An Acoustic Model of Communicative Efficiency in Consonants and Vowels taking into Account Context Distinctiveness

R.J.J.H. van Son and Louis C.W. Pols
University of Amsterdam, Institute of Phonetic Sciences/ACLC
Herengracht 338, 1016 CG Amsterdam, The Netherlands
Email: Rob.van.Son@hum.uva.nl, Louis.Pols@hum.uva.nl

ABSTRACT

Speaking is generally considered efficient in that less effort is spent articulating more redundant items. With efficient speech production, less reduction is expected in the pronunciation of phonemes that are more important (distinctive) for word identification. The importance of a single phoneme in word recognition can be quantified as the information (in bits) it adds to the preceding word onset to narrow down the context corrected lexical search. In our study, segmental information showed to correlate consistently with both duration and spectral reduction in vowels and most consonants. No such correlations were found for stops and only little for nasals. This correlation was found after accounting for speaker and vowel identity, speaking style, lexical stress, modeled prominence, position in the syllable, and position of the phoneme in the word. We conclude that speech is organized for efficiency at the level of the phoneme.

1. INTRODUCTION

Speech can be seen as an efficient communication channel: less speaking effort is spent on redundant than on informative items. Studies showed that listeners identify redundant tokens better and that speakers take advantage of this by reducing predictable items [1][2][3][4][5][8][9][16][19][21]. For example, *nine* is pronounced more reduced in the proverb *A stitch in time saves nine* than in *The next number is nine* [9].

Tractable forms of predictability are frequency of occurrence of words and N-gram language models [12]. However, word-frequency effects are partly based on features of the mental lexicon [4][5][25]. Therefore, "frequency" and "language" effects can best be studied separately. Still, word-frequencies are affected by the context [24]. As a first step, this study will be limited to the average context related frequency of words [24].

One way speakers can enhance efficiency is by manipulating the prosodic structure of the utterance. Whether there is an effect of word frequency in addition to these prosodic enhancements is the question we study in this paper.

Theories of word recognition emphasize that word recognition is an incremental task that works on a phoneme by phoneme basis [11]. Often, words are recognized on their first syllable(s) well before all phonemes have been processed [7]. In English and Dutch this is reflected in the fact that lexical stress is predominantly on the first syllable of a word [6][7]. We use a model of word recognition with competition based on an incremental match of incoming phonemes in the mental lexicon [11]. However, words are also primed by their context [24][25]. We will model this priming as an increase in apparent frequency.

Word recognition is a incremental, task [11][18]. Therefore, we will use a measure of the position-dependent segmental contribution in distinguishing words given the preceding word-onset. The lexical information I_L (in bits) of a segment ***s*** preceded by a segment sequence [*word-onset***]** is [28]:

$$I_L = -\log_2 \left(\frac{Frequency\,([\,word-onset\,]+s)}{Frequency\,([\,word-onset\,]+any\ segment)} \right)$$

Frequencies are calculated from a CELEX word-count list of Dutch, based on 39 million words. The word *frequencies* were estimated using a Katz smoothing on counts from 1-5 and an extrapolation based on Zipf's law [26]. The above equation does not account for the predictability of the word due to its distributional (contextual) properties [24][25]. It is possible to determine the average predictability of the word actually spoken in its proper context. To make this calculation tractable, we assume that only the frequency of the word actually spoken is affected by the context. Words tend to occur in certain contexts more than in others (*good idea* vs. *green idea*). This means that the frequencies of words in the neighborhood of the target word will be different from the global frequencies. This difference can be quantified as the Kullback-Leibler distance between the distribution in the context and the global distribution [24]. The resulting value is called the *Context Distinctiveness* of the word (CD(*w*)) and has a value between 0 and the $\log_2$ of the global frequency of the target word [24]. In formula:

$$CD(w) = \sum_{vocabulary} P(c_i|w) \log_2 \left(\frac{P(c_i|w)}{P(c_i)} \right)$$

Where $P(c_i)$ is the plain probability of the word c_i and $P(c_i|w)$ the conditional probability of c_i appearing in the context of *w*. On average, the relative frequency, $CF_{CGN}(w)$, of the target word *w* in the CGN is a factor $2^{CD(w)}$ higher in its normal context than in the corpus as a whole, i.e., $CF_{CGN}(w) = RelFreq_{CGN}(w) \cdot 2^{CD(w)}$. To calculate the segmental information, the formula for I_L is changed to include a correction on the frequency of the target

