Information Structure and Efficiency in Speech Production

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

Chair of Phonetics/ACLC
University of Amsterdam, the Netherlands
{Rob.van.Son, Louis.Pols}@hum.uva.nl

Abstract

Speech is considered an efficient communication channel. This implies that the organization of utterances is such that more speaking effort is directed towards important parts than towards redundant parts. Based on a model of incremental word recognition, the importance of a segment is defined as its contribution to word-disambiguation. This importance is measured as the *segmental information content*, in bits. On a labeled Dutch speech corpus it is then shown that crucial aspects of the information structure of utterances partition the segmental information content and explain 90% of the variance. Two measures of acoustical reduction, duration and spectral center of gravity, are correlated with the segmental information content in such a way that more important phonemes are less reduced. It is concluded that the organization of conventional information structure does indeed increase efficiency.

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 [1-5][10][12][20][23][27][28]. For example, *nine* is pronounced more reduced in the proverb *A stitch in time saves nine* than in *The next number is nine* [12].

Speakers can enhance efficiency by manipulating the (prosodic) structure of the utterance. The *Information Structure* of an utterance, i.e., the partitioning of the utterance according to information value, is generally reflected in the (hierarchical) phonological structure of the utterance. This phonological structure again is reflected in the way words and syllables are (de-)stressed and, therefore, (de-) emphasized in articulation. This way, more informative parts are emphasized in articulation, and less informative parts are de-emphasized, making the utterance more efficient with respect to the transfered information.

To quantify the efficiency at the articulatory level, the effort invested in the "unit of articulation" must be matched against the importance of this unit. In this paper we take the phoneme segment as the unit of articulation. The importance of an individual phoneme realization is measured in terms of the realization's (incremental) contribution to word recognition. Theories of word recognition stress that word recognition is an incremental task that works on a phoneme by phoneme basis [14]. 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 a frequency-sensitive incremental match of incoming phonemes in the mental lexicon [14][22]. However, words are also primed by their context [8][13]. We will model this priming as an increase in apparent word frequency (cf., [28]).

We use a measure of the position-dependent segmental contribution in distinguishing words given the preceding word-onset [25]. The lexical information I_L (in bits) of a segment ***s*** preceded by [*word onset*] is [24][25]:

$$I_L = -\log_2\left(\frac{Frequency([\,word\,onset\,]+s)}{Frequency([\,word\,onset\,]+any\,segment)}\right) \quad (1)$$

Frequencies are calculated from a CELEX word-count list with normative transcriptions of Dutch, based on 39 million words (N_{tot}). The word *frequencies* were estimated using a Katz smoothing on counts from 1-5 and an extrapolation based on Zipf's law [9].

Equation 1 does not account for the predictability of the word due to its distributional (contextual) properties [8][13] (cf, [1][16]). It is possible to determine the average predictability of the word spoken in its proper context. Words tend to occur in certain contexts more than in others (e.g., *very good idea* vs. *curious 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 [13]. The resulting value is called the *Context Distinctiveness* of the word w ($CD(w)$) and has a value between 0 and the $-\log_2$ of the global frequency of the target word [13]. In formula:

$$CD(w) = \sum_{vocabulary} P(c_i|w)\log_2\left(\frac{P(c_i|w)}{P(c_i)}\right) \quad (2)$$

Where $P(c_i)$ is the plain probability of the word c_i and $P(c_i|w)$ the conditional probability of word c_i appearing in the context of the target word w. On average, the relative frequency, $CF(w)$, of the target word w is a factor $2^{CD(w)}$ higher in its normal context than in the corpus as a whole, i.e., $CF(w) = RelativeFrequency(w)\cdot 2^{CD(w)}$. Equation 1 is changed to include a correction on the frequency of the target word w:

$$D(w) = CF(w)\cdot N_{tot} - Frequency(w) \quad (3)$$

Where $CF(w)$ can be based on a different corpus than N_{tot} and $Frequency(w)$. The segmental information, I_S, then becomes:

$$I_S = -\log_2\left(\frac{Frequency([word\,onset]+s)+D(w)}{Frequency([word\,onset]+any\,segment)+D(w)}\right) \quad (4)$$

2. Speech material and Methods

For I_S and the acoustic part of this study we used the IFAcorpus [11][26] which contains 5½ hours (50 kWord) of hand-aligned phonemically segmented speech from 8 native speakers of Dutch, 4 female and 4 male. 5 of the 8 available speaking styles were used: 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*). Acoustic reduction is measured on phoneme duration and on the phoneme mid-point spectral Center of Gravity (CoG) [21].

