Institute of Phonetic Sciences,
University of Amsterdam,
Proceedings 25 (2003), 171–184.

How efficient is speech?[1]

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

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 reduction according to 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 any predictable items (Aylett, 1999; Boersma, 1998; Borsky et al., 1998; Cutler, 1987;1995; Fowler, 1988; Lieberman, 1963; Van Son et al., 1998; Van Son & Pols, 1999c; Vitevitch et al., 1997; Whiteside & Varley, 1999). For example, *nine* is pronounced more reduced in the proverb *A stitch in time saves nine* than in the neutral expression *The next number is nine* (Lieberman, 1963).

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 redundant parts are de-emphasized, making the utterance more efficient with respect to the transfered information.

1 This paper is a compilation and extension of a paper presented at ICPhS2003 (Van Son & Pols, 2003a) and another one at Eurospeech2003 (Van Son & Pols, 2003b).

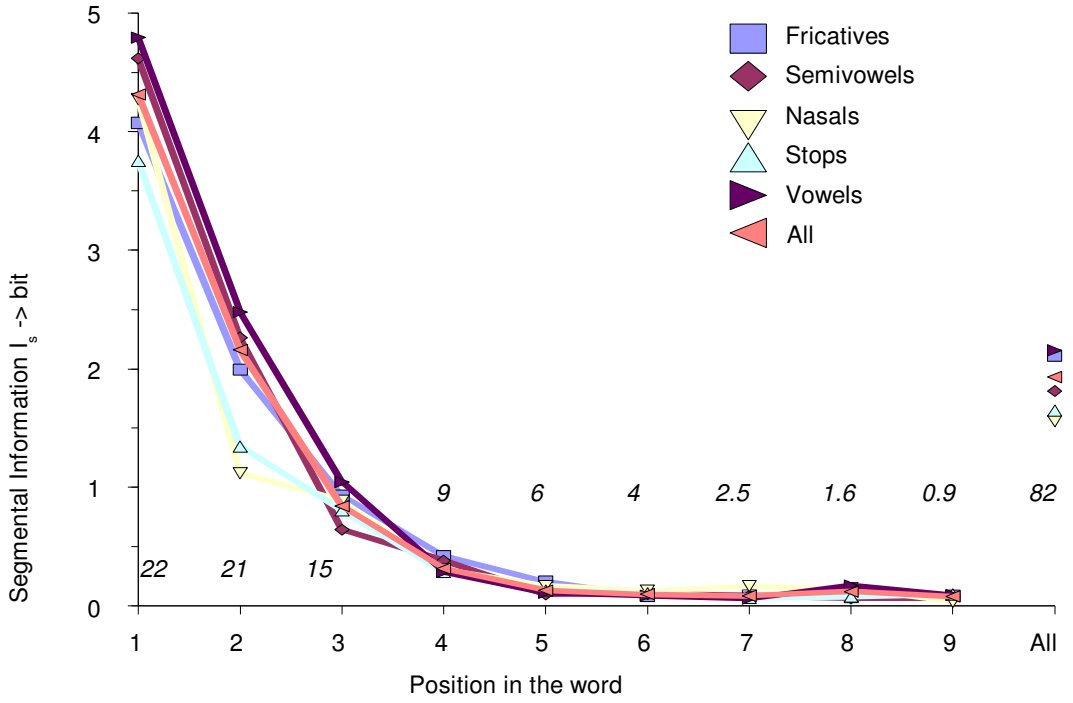

Fig. 1. Relation between average segmental information (I_s) and the position in the word grouped by manner of articulation. The total number of segments (x1000) for each position is indicated with italic numbers.

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 realizations (incremental) contribution to word disambiguation in recognition. Theories of word recognition stress that word recognition is an incremental task that works on a phoneme by phoneme basis (Norris et al., 2000). Often, words are recognized by their first syllable well before all phonemes have been processed (Cutler, 1997). In English and Dutch the importance of the first syllable is reflected in the fact that lexical stress is predominantly on the first syllable of a word (Cutler, 1987; Cutler, 1997). We use a model of word recognition with competition based on a frequency-sensitive incremental match of incoming phonemes in the mental lexicon (Norris et al., 2000; Van Son & Pols, 1999c). However, words are also primed by their context (Ferrer i Cancho & Solé, 2001, 2003; McDonald & Shillcock, 2001). We will model this priming as an increase in apparent word frequency (cf., Whiteside & Varley, 1999).