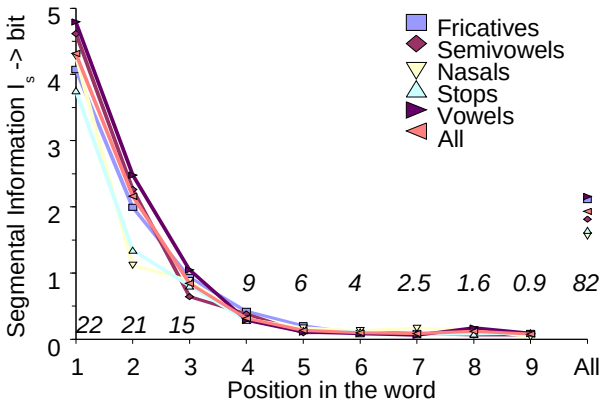


Figure 1. Relation between average segmental information (I_s) and the position in the word grouped by manner of articulation. The number of segments (x1000) for each position is indicated with italic numbers.

word $D(w) = CF_{CGN}(w) \cdot N_{CELEX} - Freq_{CELEX}(w)$. We determined $CF_{CGN}(w)$ from the 5th release of the Spoken Dutch Corpus. This way, the calculation of the CD can be done on the smaller CGN corpus and the global word frequencies can be determined on the comprehensive CELEX word list. The segmental information becomes:

$$I_s = \log_2 \left(\frac{Frequency([word-onset]+s)+D(w)}{Frequency([word-onset]+any\ segment)+D(w)} \right)$$

Where *s* is the current segment, *word-onset* the preceding phonemes in the word, and *w* the actually spoken word-form.

2. METHODS

For this study we used the IFAcorpus [20] which contains 5½ hours (50 kWord) of hand-aligned phonemically segmented speech from 8 native speakers of Dutch, 4 female and 4 male. We used 5 of the 8 speaking styles: informal face-to-face story-telling (I), retold stories (R), read text (T), read isolated sentences (S), and read semantically unpredictable pseudo-sentences (PS, e.g., *the village cooked of birds*). The IFAcorpus can be found at http://www.fon.hum.uva.nl/IFAcorpus.

Acoustic reduction is measured on duration and on the spectral Center of Gravity (CoG) [17]. For vowels we also use the position in vowel formant space. The values of the F_1 and F_2 (in semitones) were combined as the distance to a virtual target of reduction, determined for each speaker seperately as a point with an F_1 midway between the /i/ and the /u/ and an F_2 of the /a/, measured in citation speech. Reduction of a vowel results in a shorter distance to this virtual point in vowel space.

Distinctiveness (CD) was calculated over the 5th release of the Spoken Dutch Corpus (CGN), a total of 1.8 million words [27], over a window of 10 words (5 before and 5 after the target word [24]). The Context Distinctiveness increased more or less linear with the logarithm of the word frequency (R = 0.7). This was used to estimate the CD for words not in the CGN by extrapolation as $CD(w) = 2 \cdot -log_2(P(w)) - 26$ when *w* was not seen.

