EFFECTS OF STRESS AND LEXICAL STRUCTURE ON SPEECH EFFICIENCY

R.J.J.H. van Son and Louis C.W. Pols

Institute of Phonetic Sciences/IFOTT
University of Amsterdam
Herengracht 338, 1016 CG Amsterdam, The Netherlands
tel: +31 20 5252183; fax: +31 20 5252197; email: {rob, pols}@fon.hum.uva.nl

ABSTRACT

It is proposed that some of the variation in speech is the result of an effort to communicate efficiently. Speaking is considered efficient if the speech sound contains *only* the information needed to understand it. This efficiency is tested by means of a corpus of spontaneous and matched read speech, and syllable, word, and N-gram frequencies as measures of information content (1582 intervocalic consonants, and 2540 vowels). It is indeed found that the duration and spectral reduction of consonants and vowels from stressed syllables correlate with syllable and word frequencies, as does consonant intelligibility. Correlations for phonemes from unstressed syllables are generally weaker or absent. N-gram models of word predictability did not correlate with any of the factors investigated. Simple N-grams seem to be a poor model for human word prediction. It is concluded that the principle of *efficient communication* organizes at least some aspects of speech production.

1. INTRODUCTION

A large part of the variation found in speech can be described in terms of in- and decreased *articulatory precision* or *faithfulness* (*hyper*- versus *hypo*-articulation, [9]). It has been known that this variation is often planned and doesn't impede comprehension. The former is evident from research on speaking styles, speech rate, and coarticulation. Speakers have been shown to adapt the level of articulatory faithfulness to the requirements of the speaking task. On the other hand, utterances that show heavy reduction are routinely recognized with high precision, notwithstanding the fact that isolated segments or words from these same utterances show reduced intelligibility. Combined, these two aspects of articulatory variation could indicate that speakers willfully reduce the level of articulatory precision when it doesn't impede comprehension. That is, speech is efficient.

If speakers are efficient, the speech signal will only contain the information needed to understand the message: "speech is the missing information" [9]. The use of the term *efficient* implies a cost/benefit trade-off. We will limit the definition of communicative efficiency in this paper to maximal intelligibility with minimal articulatory "effort". To be able to achieve this efficiency, the speaker must estimate the ease with which the listener can understand her: "speaking for listening" [3]. Different estimates lead to different speaking styles. Ranging from over-articulated word lists to mumbled courtesies.

One aspect of efficiency, the effect of (semantic) predictability on duration and intelligibility, has been the target of previous research ([1],[2],[3],[4],[5],[7],[8],[15]). In the context of the current paper, the results of these studies can be summarized as indicating that on the one hand, listeners tend to identify whole utterances better the more predictable they are. On the other hand, speakers seem to compensate for this by better pronouncing unpredictable words.

The research presented in this paper is intended as a first step to a full quantification of efficiency in connected speech.

2. QUANTIFYING EFFICIENCY

Measures of information content are derived from Bayes' equation:

$$\mathrm{Prob}(e_i|c_i) = \mathrm{Prob}(e_i) \cdot \mathrm{Prob}(c_i|e_i) / \mathrm{Prob}(c_i) \quad (1)$$

In which e_i is a certain speech element, say a word, in a certain context c_i. *Prob(x)* is the probability of encountering *x*. *Prob(e|c)* is the conditional probability measured in missing word or *cloze* tests, i.e., the probability of observing a word (e_i) in a specific context (c_i). The information associated with the presence of a certain entity x is: $I(x) = -\log_2(Prob(x))$ (in bits). Using this we obtain equation 2:

$$I(e_i|c_i) = I(e_i) + I(c_i|e_i) - I(c_i) \quad (2)$$

For example, in the proverb "A stitch in time saves *nine*" the last word "*nine*" can be very reliably predicted from the preceding words [8]. Actually, in his sentence the word "*nine*" itself is hardly informative, $I(nine|A...saves\_) \approx 0$. Speech communication is efficient if the speech signal contains enough information to be identified, and not more. This means that, after accounting for acoustic disturbances and speaking style, each element should contain an amount of information essentially proportional to $I(e_i|c_i)$.

Earlier research has shown that the above holds qualitatively for content words ([1],[2],[3],[4],[5],[6],[8]). Therefore, the application of equation 2 to the pronunciation and intelligibility of words in utterances seems feasible. However, it is unlikely that speakers and listeners process smaller entities, like phonemes in syllables, in the same way as words in an utterance. If we ignore the effects of context, the amount of information needed to identify an element is just the logarithm of the frequency of occurrence ($I(e_i)$ in equation 2). There is evidence that this is an important factor at the level of syllables [15].

