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_*) ≈ 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].