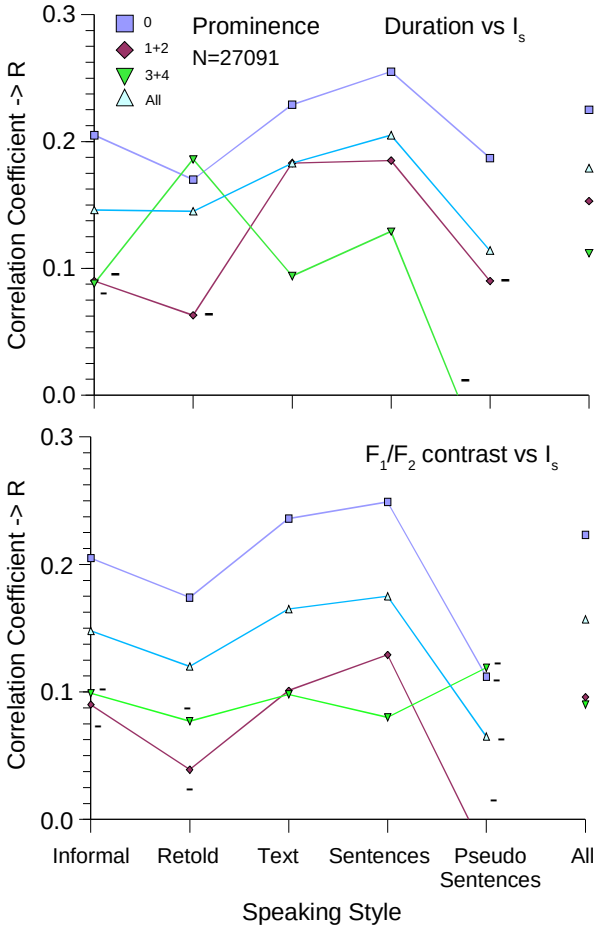
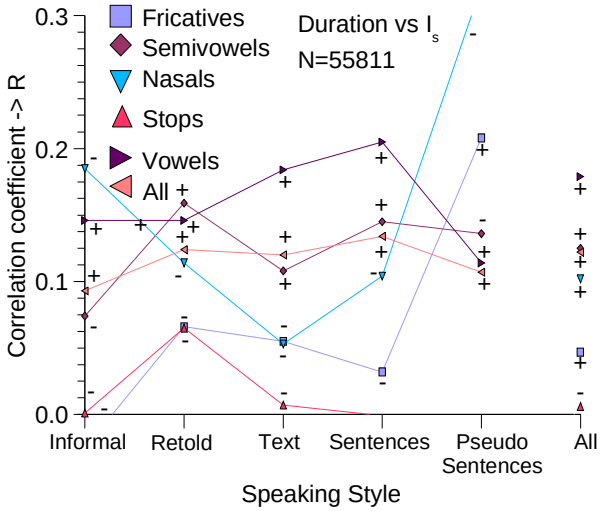


Figure 2. Correlation coefficients for vowel tokens between segmental information and duration (top) and F_1/F_2 contrast (bottom). Plotted is a breakdown on style and prominence marks. Speaker, lexical stress, vowel identity, and type of text are accounted for (see text). Excluded are schwa vowels and vowel tokens with $I_s < 1.5$ bits. All: $p < 0.001$, except those marked -: not significant.

The procedure is illustrated with an example: The segmental information, I_s, of the vowel /o/ in the Dutch word /bom/ ('boom', English 'tree'). The word 'boom' has a smoothed relative CGN frequency of $5.05 \cdot 10^{-5}$. The Context Distinctiveness of 'boom' in the CGN is $CD('boom') = 4.53$ which corresponds to an increase in relative frequency in context by a factor of $2^{CD(w)} = 2^{4.53} = 23$ to $23 \cdot 5.05 \cdot 10^{-5} = 1.2 \cdot 10^{-3}$. This corresponds to a context-corrected count of *45,402* for the word 'boom' instead of the original smoothed CELEX count of 2,226. The correction term becomes $D('boom') = 45{,}402 - 2{,}226 = 43176$. The sum of the (smoothed) counts of all 1172 word-entries in CELEX starting with /bo/ is $Frequency(/bo/) = 67710$, and for the 26186 word-entries starting with /b./ it is $Frequency(/b./) = 1544483$. For the vowel /o/ in 'boom' ('tree'), $I_L = -log_2(67710/1544483) = 4.51$ and $I_s = -log_2([67710+43176]/[1544483+43176]) = 3.84$. That is, context reduces lexical uncertainty.

Word realizations can differ from the lexical norm. The position of the *realized* phoneme in the normative *lexical* transcription is determined using Dynamic Programming.