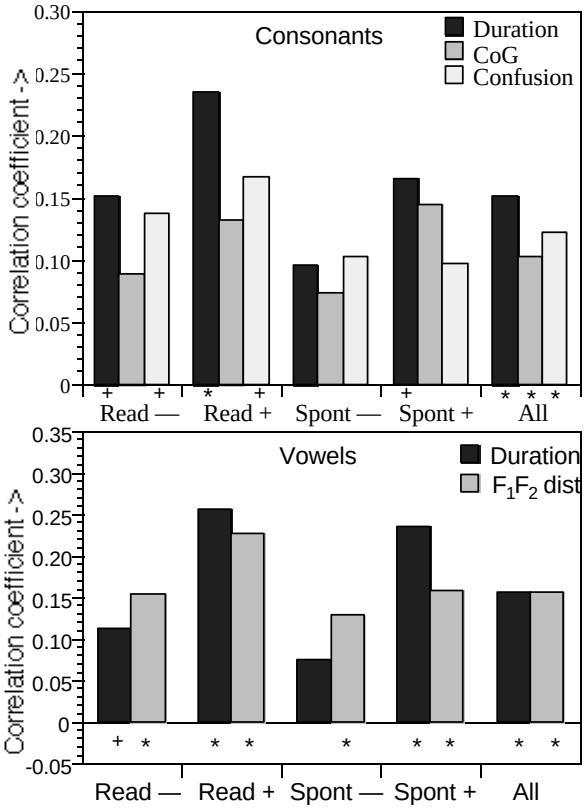

Figure 1: Correlation coefficients between I(syllable) and phoneme Duration, Spectral Center of Gravity (CoG), F_1/F_2 distance, and Confusion of consonant Identification (i.e., H(responses) per token, used with switched signs). The differences between conditions and measures were statistically not significant ($p > 0.01$). Top: Consonants (n=1582, +:308 —:483), bottom: Vowels (n=2540, +:471 —:799). Read: read, Spont: spontaneous speech, +:stressed, —:unstressed syllables, All: combined realizations. +:$p \leq 0.01$, *:$p \leq 0.001$.

It is clear that some kind of language model has to be used to evaluate the predictability of words (w_i) in context, i.e., $Prob(w_i|c_i)$. The language models currently in use for practical work are generally based on N-grams. Given the amount of text needed to determine the frequencies of longer N-grams, a full N-gram model for N>2 is generally not feasible. In this paper we will use partial models with N from 1 to 4, calculated from a relatively small corpus of Dutch newspaper texts published on the WWW.

If speech is indeed organized efficiently, we can predict that speakers adapt their speaking effort to "match" the expected effort needed for recognition. As acoustic measures of the effort and information content of speech, we use *Duration* and two measures of spectral reduction: *Spectral Center of Gravity* (CoG for consonants, i.e., the "mean" frequency in semi-tones, weighted by spectral power) and the F_1/F_2 distance to the center of vowel reduction (300, 1450 Hz for vowels) in semitones. These measures have been shown to be related to speaking effort as used here and intelligibility ([10],[11],[12],[13],[14]). The entropy of the responses to single stimulus tokens is used as a measure of *un*intelligibility, i.e., confusion. This is equivalent to the logarithm of the *perplexity* of the responses and measures the amount of information *missing* from the acoustic signal.

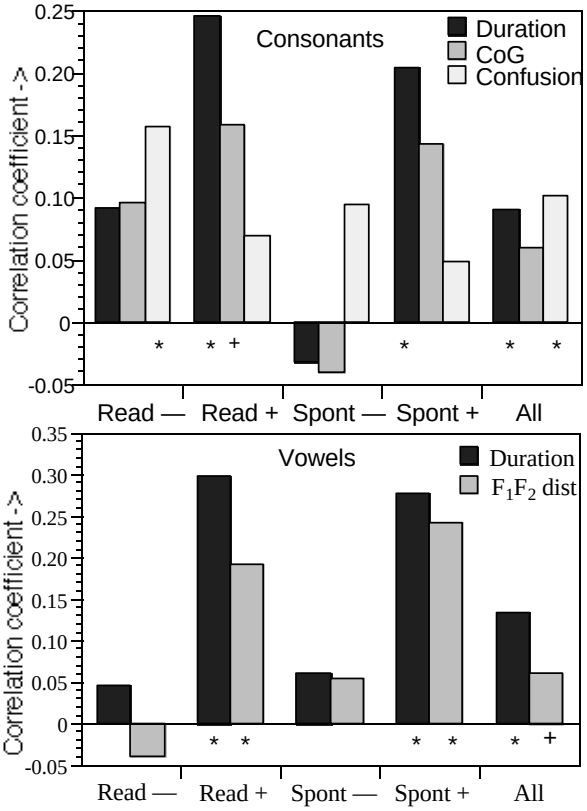

Figure 2: As figure 1 but now using I(word).