2 The importance of a segment

We use a measure of the position-dependent segmental contribution in distinguishing words given the preceding word-onset (Van Son & Pols, 2003a, b). The lexical information I_L (in bits) of a segment *s* preceded by [word onset] is:

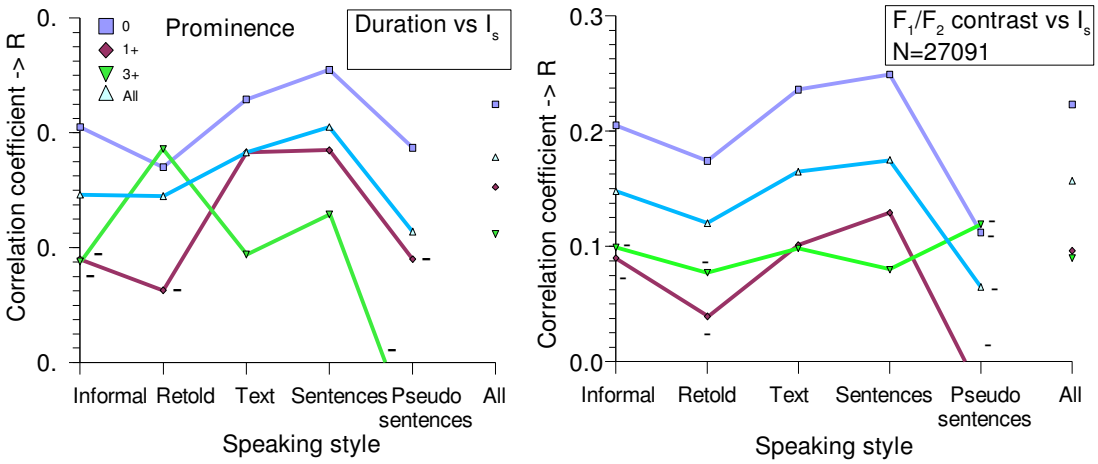

Fig. 2. Correlation coefficients for *vowel* tokens between segmental information and duration (left) and F_1/F_2 contrast (right). Plotted is a breakdown on style and prominence marks. Speaker, lexical stress, vowel identity, and type of text are accounted for (see text). Excluded are schwa vowels and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

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

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 (Ferrer i Cancho & Solé, 2003).

Equation (1) does not account for the predictability of the word due to its distributional (contextual) properties (Ferrer i Cancho & Solé, 2001; McDonald & Shillcock, 2001; cf, Aylett, 1999; Owens et al., 1997). 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* <u>*idea*</u> vs. *curious green* <u>*idea*</u>). This means that the frequencies of words in the neighbourhood 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 (McDonald & Shillcock, 2001). The resulting value is called the Context

Table 1: Factors used to describe the information structure of an utterrance. For each factor the major (sub-)component is given, if it exists.

<u>Segmental Factors</u>		
1. Phoneme position	: Position of segment in word	*Word-boundary*
2. Phoneme	: Phoneme identity	*Phoneme class*
<u>Word Level Factors</u>		
3. Nr. of syllables	: Word-length in syllables	*Mono/Polysyllabic*
4. Prominence *word*	: Automatically determined prominence	*Function/Content*
5. Lexical stress	: Lexical syllable stress	-
<u>Syllable Level Factors</u> (consonants only)		
6. Cluster length	: Length of consonant clusters	-
7. Syllable part	: Onset, Kernel, or Coda	-
<u>Other Factors</u>		
8. Word position	: Position of word in sentence (1-5, >5)	*Sentence boundary*
9. Syllable position	: Position of syllable in word	*Word boundary*