Distinctiveness (*CD*) was calculated over the 5th release of the Spoken Dutch Corpus (CGN), a total of 1.8 million words [15], over a window of 10 words (5 before and 5 after the target word [13]). 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 in the CGN, i.e., using $P(w)$ from CELEX.

As an illustration, the segmental information, I_S, is calculated for the vowel /o/ in the Dutch word /bom/ (*boom*, English *tree*, example taken from [25]).

- Relative CGN frequency of *boom*: $5.05 \cdot 10^{-5}$
- Context Distinctiveness: $CD(boom) = 4.53$ (eq. 2, CGN)
- Relative frequency in context: $2^{CD(boom)} \cdot 5.05 \cdot 10^{-5} = 1.2 \cdot 10^{-3}$
- Original smoothed CELEX word count of *boom*: 2,226
- Context-corrected CELEX count: 45,402 $(1.2 \cdot 10^{-3} \cdot 39 \cdot 10^{6})$
- Correction term: $D(boom) = 45{,}402 - 2{,}226 =$ **43,176** (eq. 3)
- Words starting with /bo/: **67,710** (1,172 CELEX entries)
- The same for /b./: **1,544,483** (26,186 CELEX entries)
- $I_L = -\log_2(67710/1544483) =$ **4.51** (eq. 1)
- $I_S = -\log_2([67710+43176]/[1544483+43176]) =$ **3.84** (eq. 4)

That is, $I_S < I_L$, so 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. The lexical normative transcription of the word-onset and phoneme identity are used to search the CELEX word-list.

To cope with the factors that affect measured quantities, the data are divided into quasi-uniform subsets. Each subset contains all observations that are uniform with respect to all relevant factors. Note that the resulting "tables" of factor values are extremely sparse. In general, far less than half of all *possible* subsets had any values in them (there can be millions of possible subsets). When all factors are accounted for, the average number of observations per *filled* subset is actually less than 2 (for duration).

Variances are calculated after equalizing the means in the subsets (i.e., mean=0) and subtracting 1 degree of freedom for each subset (ignoring subsets with only a single value). 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. All factor "values" were determined automatically from transcribed and tagged text. Prominence is assigned automatically by rules from text input based on POS tags [17][18][24]. 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) [17][18].

The importance of "distributional" factors was determined by their influence on the variance of the segmental information, I_S, duration, or spectral CoG. Starting with no or minimal subdivisions, each time a factor was selected that reduced the variance most. In the next round, from the remaining factors, the one that reduced the variance most after applying all previous factors was chosen. An F-test was used to determine whether a change in variance was statistically significant. After applying a Bonferroni correction, a level of significance of $p < 0.001$ was chosen for comparing factors.

The following distributional factors are used:

- Phoneme position : Position of segment in word
- Phoneme : Phoneme identity
- Nr. of Syllables : Word-length in syllables

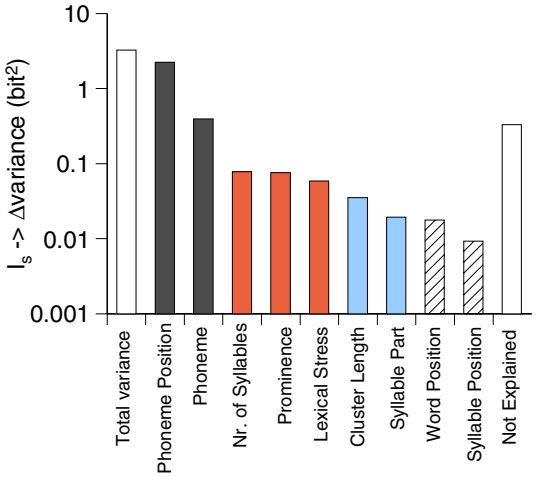

Figure 1. Reduction in variance of the segmental information (vertical scale) due to accounting for the indicated factor and all the factors to the left of it (horizontal scale). All phonemes, excluding phonemes from syllables containing a schwa. N=26,411, maximal number of subsets: 6,428. Not Explained variance: 10.2% of Total variance. All factors p<0.001 (F-test). White columns are plain variances, not differences.

- Prominence : Automatically determined prominence
- Lexical Stress : Lexical syllable stress
- Cluster length : Length of consonant clusters
- Syllable Part : Onset, Kernel, or Coda
- Word position : Position of word in sentence (1-5, >5)
- Syllable position : Position of syllable in word

3. Results

Schwa's are maximally reduced phonemes. The same will likely hold for consonants in the same syllable as a schwa. Syllables containing a schwa generally signal affixes and particles (clitics) in Dutch. It is unclear how these syllables function in word recognition and whether our simple model for segmental information content and speech efficiency is relevant to them. To prevent these maximally reduced syllables to swamp our statistics, we decided to exclude phonemes from syllables that contain a schwa.