3. MATERIALS

For this study we selected recordings of a single male speaker who read aloud a transliteration of spontaneous speech recorded earlier (20 minutes of speech each, in total 12007 syllables and 8046 words). The orthographic script was transcribed to phonetic symbols ([13],[14]). The original transcribed text was used to estimate syllable frequencies (but *not* word-frequencies, contrary to [14]). All Vowel-Consonant-Vowel (VCV) segments were located in the speech recordings (read and spontaneous). 791 VCV pairs that had both realizations originating from corresponding positions in the utterances with identical syllable structure, syllable boundary type, and sentence accent and lexical syllable stress, were selected for this study (see table 1, implying 1270 vowel pairs [13],[14]). Monosyllabic function words are marked as unstressed. Word medial consonants are considered to be syllable initial (maximal onset). The VCV pairs were randomly selected to cover all consonants present and both stress

	Velar	Pal	Alv	Lab	Total
Plos	kg 63	-	td 65	pb 61	189
Fric	X 77	SJ 3	sz 63	fv 75	218
Nasal	N 14	-	n 72	m 63	149
V-like	r 60	j 21	l¬ 94	w 60	235
Total	214	24	294	259	791

Table 1: Dutch consonants used in this paper and the number of matched Read/Spontaneous VCV pairs (ignoring voicing differences). 308 pairs were from syllables carrying lexical syllable stress, 483 from unstressed syllables.

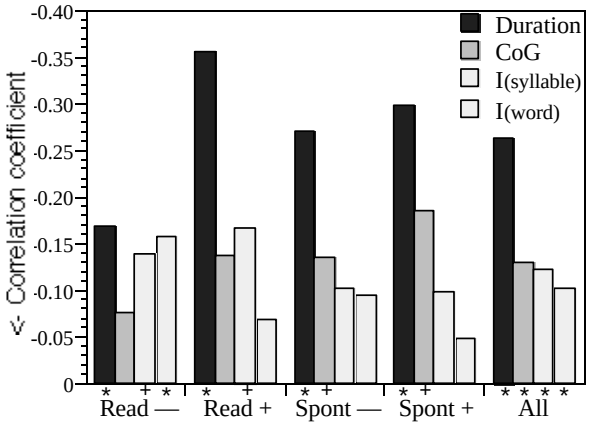

Figure 3: As figure 1 but now correlating the Duration, CoG, I(Syllable) and I(word) with the Confusion of consonants. Note the reversed vertical axis.

conditions (except for /h/, primary lexical syllable stress only). Duration and the extreme CoG frequency of all vowel and consonant realizations were measured ([13],[14]).

For this paper, 22 Dutch subjects, all native speakers of Dutch, were asked to identify these 1582 intervocalic consonant realizations in their original VCV context. The outer 10 ms of the VCV tokens were removed and smoothened with 2 ms Hanning windows to prevent interference from the adjacent consonants and transient clicks. The order of presentation was (pseudo-) random and different for each subject. The subjects had to select the Dutch orthographic symbol on a computer CRT screen that corresponded to the sound heard (this causes no ambiguity in Dutch). For each token, the entropy of the 22 responses was calculated and used as a measure of confusion (H(responses) = log(Perplexity) i.e., the missing information).

Obtaining a reasonable estimates of word- and N-gram frequencies requires large amounts of text. Therefore, we decided to use an separate text corpus to estimate word-frequencies and N-grams. From around 1400 "normalized" (i.e., pre-processed) Dutch newspaper texts collected from the WWW (around 890,000 words), we counted N-gram frequencies for N=1 (word-frequencies) to N=4. We included the transcription of the speech recordings in the corpus to suppress out-of-vocabulary words.

For each word in the transcription, we determined *I(word|context)* for a given N-gram length as the minimum value up to that length. N-grams (N≥2) were limited to those occurring more than once, consisting of words found at least 5 times in the corpus (6 times for N=4). Coverage decreased from 55% for N=2, to 22% for N=3, and only 5% for N=4.

