EFFICIENCY AS AN ORGANIZING PRINCIPLE OF NATURAL SPEECH

R.J.J.H. van Son, Florien J. Koopmans-van Beinum, and Louis C.W. Pols

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

ABSTRACT

A large part of the variation in natural speech appears along the dimensions of articulatory precision / perceptual distinctiveness. We propose that this variation 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 and word frequencies as measures of information content (12007 syllables, 8046 word forms, 1582 intervocalic consonants, and 2540 vowels). It is indeed found that the duration and spectral reduction of consonants and vowels correlate with the frequency of syllables and words in this corpus. Consonant intelligibility correlates with both the acoustic factors and the syllable and word frequencies. It is concluded that the principle of *efficient communication* organizes at least some aspects of speech production.

1. INTRODUCTION

Speech can be considered an efficient means of communication. Speakers will not articulate more accurately than they consider necessary to be understood. This means that the speech signal will only contain the information needed to understand the message: "speech is the missing information" [10]. Much of the variation that is normally found in speech can be interpreted as ways to increase the efficiency of communication. Especially variation as a result of speaking style, assimilation, coarticulation and reduction.

The use of the term "efficient" implies a cost/benefit trade-off. The maximal amount of information transmitted at the minimum "cost". To be able to achieve this, the speaker must estimate the ease with which the listener can understand her: "speaking for listening" [4]. 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 [2,3,4,5,6,8,9,17]. 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 utterances better the more predictable they are. On the other hand, speakers seem to compensate for this by better pronouncing unpredictable words.

The actual "level of efficiency" is less important than which factors are used to determine the optimal level of "articulatory faithfulness" and how speakers "manipulate" the speech sound to achieve optimal intelligibility [1,11]. Note that a full investigation of these questions constitutes a major research effort.

In this paper we will start with a demonstration of principle. First, we will develop a quantitative basis for determining the level of efficiency. Then we will present results of an evaluation of the extent to which syllable and word frequencies affect the acoustic realization of consonants and vowels, and the intelligibility of consonants. Finally, we will discuss these results.

2. QUANTIFYING EFFICIENCY

Measures of information content are derived from Bayes' equation:

$$\mathrm{Prob}(e_i,c_i) = \mathrm{Prob}(c_i|e_i) \cdot \mathrm{Prob}(e_i) = \mathrm{Prob}(e_i|c_i) \cdot \mathrm{Prob}(c_i) \qquad (1)$$

In which e_i is a certain speech element, say a word or a phoneme, in a certain context c_i. *Prob(x)* is the probability of encountering x. *Prob(x|y)* is the conditional probability of encountering x if y is present and *Prob(x,y)* is the probability of encountering both x and y together. Using equation 1, we can rewrite the probability of encountering e_i as the product of the probability of finding e_i in a certain context, c_i, and the probability of finding this context:

$$\mathrm{Prob}(e_i) = \mathrm{Prob}(e_i|c_i) \cdot \mathrm{Prob}(c_i) \,/\, \mathrm{Prob}(c_i|e_i) \qquad (2)$$

Prob(e_i|c_i) is the probability measured in missing word or *cloze* tests, i.e., the probability of observing a word in a specific context. 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 3:

$$I(e_i) = I(e_i|c_i) + I(c_i) - I(c_i|e_i) \qquad (3)$$

Averaging equation 3 over all possible elements, e_i, and contexts, c_i in a language, we obtain the conventional average information content:

$$H(e) = H(e|c) + H(c) - H(c|e) = H(e|c) + T(e,c) \qquad (4)$$

In which $H(x) = \Sigma_i\{-Prob(x_i)\cdot log2(Prob(x_i))\}$ summed over all possible x_i, is the average information of x. *H(e|c)* is the logarithm of the perplexity of the language and $T(e,c) = H(e)+H(c)-H(e,c)$ is called the Mutual Information. Equation 4 describes a way to divide the average information, H(e), needed to identify an element e (e.g., a word) into a part carried by the element itself, i.e., *H(e|c)*, and a part carried by the context, i.e., *T(e,c)*.

