EVIDENCE FOR EFFICIENCY IN VOWEL PRODUCTION

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
Rob.van.Son@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 lexical search. In our study, segmental information showed to correlate consistently with two measures of reduction: vowel duration and formant reduction. This correlation was found after accounting for speaker and vowel identity, speaking style, lexical stress, modeled prominence, and position of the syllable in the word. However, consistent correlations are only found in high−frequency words. Furthermore, the correlation is strongest in normal reading and weaker in spontaneous and anomalous read speech. Combined, these facts suggest that this type of efficiency in production might rely on retrieving stored words from memory. Efficiency in vowel production seems to be less or absent when words have to be assembled on−line.

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].

It is very difficult to quantify redundancy in normal texts or utterances. 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]. Therefore, "frequency" and "language" effects can best be studied separately. As a first step, this study will be limited to effects that are related to the lexical frequency of words.

One way speakers can enhance efficiency is by manipulating the prosodic structure of the utterance. It has long been known that speakers will place important and unpredictable words in focus. Such words tend to get a sentence (pitch) accent and are emphasized considerably [1][10]. Furthermore, in languages that have lexical stress (e.g., English and Dutch), the stressed syllable tends to be the most informative, i.e., unpredictable, of the word [23]. Whether there is an effect of lexical frequency in addition to these prosodic enhancements is the question we study in this paper.

An important question is to what extent the way articulation is organized and controlled influences the efficiency of speech [1]. There are suggestions that "assembling" the articulation of words is a time limiting process in speech [22]. Words whose articulations are retrieved from memory are pronounced faster, and most likely, more reduced than words that have to be assembled "from scratch". This could result in differences in the trade−off's that underly efficient speaking.

Attention has generally been directed at the word−level. However, theories of word recognition emphasize that it is an on−going task that works on a phoneme by phoneme basis, possibly without a need for feed−back (top−down processes) [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].

If speech is efficient at the segmental level, we can expect that speakers will emphasize informative segments and de−emphasize redundant segments. De−emphasis is generally considered to result in acoustic reduction. So speech efficiency at the segmental level can be studied by correlating the importance of a segment with the level of acoustical reduction.

The effects of acoustic reduction on consonants are less well studied than that of vowels and is also more intricate [17]. Furthermore, phonotactic constraints limit the occurrences of specific consonants and link them to the position in the word. These complications and the limited size of the IFAcorpus (50,000 words) would be prohibitive. Therefore, we decided to limit the current study to vowels and leave consonants to a follow−up study. As the schwa is a completely assimilated vowel [15], acoustic reduction is not defined for the schwa. Therefore, we excluded the schwa and used only full vowels. As unstressed syllables in Dutch are generally pronounced with full vowels this still leaves us with enough vowel tokens to do this study.

2. METHODS

2.1 Speech Material and reduction

For this study we used the IFAcorpus [20] which contains 5 1/2 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 www.fon.hum.uva.nl/IFAcorpus.

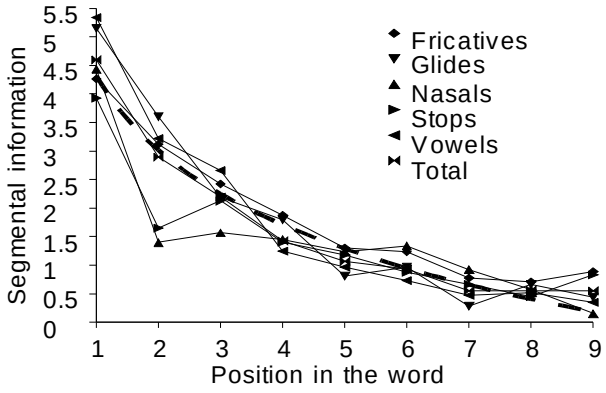


Figure 1. Relation between segmental information and the position in the word grouped by manner of articulation for comparison. The pooled values (Total) have been fitted with a logarithmic line (dashed line).

Acoustic reduction is measured on duration and on 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, fixed for all speakers: midway between the /i/ and the /u/ (F_1 = 350 Hz, F_2 = 1450 Hz). Reduction of a vowel results in a shorter duration and a shorter distance to this virtual point in vowel space. Although the *real* target of reduction will differ between speakers, it will not be far from the chosen point and we will compare only changes in this distance.

2.2 Segmental Importance

The importance of an individual segment in a word is generally calculated by comparing the word frequency with the combined lexical frequencies of all words that differ only at the position of the segment studied. This is a kind of lexical *cloze* test. This definition does not take into account that word recognition is a serial, left−to−right, task [11][18]. Therefore, we will use a measure of the position−dependent segmental contribution in distinguishing words given the preceding word−onset. Assuming perfect recognition of the preceding word−onset. Formally, the segmental information I_s (in bits) of a segment ***s*** preceded by a segment sequence [*word−onset*] is:

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

Frequencies are calculated from a CELEX word−count list of Dutch, based on 30 million words, combined with the relevant word−counts of the IFAcorpus to prevent out−of−vocabulary words. The partial−word *frequencies* were estimated from the matching word−*counts* by summing (*wordcount+1*) instead of the plain word counts themselves. This is a better frequency estimate for low−frequency words.