4. RESULTS

To compensate for the large variation in values between our phonemes, we calculated the correlation coefficients after subtracting the individual mean values from each quasi-homogeneous group of phoneme realizations (homogeneous with respect to phoneme identity, speaking style, and syllable stress). The degrees of freedom in the statistical tests were reduced accordingly to compensate for this procedure.

The results are represented in the figures 1-3. Figure 1 shows the correlation between Duration, spectral reduction

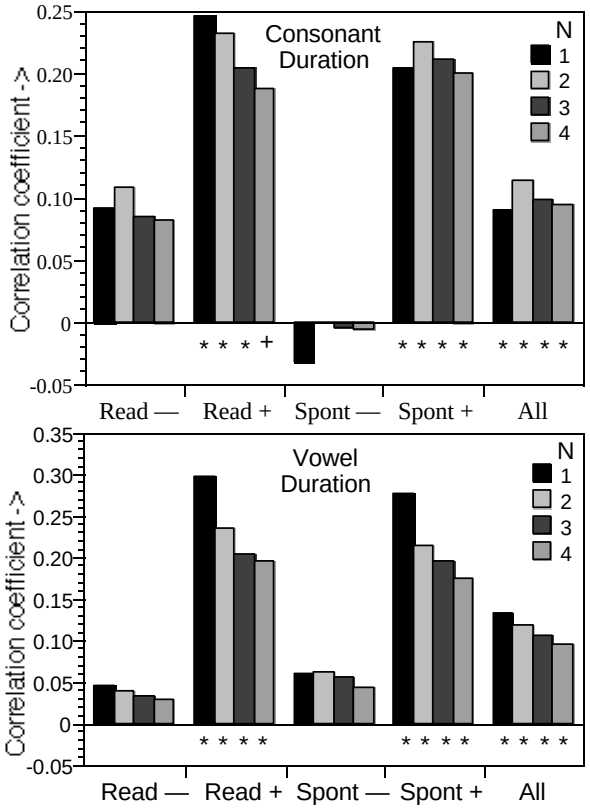

Figure 4: As figure 1 but now correlating the Duration, with I(word|context) for different size N-grams. N=1 equals plain word-frequencies.

(respectively, CoG and F_1/F_2 distance), or the Confusion of our listeners for both consonants and vowels with the negative logarithm of the syllable frequency, *I(syllable)*. Figure 2 shows the results for a correlation with the negative logarithm of the word frequency, *I(word)*. High correlation was largely limited to the stressed syllables ($p \leq 0.01$, R_+vs.R_-). Figure 3 shows the correlation of all other values with the confusion in the listening experiment, i.e., the intelligibility of the consonants. From figure 3 it becomes clear that duration was most strongly linked to intelligibility ($p \leq 0.001$). Figure 4 shows the decreasing effects on the correlation strength between duration and predictability of including context (longer N-grams) in the calculation of *I(word|c_i)*. Figure 5 shows the decreasing correlation between *I(syllable)* and I(*word|c_i)* as a function of the length of the context.

5. DISCUSSION AND CONCLUSIONS

Although the correlation coefficients found in our data are generally statistically significant, they are also quite small ($R^2 < 0.07$). Part of this weakness can be attributed to large measuring errors in determining the relevant parameters. A more important problem is that syllable and word frequencies are only a first step in evaluating predictability. It must be noted that the correlations with *I(word)* taken from the larger corpus were weaker than using word-frequencies from the spoken text itself (not shown, c.f., figures 2-3 with [14]). It is remarkable that including context actually decreased the correlation strength (figure 4). This suggests that N-gram frequencies taken from large corpora might be worse models for expected ease of identification than plain word-frequencies

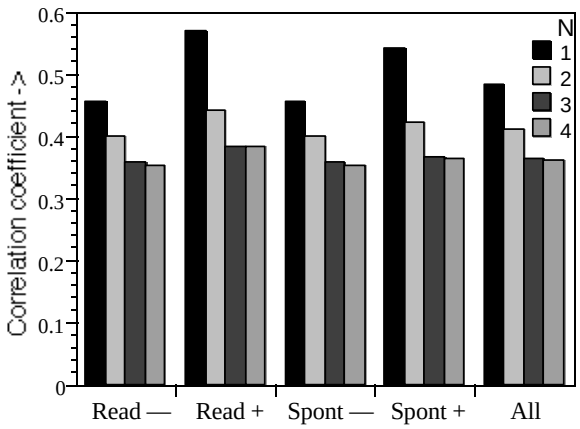

Figure 5: As figure 4 but now correlating I(syllable), with I(word|context) for different size N-grams. All correlations are significant ($p \leq 0.001$)

taken from the transcription itself. Therefore, we think the position of a word in an utterance should be evaluated using more elaborate models including grammar, prosody, and preferably, semantics.