For example, to identify a word, not all information has to be present in the word itself, part of it can be extracted from the context. An extreme example is the sentence "A stitch in time saves *nine*". The last word "*nine*" can be very reliably predicted from the preceding words [9]. The word itself is hardly informative, $I(nine|A\ldots saves_) \approx 0$.

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 $I(e_i|c_i)$. For content words this has been qualitatively found [2,3,4,5,6,7,9]. Therefore, the application of equation 3 on 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 3). There is evidence that this is an important factor at the level of syllables [17] and we will use this measure in the present study.

If speech is indeed organized efficiently, we can predict that speakers adapt their speaking effort to "match" the expected effort needed for recognition. So we should find a correlation between acoustic measures of effort and information content. The effect of this correlation is acoustic reduction of the phonemes in "predictable" positions in the utterance and a strengthening in "unpredictable" positions. The intelligibility of the isolated phonemes should follow the acoustic reduction and strengthening. 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, weighted by spectral power) and the F_1/F_2 distance to the center of vowel reduction (300, 1450 for vowels) in semitones. These measures have been shown to be related to speaking effort and intelligibility [15,16,12,13,14]. The entropy of the responses to single stimulus tokens was 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.

	Velar	Pal	Alv	Lab	Total
Plos	kg 63	-	td 65	pb 61	189
Fric	χ 77	ʃʒ 3	sz 63	fv 75	218
Nasal	ŋ 14	-	n 72	m 63	149
V-like	r 60	j 21	l ɬ 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.

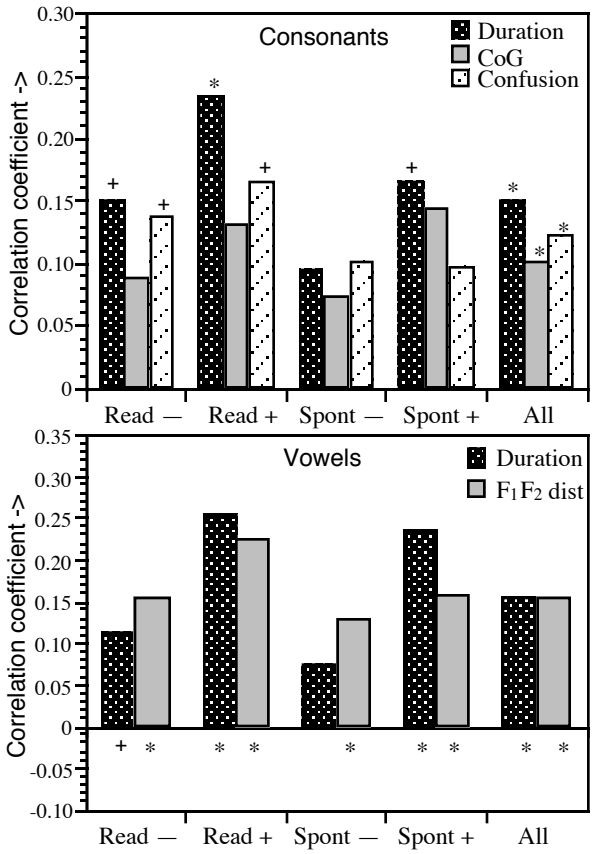

Figure 1: Correlation coefficients between I(syllable) and phoneme Duration, Spectral Center of Gravity (CoG), F_1/F_2 distance, and Confusion of Identification (i.e., H(responses) per token, used with switched signs). The differences between conditions and measures were statistically not significant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

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, 12007 syllables and 8046 word forms). The orthographic script was transcribed to phonetic symbols and each recording was checked against this transcription and marked for sentence accent by one of us [15,16]. The original transcribed text was used to estimate word and syllable frequencies, circumventing the scarcity of data on spontaneous word frequencies. From the phonetic transcription, 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, 1770 distinct vowel pairs, these are