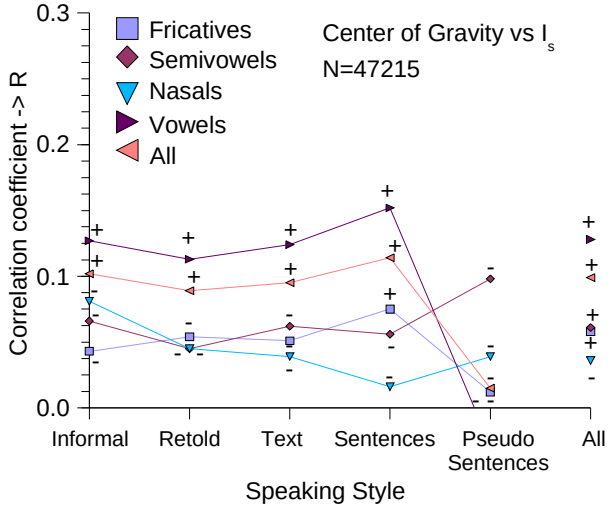


Figure 3. Correlation coefficients between segmental information and duration (left) and Center of Gravity (right). Plotted is a breakdown on style and manner of pronunciation. Speaker and phoneme identity, prominence, lexical stress, and position in the syllable are accounted for (see text). +: p < 0.001, -: not significant.

The lexical normative transcription of the word-onset and phoneme identity are used to search the lexicon.

To cope with the factors that affect acoustic measure, the data are divided into quasi-uniform subsets. Each subset contains all observations that are uniform with respect to all relevant factors. Correlations are calculated after normalizing the values to zero mean value and unit standard deviation (i.e., mean=0, SD=1) within each quasi-uniform subset. The degrees of freedom are reduced by 2 for each subset to account for the normalization. In all analyses, we account for speaker and phoneme identity, speaking style, text type (fixed story or a speaker' sown words), lexical stress, automatically determined prominence, and position in the syllable (onset, kernel, coda). After applying a Bonferroni correction, a level of significance of $p < 0.001$ was chosen. Prominence is assigned automatically by rules from text input based on POS tags [13][14][28]. Function words receive 0, content words 1-4 marks. Prominence marks were combined and words were divided into three classes based on the prominence marks: 0, 1-2, and 3-4. Rule-based prominence marks correlated well with human transcribers (Cohen' s Kappa = 0.62) [13][14].

3. RESULTS

Figure 1 shows the distribution of segmental information over words for the different phoneme classes. We see the expected (sharp) decrease in segmental information value with increasing length of the word-onset caused by the incremental word recognition model used [28]. There seem to be no fundamental differences between the different phoneme classes. The fact that all classes contribute equally to word recognition is itself a form of efficiency.

The schwa is a completely assimilated vowel [15]. Therefore, we excluded the schwa and used only full vowels. The consonants /s n t/ are in many respects the consonantal counterparts of the schwa as the most reduced consonants of their class in Dutch. Therefore, we excluded these consonants too. We excluded all segments with a segmental information below 1.5 bits as our earlier study had shown that there is a floor in the effect of segmental information [28].

Figure 2 displays the correlation between segmental information and vowel duration and formant contrast grouped on prominence. It is obvious that there is a consistent, and statistically significant, correlation between vowel reduction and segmental information. The analysis was repeated for *all* phonemes (excluding /@ s n t/) for both duration and spectral Center of Gravity (CoG, sign reversed for semivowels and nasals). Figure 3 presents the results separated on speaking style and manner of articulation. The results are largely the same as for vowels alone. However, there is a lot more "noise" in the data and not all results are statistically significant. For *stops*, no CoG could be calculated. No relation between duration and I_s could be found for *stops*. For the *nasals*, only for the duration there was a statistically significant correlation with I_s (All category).

4. DISCUSSION AND CONCLUSIONS

Figure 1 clearly shows the importance of "early" phonemes. Dutch (and English) increase recognition efficiency by a prevalence for word-initial lexical stress [6] (73% of word-forms in the IFAcorpus [28]). The strong correlation between position in the word and I_s prevents us from separating these two. A repeated analysis revealed a statistically significant correlation between vowel duration and I_s after accounting for position in the word (positions 1-3, not shown) .

Figures 2 and 3 show that segmental redundancy correlates consistently with acoustic reduction in a wide range of phoneme classes and conditions, both for duration and spectral reduction. However, the correlation coefficients are small and only partially explain the

variance. This is hardly surprising as on one hand, we have "removed" the most important conventional factors that implement efficiency: Prosody and Syllable structure. Furthermore, most of these factors were determined automatically, introducing a lot of errors. The segmentation of the phonemes has its own errors which affects the reduction measures. All these errors induce "noise" which reduces the correlations. In addition, earlier studies have shown that consonant reduction is considerably more difficult to measure than vowel reduction [17]. Together, these factors make using a large corpus necessary to pick up the correlation from the noise.