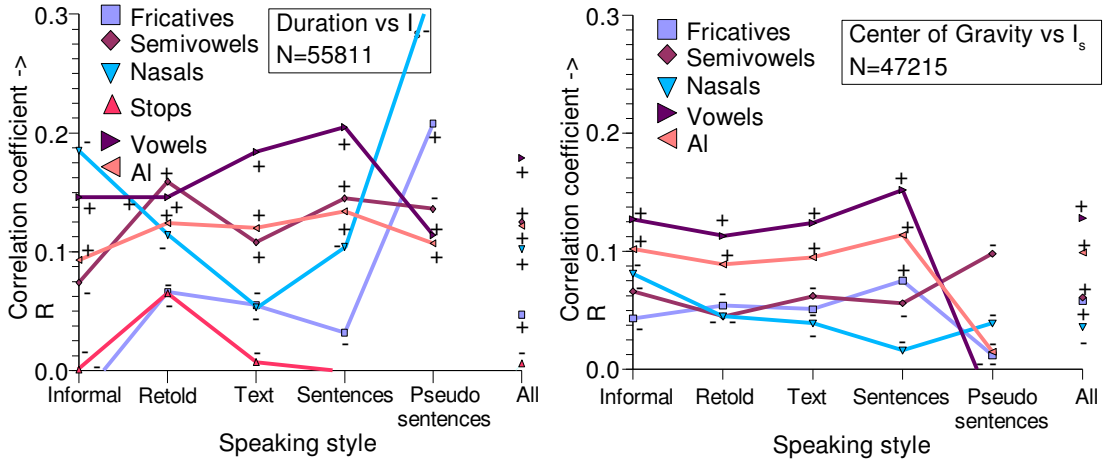

Fig. 3. Correlation coefficients between segmental information and duration (left) and Center of Gravity (right). Plotted is a breakdown on style and manner of pronunciation. Speaker and phoneme identity, prominence, lexical stress, and position in the syllable are accounted for (see text). +: $p < 0.001$,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

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 (McDonald & Shillcock, 2001). 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*)·$2^{CD(w)}$. Equation (1) is changed to include a correction on the absolute 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:

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

Context Distinctiveness (CD) was calculated over the 5th release of the Spoken Dutch Corpus (CGN), a total of 1.8 million words (Oostdijk et al., 2002), over a window of ten words (five before and five after the target word, see McDonald & Shillcock, 2001). 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 from Van Son & Pols, 2003a).

- Word tokens 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** (c.f. eq. (1))

- Relative CGN frequency of *boom*: **$5.05 \cdot 10^{-5}$**
- Context Distinctiveness: CD(*boom*) = **4.53** (eq. 2, CGN)
- Relative frequency in context: $2^{CD(boom)} \cdot 5.05 \cdot 10^{-5} =$ **$1.2 \cdot 10^{-3}$**
- CELEX word count of *boom*: **2,226** (smoothed count)
- 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))

$I_S = -\log_2([67710+43176]/[1544483+43176]) =$ **3.84** (c.f. 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 a Dynamic Programming algorithm. The lexical normative transcription of the word-onset and phoneme identity are used to search the CELEX word-list.

For I_S and the acoustic part of this study we used the IFAcorpus (IFAcorpus, 2001; Van Son et al., 2001) which contains 5½ hours (50 kWord) of hand-aligned phonemically segmented speech from eight native speakers of Dutch, four female and four male. Five of the eight 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 can be measured on such features as duration and on the spectral Center of Gravity (CoG, c.f., Van Son & Pols, 1999a). 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.

3 Correlational analysis

3.1 Methods

To cope with the many different factors that affect acoustic measures, the data are divided into quasi-uniform subsets. Each subset contains all observations that are uniform with respect to all relevant factors (see Table 1, e.g., a realization of an /a/ in a mono-syllabic word with prominence 0, in the nucleus of a stressed syllable, etc., versus another phoneme /x/ in a bi-syllabic word with prominence 2, in the coda of an unstressed second syllable, etc., from informal speech of female speaker N and read speech of male speaker O, respectively). All factor "values" were determined automatically from transcribed and tagged text. 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. 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