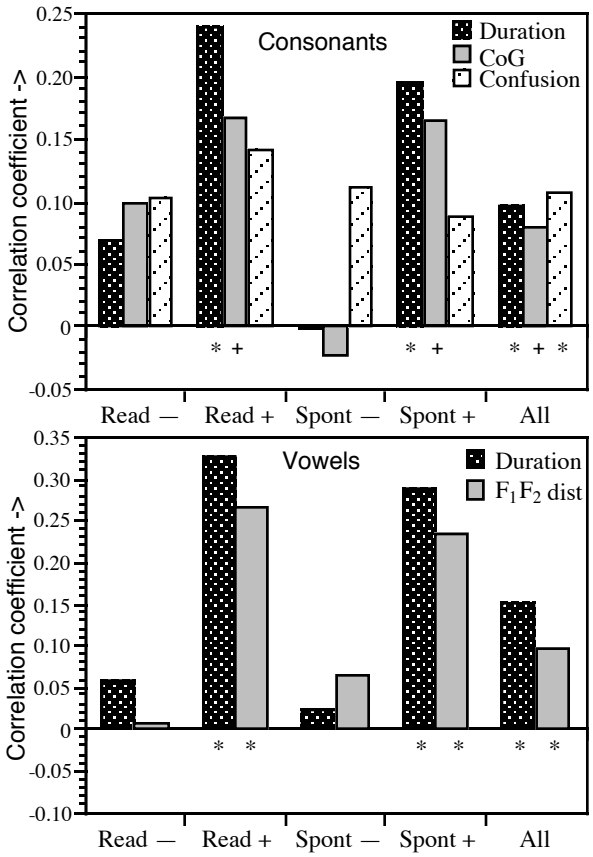

Figure 2: As figure 1 but now using I(word).

the same realizations as used by [15,16]). Monosyllabic function words are marked as unstressed. Word medial consonants are considered to be syllable initial (maximal onset). The VCV pairs were selected to cover all consonants and stress conditions present (except for /h/). The pairs were selected randomly for each individual consonant and stress condition (lexical syllable stress only, 308 pairs from stressed and 483 from unstressed syllables). Duration and the extreme CoG frequency of all vowel and consonant realizations were measured [15,16].

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 (i.e., missing information).

4. RESULTS

$I(phoneme_i|syllable_i)$ nor $I(phoneme_i|word_i)$ (equation 3) was correlated with phoneme duration, CoG, nor with intelligibility (not shown). Therefore, we will use the

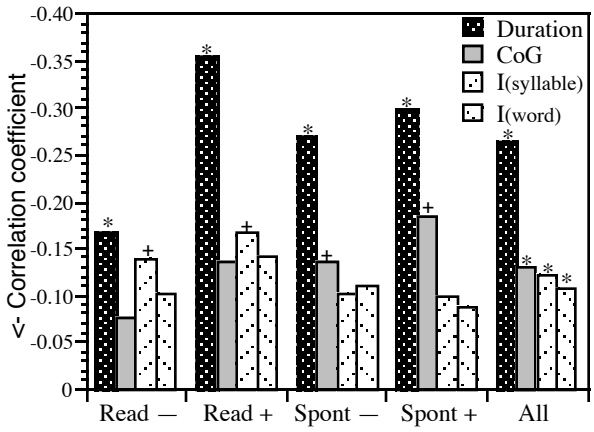

Figure 3: As figure 1 but now correlating the Duration, CoG, I(syllable) and I(word) with the Confusion of consonants. Note the reversed vertical axis.

frequency related measures of $I(syllable_i)$ and $I(word_i)$ as the parameters of interest in the remainder of this paper

To compensate for the large variation in intrinsic 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, but ignoring voicing). 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 the negative logarithm of the syllable frequency, I(syllable), and Duration, spectral reduction (respectively, CoG and F_1/F_2 distance), and the Confusion of our listeners for both consonants and vowels (Confusion for consonants only). Figure 2 shows the results for a correlation with the negative logarithm of the word frequency. This 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. From figure 3 it becomes clear that duration was most strongly linked with intelligibility ($p \leq 0.001$).