Figure 1 gives the distribution of explained variance for the segmental information, I_S, using non-repeated material: *Retold* speech and *read sentences* from speaker's own stories [26]. The variance explained by a factor is calculated by subtracting the variance calculated *including* the target factor (and all factors to the left of it in the graph) from the variance calculated *without* the target factor. For example, the column for *Prominence* in Figure 1 (0.076 bit^2) is the difference between the variance calculated using quasi-uniform subsets for *Phoneme position*, *Phoneme* (identity), and *Number of Syllables* (0.549 bit^2, not shown) and the variance calculated when these subsets are again subdivided on *Prominence* (0.473 bit^2, not shown).

All factors together explain 90% of the variance in I_S. Each factor's reduction of the variance was statistically significant ($p<0.001$, F-test). The two principle segmental factors are the position of the phoneme in the word and the phoneme identity (two dark-gray columns in Figure 1). Together they explain 81% of the total variance in segmental information content, I_S. This can be explained by the fact that

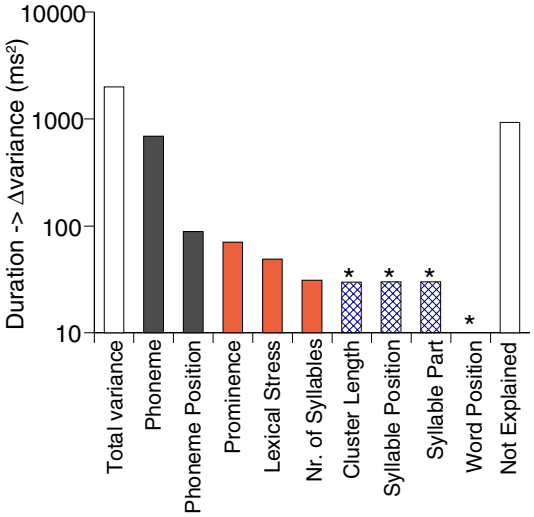

Figure 2. As Fig 1, but now for phoneme duration. All continuant phonemes (no Stops), excluding phonemes from syllables containing a schwa. N=85,922, maximal number of subsets: 43,799. Factors up to Number of syllables: p<0.001 (F-test). Crosshatched columns (, not significant) are differences calculated with respect to the Lexical Stress column. The Not Explained variance is 53.5% of the Total variance, calculated with respect to the Number of syllables column.*

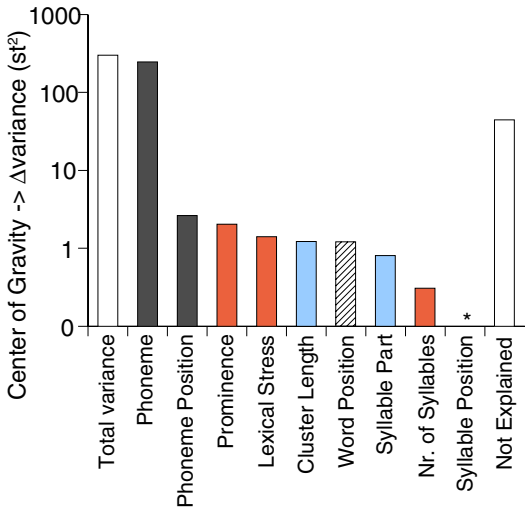

Figure 3. As Fig 2, but now for spectral Center of Gravity (in semitones). N=85,890, maximal number of subsets: 38,795. All factors p<0.001 (F-test) except Syllable position (, not significant). Not Explained variance: 14.9% of Total variance.*

the definition of the segmental information content, I_S, is based on the phoneme and searches the phonemically transcribed word-list from the start of the word. The other factors are evaluated with respect to the remaining variance after accounting for these two segmental factors.

The next three factors model word-level aspects of the speech, word-length in syllables, prominence, and lexical stress (red/gray columns), together these three explain 34% (9-12% each) of the remaining variance. These three factors are all correlated to the frequency of occurrence of the words and syllables. Longer words are less common. Prominence separates common function words from rare content words, and within content words, it favors the (low-frequency) Nouns and Adjectives[17][18]. Lexical stress tends to fall on the most informative (least common) syllable [29].

The blue/light-gray columns in Figure 1 mark sub-syllabic factors (length of the consonant cluster and part of the syllable), that together account for 8.7% of the remaining variance. The last two positional factors (hatched columns) together explain only 4.3% of the remaining variance. Together, all 7 supra-segmental factors explain 46.5% of the remaining variance of the segmental information content.