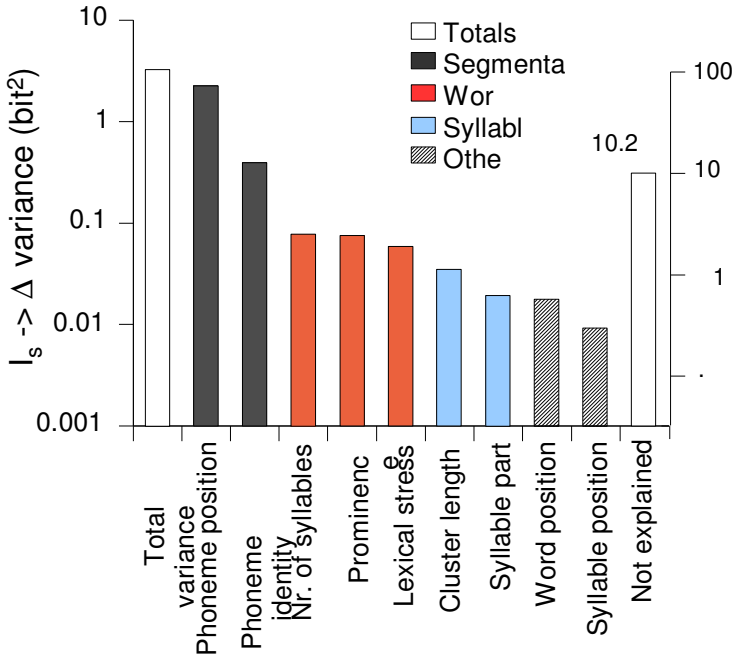

Fig. 4. Variance (first and last column) and reduction in variance of the segmental information I_S (vertical scale, % of total on the right hand scale) due to accounting for the indicated factor and all the factors to the left of it (horizontal scale). The order of the factors is the same as in table 1. The data concern 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.

accounted for, the average number of observations per filled subset, i.e., excluding empty sub-sets, is actually less than two (for duration).

In all analyses, we account for speaker and *phoneme identity*, *speaking style*, *text type* (fixed story or a speaker's own 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 (Streefkerk et al., 2001; Streefkerk, 2002; Van Son & Pols, 2002). 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, Streefkerk et al., 2001; Streefkerk, 2002).

3.2 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 (Van Son & Pols, 2002). 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.

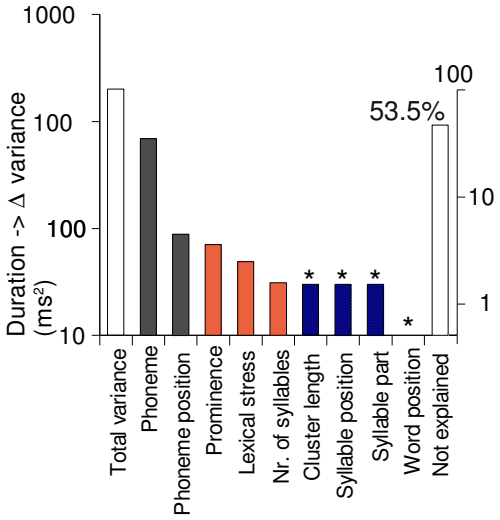

Fig. 5. As Fig 4, but now for **phoneme duration**. Data concerning all continuant phonemes (no Stops), excluding phonemes from syllables containing a schwa. N=85,922, maximal number of quasi-uniform subsets: 43,799. Black and gray columns: p<0.001 (F-test,). Crosshatched columns (*): not significan. The Not Explained variance is 53.5% of the Total variance, calculated with respect to the Number of syllables column.

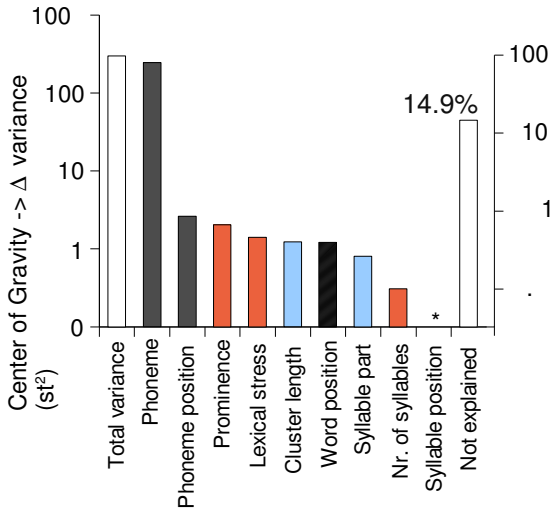

Fig. 6. As Fig 5, but now for **spectral Center of Gravity** (in semitones). N=85,890, maximal number of subsets: 38,795. All factors p<0.001 (F-test) except Syllable position (*, not significant). Not Explained variance: 14.9% of Total variance.