5. DISCUSSION AND CONCLUSIONS

Although the correlation coefficients found in our data are generally statistically significant, they are also quite small ($R^2 < 0.07$). There are several explanations for this weak correlation. First of all, there were large errors in determining syllable and word frequencies from such a small corpus and determining intelligibility using only 22 listeners. Measuring phoneme durations in connected (spontaneous) speech too has a high level of errors. Together with the small number of realizations, we expected a large level of "noise" in our data. More important, the frequency of occurrence is only a first step in evaluating predictability. To really express the importance of a word, its position in the utterance should be evaluated using models of grammar, prosody, and preferably, semantics.

On the whole, our results support the idea that the individual components of speech contain only the information needed to identify them. When the predictability of syllables and words is expressed in terms of information content, it correlated with duration, spectral reduction and intelligibility of individual phonemes. This correlation was found *after* normalization for the effects of phoneme identity, speaking style differences, and lexical stress. We know that predictability is strongly correlated with ease of identification [5,7]. Therefore, we can conclude that speakers anticipate the efforts listeners have to expend in recognition and try to strike a balance with their own efforts. The result can be seen as *efficient communication*.

Figure 2 shows that the effects of word frequency (*I(word)*) are limited to the stressed syllables. This can be partly explained by noting that rare syllables tend to occur as the stressed syllables of rare words, and unstressed syllables are generally high-frequency syllables [5,18]. This indicates that syllable frequencies dominate any effect of word frequency.

On a more detailed level, it was found that phoneme duration was the factor most strongly related to both information content and intelligibility. Speakers seem to smooth the distribution of information over time by varying segmental (or syllabic) durations. The weaker correlations of both spectral reduction and intelligibility with information content and with each other suggests that *time* is the limiting factor in speech [16]. Speakers seem to detest spending time speaking while listeners need some time to understand it [1], but also should not be bored [4]. If listeners do not have enough time to identify all aspects of a phoneme, it doesn't "pay" to articulate them properly. Hence, the spectral reduction found in predictable syllables and words could be a secondary result of the shorter durations.

Combining our data with those presented in the literature, we can conclude that speakers 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. REFERENCES

1. Boersma, P.B. *Functional Phonology, formalizing the interactions between articulatory and perceptual drives*, Ph.D. thesis University of Amsterdam, 1998.

2. 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.

3. Charles-Luce, J. "Cognitive factors involved in preserving a phonemic contrast", *Language and Speech* 40, 229-248, 1997.

4. 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.

5. 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.

6. Fowler, C.A. "Differential shortening of repeated content words in various communicative contexts", *Language and Speech* 31, 307-319, 1988.

7. Hunnicutt, S. "Intelligibility versus redundancy - conditions of dependency", *Language and Speech* 28, 47-56, 1985.

8. Kang, H-S. "Acoustic and intonational correlates of the informational status of referring expressions in Seoul Korean", *Language and Speech* 39, 307-340, 1996.

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

10. Lindblom, B. "Role of articulation in speech perception: Clues from production", *J. Acoust. Soc. Am.* 99, 1683-1692, 1996.

11. Schwartz, J.-L., Boë, L.-J., Vallée, N., and Abrt, C. "The dispersion-focalization theory of vowel systems", *J. Phonetics* 25, 255-286, 1997.

12. Sluyter, A.M.C., en Van Heuven, V.J. "Spectral balance as an acoustic correlate of linguistic stress", *J. Acoust. Soc. Am.* 100, 2471-2485, 1996.

13. Sluyter, A.M.C., Van Heuven, V.J., en Pacilly, J.J.A. "Spectral balance as a cue in the perception of linguistic stress", *J. Acoust. Soc. Am.* 101 (1), 503-513, 1997.

14. Sluyter, A.M.C. *Phonetic correlates of stress and accent*, HIL dissertations 15, PhD thesis, University of Leiden, 1995.

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

16. Van Son, R.J.J.H., and Pols, L.C.W. "The correlation between consonant identification and the amount of acoustic consonant reduction", *Proc. Eurospeech'97*, Rhodos, 2135-2138, 1997.

17. 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.

18. Zue, V.W. "The use of speech knowledge in automatic speech recognition", *Proc. of the IEEE*, 73 (11), 1602-1615, 1985.