

Figure 7.a.


/m/
M+ & I+–
n=377
M– & F+–
n=387
(weak) M– & F+– → #
M+ & I+– → # (strong)
0
20
40
60
80
100
120
140
160
Duration → ms

5

Figure 7.b.

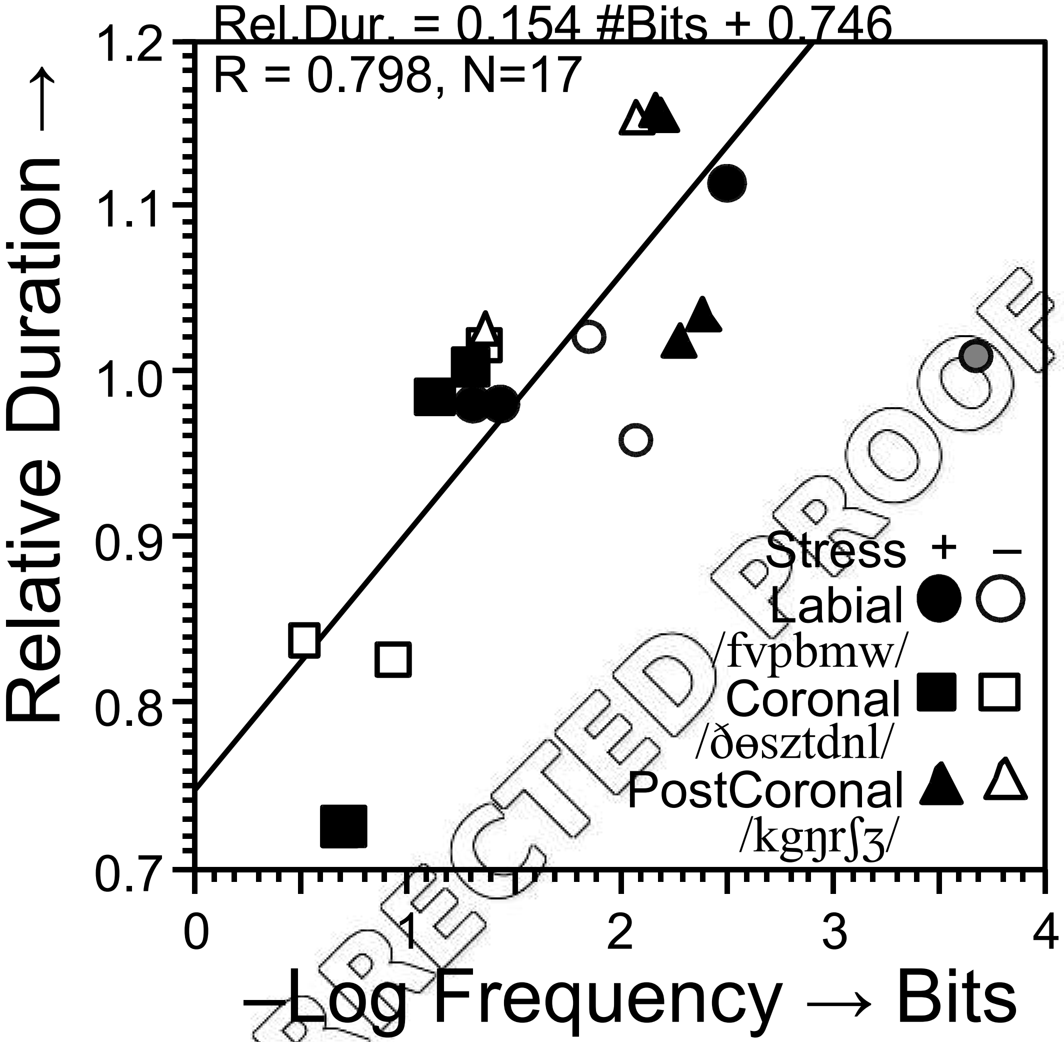

Rel.Dur. = 0.154 #Bits + 0.746
R = 0.798, N=17
Relative Duration →
1.2
1.1
1.0
0.9
0.8
0.7
0
1
2
3
4
–Log Frequency → Bits
Stress + –
Labial ● ○
/fvpbmw/
Coronal ■ □
/ðθsztdnl/
PostCoronal ▲ △
/kgŋrʃʒ/

Figure 8.a.