Figures 2 shows that the effects of word frequency, *I(word)*, on acoustic parameters are limited to the stressed syllables. Figure 5 shows that this cannot be completely explained by noting that rare syllables tend to be the stressed parts of rare words and *vice versa* ([4],[16]). For unstressed syllables, which include monosyllabic function words, there are statistically significant correlations between *I(syllable)* and acoustic parameters and confusion but not for *I(word)* (compare figures 1 and 2). The differences between figures 1 and 2 are much larger than the corresponding correlations from figure 5 would suggest. This indicates that the effects of word predictability might be somehow limited to stressed syllables.

To some extent, our results support the idea that the articulatory "content" of individual components of speech correlates with the information needed to identify them. The syllable and word frequencies are correlated with the duration, spectral reduction and intelligibility of individual phonemes. This confirms the correlation between predictability and ease of identification as found in the literature ([4],[6]). However, using N-gram frequencies, we were unable ascertain that speakers actually use the predictability of words *in context*.

Combining our data with those presented in the literature, we can conclude that speakers, at least to some extent, anticipate the efforts needed to understand their message. They adapt some aspects of their speech to strike a balance between their own efforts and those of their audience. This adaptation increases the efficiency of communication.

6. ACKNOWLEDGMENTS

This research was made possible by grants 300-173-029 and 355-75-001 of the Netherlands Organization of Research.

7. REFERENCES

1. Borsky, S., Tuller, B. and Shapiro, L.P. "'How to milk a coat:' The effects of semantic and acoustic information on phoneme categorization". *J. Acoust. Soc. Am.* 103, 2670-2676, 1998.
2. Charles-Luce, J. "Cognitive factors involved in preserving a phonemic contrast", *Language and Speech* 40, 229-248, 1997.
3. Cutler, A. "Speaking for listening", in A. Allport, D. McKay, W. Prinz and E. Scheerer (eds.) *Language perception and production*, London; Academic Press, 23-40, 1987.
4. Cutler, A. "Spoken word recognition and production", in J.L. Miller and P.D. Eimas (eds.) *Speech, Language, and Communication*. *Handbook of Perception and Cognition*, 11, Academic Press, Inc, 97-136, 1995.
5. Fowler, C.A. "Differential shortening of repeated content words in various communicative contexts", *Language and Speech* 31, 307-319, 1988.
6. Hunnicutt, S. "Intelligibility versus redundancy - conditions of dependency", *Language and Speech* 28, 47-56, 1985.
7. Kang, H-S. "Acoustic and intonational correlates of the informational status of referring expressions in Seoul Korean", *Language and Speech* 39, 307-340, 1996.
8. Lieberman, P. "Some effects of semantic and grammatical context on the production and perception of speech", *Language and Speech* 6, 172-187, 1963.
9. Lindblom, B. "Role of articulation in speech perception: Clues from production", *J. Acoust. Soc. Am.* 99, 1683-1692, 1996.
10. Sluyter, A.M.C. and Van Heuven, V.J. "Spectral balance as an acoustic correlate of linguistic stress", *J. Acoust. Soc. Am.* 100, 2471-2485, 1996.
11. Sluyter, A.M.C., Van Heuven, V.J., and Pacilly, J.J.A. "Spectral balance as a cue in the perception of linguistic stress", *J. Acoust. Soc. Am.* 101 (1), 503-513, 1997.
12. Sluyter, A.M.C. *Phonetic correlates of stress and accent*, HIL dissertations 15, PhD thesis, University of Leiden, 1995.
13. Van Son, R.J.J.H. and Pols, L.C.W. "An acoustic description of consonant reduction", *Speech Communication* (in press).
14. Van Son, R.J.J.H., Koopmans-van Beinum, F.J., and Pols, L.C.W. "Efficiency as an organizing factor in natural speech", *Proc. ICSLP'98,* Sydney, Vol 6, 2375-2378, 1998.
15. Vitevitch, M.S., Luce, P.A., Charles-Luce, J., and Kemmerer, D. "Phonotactics and syllable stress: Implications for the processing of spoken nonsense words", *Language and Speech* 50, 47-62, 1997.
16. Zue, V.W. "The use of speech knowledge in automatic speech recognition", *Proc. of the IEEE*, 73 (11), 1602-1615, 1985.