The schwa, /@/, is a completely assimilated vowel (Van Bergem, 1993). 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 frequent and 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, since our earlier study had shown that there is a floor in the effect of segmental information (Van Son & Pols, 2002).

Figure 2 displays the correlation between segmental information, I_S, and vowel duration (left) and formant contrast (right) grouped on prominence. It is obvious that there is a consistent, and statistically significant, correlation between vowel reduction (in terms of duration and F_1/F_2 contrast) and segmental information (in terms of I_S). 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 the vowel segments 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).

3.3 Discussion

Figure 1 clearly shows the importance of "early" phonemes. Dutch (and English) increase recognition efficiency by a prevalence for word-initial *lexical stress* (Cutler, 1987), i.e., 73% of word-forms in the IFAcorpus have word-initial *lexical stress* (Van Son & Pols, 2002). The strong correlation between position in the word and I_S prevents us from separating these two. A separate 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, I_S, 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 (Van Son & Pols, 1999a). In the next section we will investigate the contributions of these individual factors to efficiency in terms of variance explained.

To summarize, we do find a consistent correlation between the distinctive (information) importance of a phoneme for (incremental) word recognition 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.

4 Factor Contributions

4.1 Speech material and Methods

The factors chosen for this study are presented in Table 1. 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 from Table 1 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.

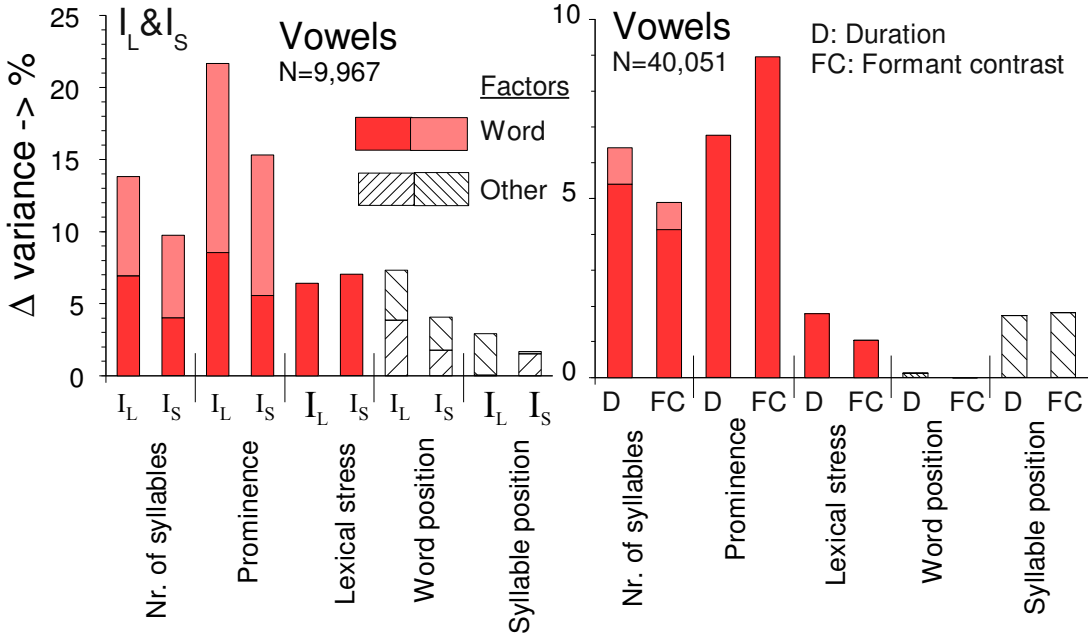

Fig. 7. Explained variance in percent with respect to the remaining variance after accounting for the segmental factors *Postition in the word* and *Phoneme identity*. Left, information content in I_L and I_S. Right, Reduction in terms of Duration and Formant contrast. Within each individual factor, the contribution of the main component (rightmost column in Table1) is displayed as a darker/lower sub-bar. The explained variance after using the additional finer distinctions are displayed in the lighter/upper bars All indicated bars are significant (F-test, $p<0.001$).