To evaluate acoustic reduction, we also accounted for speaker, speaking style, and recording session (duration only). These factors have large influences on speech acoustics and speaking rate, but are not modeled in this study. Figure 2 shows the results for phoneme duration (excluding *Stops*). The *order* of importance of the factors found for duration correlates well with that found for I_S (Spearman's Rank Correlation, $R=0.833$, p<0.003, no Bonferroni correction).

For duration, the two segmental factors (dark gray columns) account for 38.9% of the variance. Maximally 53.5% of the variance is accounted for by the 9 factors used. The 46.5% of variance not explained can be traced to segmentation "noise" and contextual (phonological) factors not modeled here. The word-level factors (red/gray columns) together explain 12.4% of the remaining variance. The variance is at its minimum when the *Number of Syllables* is accounted for. However, it was still possible to determine an order of minimal variance in the remaining factors. The reduction of the variance for all factors up to *Number of Syllables* is statistically significant (p<0.001, F-test). The other factors do not reduce the variance (not significant, cross-hatched).

Figure 3 shows the results for spectral Center of Gravity (CoG, semitones). The selections are like figure 2. In Figure 3, only the *Syllable Position* factor does *not* decrease the variance (not significant). The *Number of Syllables* has the lowest effect on the variance of the CoG (p<0.001). The total variance explained is 85%, but this is almost completely due to the phoneme identity (82%). Of the variance remaining after accounting for the two segmental factors (dark-gray), only 13.6% can be explained by the other factors used. This low explanatory power can, at least partly, be traced to the inherently noisy character of CoG measurements [21][25].The order of the factors found for CoG correlates with that found for I_S (Spearman's Rank Correlation, $R=0.716$, p<0.04, no Bonferroni correction).

For every set of factors accounted for, there is a positive correlation between I_S and both phoneme Duration and CoG (normalized correlations, $R \approx 0.03$, $p < 0.001$, cf, [21][25]). These weak, positive correlations still show that the relation between I_S and reduction is always towards greater efficiency.

4. Discussion and Conclusions

The definition of the *communicative* importance as the segmental information content, I_S, with respect to incremental word-recognition (eq. 4) agrees very well with the common factors used to describe Information Structure. With 9 factors, 90% of the variance in I_S can be explained. All factors used in this study were determined automatically from tagged text (including *Lexical Stress* and *Prominence*). It is therefore surprising how well the order of factors in Figure 1 match that determined for the measured duration and CoG ($R=0.833$ and 0.716). Obviously, phoneme identity is the single most important determinant of segmental duration and spectral Center of Gravity. But the fact that the position in the word is the second most important factor to explain the variance of

duration and spectral Center of Gravity is not obvious unless efficiency is taken into account.

This paper confirms the importance of *Prominence*, *Lexical Stress*, and W*ord-Length* for efficient speech [19]. Acoustic measurements are inherently noisy [21]. Furthermore, phoneme duration is influenced by many contextual factors, e.g., next phoneme, which cannot be simultaneously modeled on this limited material. Still, figures 2 and 3 show that, at all levels, the factors that distinguish important from redundant parts in an utterance also distinguish reduced from emphasized parts.

The fact that there is always a positive correlation between segmental information content and measures of reduction points towards an efficient organization of speech.

We conclude that many factors that determine the suprasegmental information structure of speech separate the important from the redundant parts of the utterances. The same factors also separate more and less reduced phonemes with respect to phoneme duration and spectral center of gravity. That is, the conventional information structure indeed increases the efficiency of speech at the segmental level.

5. 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.

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

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

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

[11] IFAcorpus, http://www.fon.hum.uva.nl/IFAcorpus Available under the GNU General Public License.

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

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

[14] 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.

[15] Oostdijk, N., Goedertier,W. ,Van Eynde,F. , Boves,L. , Martens,J.P., Moortgat,M., and Baayen,H., "Experiences from the Spoken Dutch Corpus Project", Proceedings of the third International Conference on Language Resources and Evaluation, 340-347, 2002.

[16] 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.

[17] 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.

[18] Streefkerk, B.M., "Prominence. Acoustical and lexical/syntactic correlates", Ph.D. Thesis University of Amsterdam, p169, 2002.

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

[20] 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.

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

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

[23] 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.

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

[25] Van Son, R.J.J.H. & Pols, L.C.W., "An Acoustic Model of Communicative Efficiency in Consonants and Vowels taking into Account Context Distinctiveness", Proceedings of ICPhS2003, Barcelona, Spain, 2003.

[26] 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.

[27] 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.

[28] 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.

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