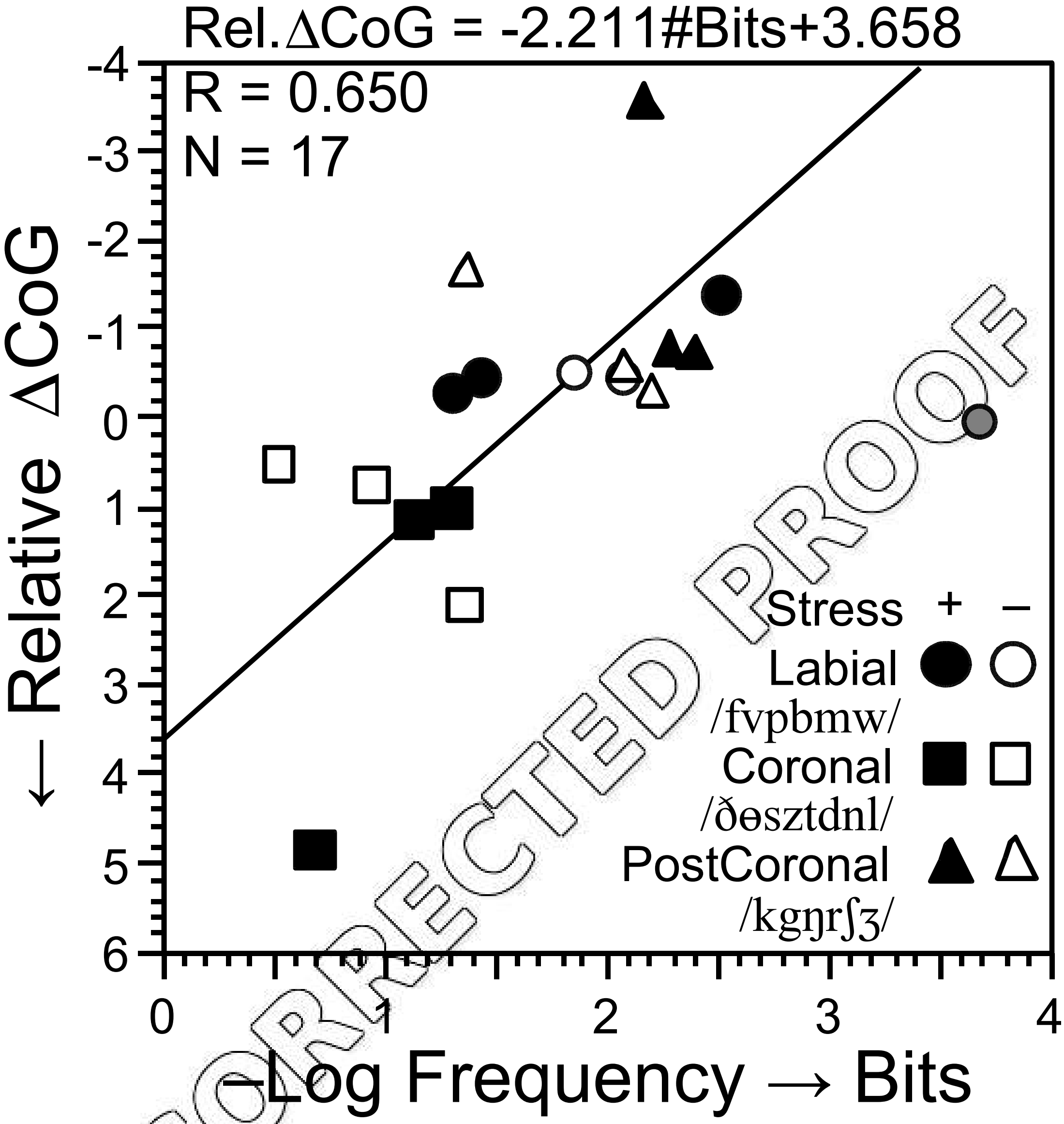

Rel.ΔCoG = -2.211#Bits+3.658
R = 0.650
N = 17
← Relative ΔCoG
-4
-3
-2
-1
0
1
2
3
4
5
6
0
1
2
3
4
–Log Frequency → Bits
Stress + –
Labial ● ○
/fvpbmw/
Coronal ■ □
/ðɵsztdnl/
PostCoronal ▲ △
/kgŋrʃʒ/

Figure 8.b.