4.2 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 that these maximally reduced syllables will swamp our statistics, we decided to exclude phonemes from syllables that contain a schwa.

Figure 4 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 (Van Son et al., 2001). 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 4 (0.076 bit^2) is the difference between the variance calculated using quasi-uniform subsets for all the factors to the left of it, i.e., *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 leave 10% of the variance in I_S unexplained. 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 4). Together they explain 81% of

the total variance in segmental information content, I_S. This can be explained by the fact that 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* (gray columns). Together these three explain 34% (9-12% each not shown, ~2% each of total variance) of the remaining variance after accounting for the first two factors. 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 (Streefkerk et al., 2001; Streefkerk, 2002). Lexical stress tends to fall on the most informative (least common) syllable (Zue, 1985).

The two light-gray columns in Figure 4 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 seven supra-segmental factors explain 46.5% of the remaining variance of the segmental information content (i.e., after accounting for *phoneme position* in the word and *phoneme identity*).

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 5 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 46.5% of the variance is accounted for by the nine factors used. The 53.5% of variance not explained can be traced to segmentation "noise" and contextual (phonological) factors not modeled here. The three word-level factors (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 6 shows the results for spectral Center of Gravity (CoG, semitones). The selections are like Figure 5. In Figure 6, 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 (Van Son & Pols, 1999a; Van Son & Pols, 2003).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, $R \sim 0.03$, $p < 0.001$, (cf,

Van Son & Pols, 1999a; Van Son & Pols, 2003a). These weak, positive correlations still show that the relation between I_S and reduction is always towards greater efficiency.

4.3 Dissecting factors

The factors used to explain the information structure of utterrances generally are dominated by one major determinant and a set of finer distinctions. For example, the major determinant of our automatically assigned prominence is whether the word is a function word or a content word. Less effects are expected from distinguishing more levels of prominence. For the vowel data, Figure 7 presents the contribution of the major determinant (in %) to the overall variance with respect to the variance that remains after accounting for the two segmental factors, *phoneme identity* and *position in the word.* The lexical information content was included for comparison.

The results are most clear for the information content, I_S and I_L (see left-hand side of Figure 7). Except for the *syllable position*, the complete factors, including the finer distinctions, explain significantly more variance than the coarser distinction of the major component alone. That is, the effect of the *number of syllables* in a word, including differences between polysyllabic words, is more than the difference between mono- and polysyllabic words, and *prominence,* including higher levels*,* is more than a distinction between function and content words.

For *reduction* (in terms of duration and formant contrast, see right-hand side of Figure 7) it is found that the relative amount of variance explained is roughly half the amount from the information content. Furthermore, only the *number of nyllables* is more than its principal determinant, mono- vs. polysyllabic. For all other factors the effects on reduction can be explained by the major determinant alone and finer distinctions are unnecessary.

From Figure 7 it would be possible to conclude that speech only uses a rather coarse information structure based on only the major distinctions, e.g., function/content words. However, Figures 5 and 6 show that there is a rather large amount of unexplained variance in the reduction measurements. Which aspects of each factor are relevant can only be decided when the “noise” level in the acoustic measurements can be brought down.

4.4 Discussion

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 nine 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 4 match that determined for the measured duration and CoG (R=0.833 and 0.716, respectively). 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 study confirms the importance of *prominence*, *lexical stress*, and *word-length* for efficient speech (Van Bergem, 1993; Van Son & Pols, 1999b). Acoustic measurements are inherently noisy (Van Son & Pols, 1999a). Furthermore, phoneme duration is influenced by many contextual factors, e.g., next phoneme, which cannot be simultaneously modeled on this limited material. Still, Figures 5 and 6 show that, at all levels, the factors that distinguish *important* from *redundant* parts in an utterance also distinguish *reduced* from *emphasized* parts.