To summarize, we do find a consistent correlation between the distinctive (information) importance of a phoneme for (incremental) word identification and its acoustic reduction in terms of duration and spectral contrast. This correlation is found after accounting for speaker and vowel identity, speaking style, lexical stress, (modeled) prominence, and position of the phoneme in the syllable. We even found this correlation after accounting for the position of the phoneme in the word (not shown). However, data-sparsity prevented us from analysing this further. We conclude that speech is structured efficiently, even after accounting for the effects of prosodic structure and predictability in average context.

ACKNOWLEDGMENTS

We thank David Weenink for his implementation of the Dynamic Programming algorithm. This research was made possible by grant and 355-75-001 of the Netherlands Organization of Research. The IFAcorpus is licensed under the GNU GPL by the Dutch Language Union.

REFERENCES

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

[2] Boersma, P.B. Functional Phonology, formalizing the interactions between articulatory and perceptual drives, Ph.D. thesis University of Amsterdam, 493 pp, 1998.

[3] 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.

[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] Cutler, A. and Carter, D.M.. "The predominance of strong initial syllables in English vocabulary", Computer Speech and Language 2, 133-142, 1987.

[7] Cutler A.. "The comparative perspective on spoken-language processing", Speech Communication 21, 3-15, 1997.

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

[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] Norris D., McQueen J.M., and Cutler A.. "Merging information in speech recognition: Feedback is never necessary". Behavioral and Brain Sciences 23, 299-325, 2000.

[12] Owens, M., O' Boyle P., McMahon, J., Ming, J. and Smith, F.J. "A comparison of human and statistical language model performance using missing-word tests", Language and Speech 40, 377-389, 1997.

[13] Streefkerk, B.M., Pols, L.C.W., and ten Bosch, L.F.M.. "Acoustical and lexical/syntactic features to predict prominence", Proceedings of the Institute of Phonetic Sciences, University of Amsterdam 24, 155-165, 2001.

[14] Streefkerk, B.M. "Prominence. Acoustical and lexical/syntactic correlates", Ph.D. Thesis University of Amsterdam (In Press).

[15] Van Bergem, D.R. 1993. "Acoustic vowel reduction as a function of sentence accent, word stress, and word class". Speech Communication 12, 1-23, 1993.

[16] 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, 2375-2378, 1998.

[17] Van Son, R.J.J.H. and Pols, L.C.W. "An acoustic description of consonant reduction", Speech Communication 28, 125-140, 1999.

[18] Van Son, R.J.J.H. and Pols, L.C.W. "Perisegmental speech improves consonant and vowel identification", Speech Communication 29, 1-22, 1999.

[19] Van Son, R.J.J.H. and Pols, L.C.W.. "Effects of stress and lexical structure on speech efficiency" Proc. EUROSPEECH' 99, Budapest, 439-442, 1999.

[20] Van Son, R.J.J.H., Binnenpoorte, D., van den Heuvel, H. and Pols, L.C.W.. "The IFA corpus: a phonemically segmented Dutch Open Source speech database", Proc. EUROSPEECH 2001, Aalborg, Denmark, Vol. 3, 2051-2054, 2001.

[21] 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.

[22] Whiteside, S.P. and Varley, R.A. "Verbo-motor priming in the phonetic encoding of real and non-words", Proc. EUROSPEECH' 99, Budapest, 1919-1922, 1999.

[23] Zue, V.W. "The use of speech knowledge in automatic speech recognition", Proc. IEEE 73, 1602-1616, 1985.

[24] McDonald, S.C. and Shillcock, R.C. (2001). "Rethinking the word frequency effect: The neglected role of distributional information in lexical processing", Language and Speech 44, 295-323.

[25] Ferrer i Cancho, R. and Solé R.V. (2001). "The small world of human language", Proceedings of the Royal Society of London B 268, 2261-2265.

[26] Ferrer i Cancho, R. and Solé R.V. (2003). "Least effort and the origins of scaling in human language", PNAS 100, 788-791.

[27] Oostdijk, N. (2000). "The Spoken Dutch Corpus, overview and first evaluation", Proceedings of LREC-2000, Athens, Vol. 2, 887-894.

[28] Van Son, R.J.J.H. & Pols, L.C.W. (2002). "Evidence for Efficiency in vowel production", Proceedings of ICSLP2002, Denver, USA, .