Word frequencies are calculated by matching the phonemic representation of the spoken words with the normative transcription in the lexicon. However, for numerous reasons, the actual transcriptions of the spoken words in the corpus often do not match the normative lexical transcriptions in the lexicon. Therefore, the position of the realized phoneme in the lexical transcription is determined using Dynamic Programming (i.e., mapping the transcribed phonemes onto the normative transcription). Then, instead of the realized transcription, the lexical normative transcription of the word−onset and phoneme identity are used to search the lexicon.

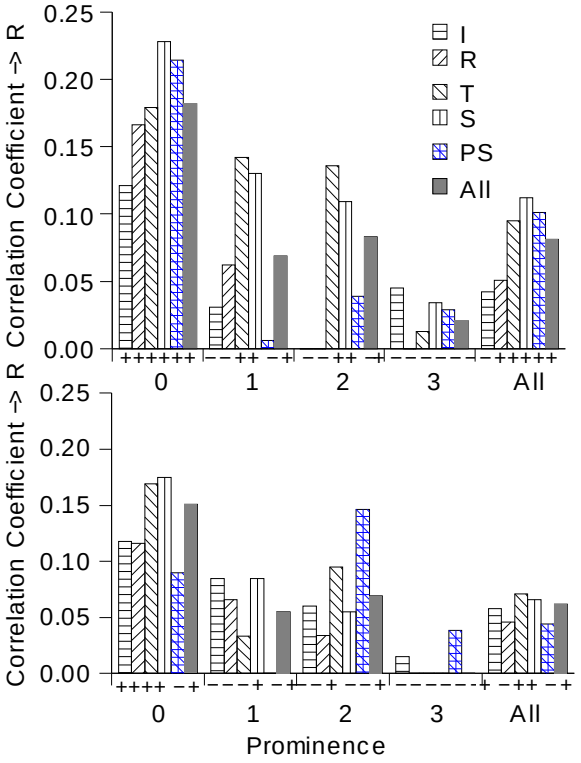


Figure 2. Correlation strength between segmental information and duration (top) and F_1/F_2 contrast (bottom). Plotted is a breakdown on style and prominence marks. Speaker, lexical stress, and vowel identity are accounted for (see text). I: Informal speech, R: retold story, T: text reading, S: isolated sentences, PS: pseudo−sentences. Statistics +: $p < 0.001$, −: not significant. Total N =40,385 tokens

2.3 Statistics

Speech efficiency is determined by correlating measures of phoneme reduction with segmental information values. A problem with this approach is that the acoustic measures of reduction are all highly dependent of many factors, which are unevenly distributed. However, all these factors have to be accounted for. To cope with the uneven data distribution, 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 by using the within (co−)variance of the quasi−uniform subsets (reducing the degrees of freedom to account for the subdivisions). In all analyses, we account for speaker and vowel identity, speaking style, lexical stress, and prominence. After applying a Bonferroni correction, a level of significance of $p < 0.001$ was chosen.

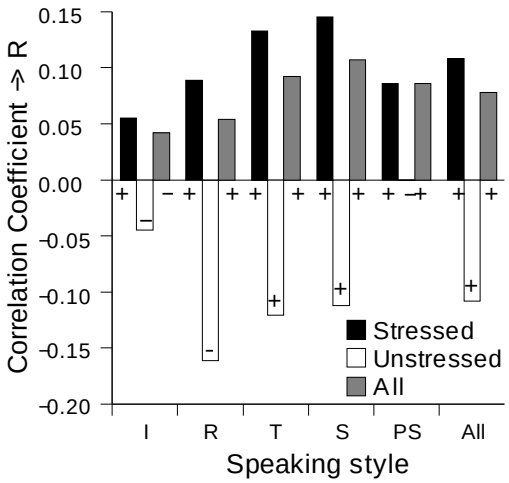


Figure 3. Correlation strength between segmental information and duration (left) and F1/F2 contrast (right). Plotted is a breakdown on style and lexical stress. Speaker and vowel identity, and prominence are accounted for (see text). I: Informal speech, R: retold story, T: text reading, S: isolated sentences, PS: pseudo−sentences. +: p < 0.001, −: not significant. Total N =40,385 tokens

2.4 Prosody

The three relevant factors for vowel reduction in Dutch are syllable stress, sentence accent/prominence, and prosodic boundaries. The IFA corpus marks lexical stress. Prominence is assigned automatically by rules from text input [13][14]. Prosodic boundaries are not yet marked so we will ignore them here. The rules for assigning prominence used are:

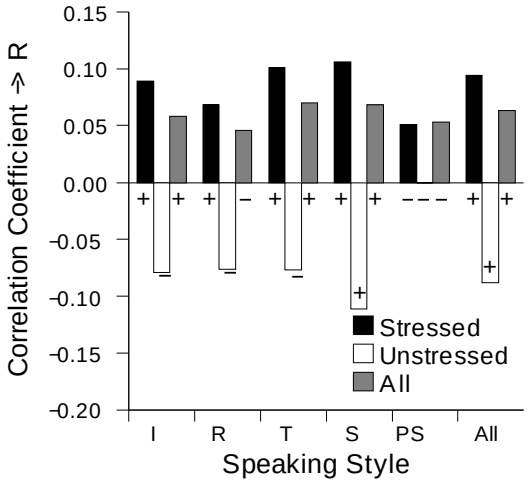
I. Each content word receives 2 marks
II. Each word from the classes [Noun, Adjective, Numeral, Negation] receives an additional mark
III. Polysyllabic content−words from the classes [Pronoun, Verb, Adverb] receive an additional mark
IV. The first content word in a sentence receives an additional mark (only implemented for the first 3 words)
V. Each Noun preceded by an Adjective loses one mark

Function words receive 0, content words 1−4 marks. Too few words received 4 marks to allow for statistical analysis. So only words with 0−3 prominence marks are used. Rule−based prominence marks correlated well with human transcribers (Cohen's Kappa = 0.62) [13][14].

3. RESULTS

3.1 Position in the word

Figure 1 shows the distribution of segmental information over words for the different phoneme classes. We see the expected decrease in segmental information value with increasing length of the word−onset. There seem to be no fundamental differences between the different phoneme classes.



3.2 Prominence and stress

Figure 2 displays the correlation between segmental information and duration and spectral contrast grouped on prominence. It is obvious that there is a consistent, and statistically significant, correlation between vowel reduction and segmental information, except when prominence is 3. This correlation is strongest in function words (prominence is 0) and for meaningful reading styles. More prominent words and spontaneous speaking styles as well as reading of pseudo−sentences were associated with weaker (often statistically not significant) correlations with segmental information.

Grouping the results on lexical stress (Figure 3) gives the same results as grouping on prominence for stressed vowels. However, for unstressed vowels, the correlation changes sign, indicating less reduction with higher redundancy. This "anti−efficiency" change of sign disappears when we either account for the position of the syllable in the word or for a hard floor in the reduction for low (<1 bit) segmental information content (not shown). Due to the increased number of factors and, therefore, subdivisions, the data became too sparse to allow reliably statistics.

3.3 Other factors

Correlations calculated using the position in the word were consistently lower, generally statistically not significant and often changed sign (not shown). Correlations between plain word frequencies from the Celex list and acoustic reduction were generally strong, but had different signs for function and content words. The correlations found would mean that content words with lower frequencies were more reduced (very *in*efficient). This effect proved to be completely determined by the low−frequency words. High−frequency words behaved as expected. The fact that for rare words a lower frequency combines with more reduction could be a word−length effect, rare words tend to be longer with more trailing redundant vowels.

4. DISCUSSION AND CONCLUSIONS

Figure 1 gives a comparative overview of the distribution of segmental information over word−internal positions. It 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 and 88% of word−tokens in the IFAcorpus have word−inital stress).

Both Figures 2 and 3 show that segmental redundancy correlates consistently with acoustic reduction, both for duration and spectral (formant) reduction. It is clear that the effects are strongest for read speech, either full text or isolated sentences. This can possibly be explained from the fact that the prominence marks were modeled after read sentences [13][14]. Other speech styles might fit less well. Furthermore, the unexpected semantic content of the pseudo−sentences (PS) might interfere with normal speech planning. The effect of redundancy is strongest on function words (prominence 0) and lowest on words with the highest prominence markings. This difference disappears when we repeat the analysis on only the high−frequency words (overall $R= 0.119$ and frequency $> 2^{-12}$; $p<0.001$, not shown). This suggests that the processing demands for assembling (low−frequency) words on−line can interfere with efficiency in speaking.

To summarize, we do find a consistent correlation between the distinctive (information) importance of a vowel for word identification and its acoustic reduction in terms of duration and spectral distinctiveness. This correlation is found after accounting for speaker and vowel identity, speaking style, lexical stress, (modeled) prominence, and position of the syllable in the word. The effect of the last factor could not be studied in detail due to data−sparseness. The correlation is only found for high−frequency words and strongest for "normal" reading tasks. This suggests that the efficiency of speech (vowel) production might be limited by the processing demands of word assembly. That is, our study suggest that efficiency in vowel production might rely on retrieving stored words from memory.

5. ACKNOWLEDGEMENTS

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 Organization.

6. 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.