5 General Discussion

For maximal efficiency, the effort invested in each phoneme of an utterrance should be determined by its "unexpectedness", i.e., its information content. It is obvious that there are only surrogate measures available for both the invested effort, e.g., reduction, and the predictability or information content of a phoneme, e.g., our measures I_L and I_S. However, if speech is efficient, it should be possible to link the amount of reduction to the redundancy of a phoneme.

The relation between redundancy and reduction was investigated at two levels. First, it was shown that after accounting for all relevant known factors we do find a consistent correlation between the distinctive (information) importance of a phoneme for (incremental) word recognition and its acoustic reduction in terms of duration, Center of Gravity, and formant contrast. This correlation is found after accounting for speaker and *phoneme 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 analyzing this further. We conclude that speech is structured efficiently, even after accounting for the effects of prosodic structure and predictability in average context.

Second, the distribution of information content and reduction over the relevant factors was exactly what would have been expected in efficient speech: Factors, like *position in the word* or *number of syllables*, that are important for the information content are also important for the level of reduction. This is especially striking for *position in the word.* This non-linguistic factor (language rules never count), is the most important determinant of the information content, and the second most important determinant of reduction. 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.

Showing that speech is, overall, organized efficiently, does not tell us how this level of efficiency is achieved. It is unlikely that speakers count syllables or use Part-of-Speech labels like we did. Each factor can be decomposed into separate parts. For instance, the major determinant of the automatically determined *prominence* is the

distinction between *function* and *content* words. For the *number of syllables*, the distinction between *monosyllabic* and *polysyllabic* words is the most important. For vowels, the efficiency obtained by using these coarser, major determinant factors was compared with the gain in efficiency obtained by using the full factor distinctions. The result showed that the full factors did a significantly better job in explaining the variance in information content. However, the acoustic reduction was only marginally better explained by using the full factors (i.e., only for *number of syllables*). So, our results do not indicate that speakers use an elaborated mechanism for obtaining efficiency. They could have used only very coarse distinctions, e.g., between function and content words. This lack of significance of the fine-grained factor values could be caused by a high level of "noise" in the reduction measurements. It is necessary to lower the "noise" in the redundancy and reduction measurement to decide whether speakers can use fine-grained distinctions to enhance efficiency. A first requirement would be to obtain a larger text-corpus to calculate the redundancy measures.

6 Conclusions

Based on an incremental word recognition model with competition, the efficiency of speech as a communication channel was investigated. Our data indicate that the relation between the conventional information structure of speech utterances and the acoustic reduction of the individual phonemes enhances the efficieny of speech as a communication channel. The importance of a structural factor like *prominence* or *position in the word* for determining the redundancy predicts the importance of such factors for acoustic phoneme reduction. Furthermore, there is always a correlation between the redundancy of a phoneme and the level of acoustic reduction of this phoneme, even if all relevant factors are accounted for. The noisiness of reduction measures and a lack of a large text corpus prevent us from specifying the details of the relation between information structure and reduction.

Acknowledgments

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

References

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

Boersma, P.B. (1998). *Functional phonology, formalizing the interactions between articulatory and perceptual drives*, Ph.D. thesis, University of Amsterdam, 493 pp.

Borsky, S., Tuller, B. & Shapiro, L.P. (1998). "'How to milk a coat': The effects of semantic and acoustic information on phoneme categorization". J. Acoust. Soc. Am. **103**, 2670-2676.

Cutler, A. (1987). "Speaking for listening", in A. Allport, D. McKay, W. Prinz & E. Scheerer (eds.) *Language perception and production*, London; Academic Press, 23-40.

Cutler, A. (1995). “Spoken word recognition and production”, in J.L. Miller & P.D. Eimas (eds.) *Speech, Language, and Communication. Handbook of Perception and Cognition, 11*, Academic Press, Inc, 97-136.
Cutler, A. & Carter, D.M. (1987). “The predominance of strong initial syllables in English vocabulary” Computer Speech and Language **2**, 133-142.
Cutler A. (1997). “The comparative perspective on spoken-language processing”, Speech Communication **21**, 3-15.
Ferrer i Cancho, R. & Solé R.V. (2001). “The small world of human language”, Proceedings of the Royal Society of London B **268**, 2261-2265.
Ferrer i Cancho, R. & Solé R.V. (2003). “Least effort and the origins of scaling in human language”, PNAS **100**, 788-791.
Fowler, C.A. (1988). “Differential shortening of repeated content words in various communicative contexts”, Language and Speech **31**, 307-319.
IFAcorpus, http://www.fon.hum.uva.nl/IFAcorpus, Available under the GNU General Public License.
Lieberman, P. (1963). “Some effects of semantic and grammatical context on the production and perception of speech”, Language and Speech **6**, 172-187.
Lindblom, B. (1996). “Role of articulation in speech perception: Clues from production”, J. Acoust. Soc. Am. **99**, 1683-1692.
McDonald, S.C. & Shillcock, R.C. (2001). “Rethinking the word frequency effect: The neglected role of distributional information in lexical processing”, Language and Speech **44**, 295-323.
Norris D., McQueen J.M. & Cutler A. (2000). “Merging information in speech recognition: Feedback is never necessary”, Behavioral and Brain Sciences **23**, 299-325.
Oostdijk, N., Goedertier, W.,Van Eynde, F., Boves, L. , Martens, J.P., Moortgat, M. & Baayen, H. (2002). “Experiences from the Spoken Dutch Corpus project”, Proceedings of the third International Conference on Language Resources and Evaluation, 340-347.
Oostdijk, N. (2000). “The Spoken Dutch Corpus, overview and first evaluation”, Proceedings of LREC-2000, Athens, Vol. **2**, 887-894.
Owens, M., O'Boyle, P., McMahon, J., Ming, J. & Smith, F.J. (1997). “A comparison of human and statistical language model performance using missing-word tests”, Language and Speech **40**, 377-389.
Streefkerk, B.M., Pols, L.C.W. & ten Bosch, L.F.M. (2001). “Acoustical and lexical/syntactic features to predict prominence”, Proceedings of the Institute of Phonetic Sciences, University of Amsterdam **24**, 155-165.
Streefkerk, B.M. (2002). *Prominence. Acoustical and lexical/syntactic correlates*, Ph.D. Thesis, University of Amsterdam, p169.
Van Bergem, D.R. (1993). “Acoustic vowel reduction as a function of sentence accent, word stress, and word class”. Speech Communication **12**, 1-23.
Van Son, R.J.J.H., Koopmans-van Beinum, F.J. & Pols, L.C.W. (1998). “Efficiency as an organizing factor in natural speech”, Proc. ICSLP'98, Sydney, 2375-2378.
Van Son, R.J.J.H. & Pols, L.C.W. (1999a). “An acoustic description of consonant reduction”, Speech Communication **28**, 125-140.
Van Son, R.J.J.H. & Pols, L.C.W. (1999b). “Effects of stress and lexical structure on speech efficiency” Proc. Eurospeech'99, Budapest, 439-442.
Van Son, R.J.J.H. & Pols, L.C.W. (1999c). “Perisegmental speech improves consonant and vowel identification”, Speech Communication **29**, 1-22.
Van Son, R.J.J.H., Binnenpoorte, D., van den Heuvel, H. & Pols, L.C.W. (2001). “The IFAcorpus: a phonemically segmented Dutch open source speech database”, Proceedings of Eurospeech 2001, Aalborg, Denmark, Vol. **3**, 2051-2054.
Van Son, R.J.J.H. & Pols, L.C.W. (2002). “Evidence for efficiency in vowel production”, Proceedings of ICSLP2002, Denver, USA, Vol **I**, 37-40.
Van Son, R.J.J.H. & Pols, L.C.W. (2003a). “An acoustic model of communicative efficiency in consonants and vowels taking into account context distinctiveness”, Proceedings of ICPhS2003, Barcelona, Spain.
Van Son, R.J.J.H. & Pols, L.C.W. (2003b). “Information structure and efficiency in speech production”, Proceedings of Eurospeech 2003, Geneva, Switzerland.

Vitevitch, M.S., Luce, P.A., Charles-Luce, J. & Kemmerer, D. (1997). "Phonotactics and syllable stress: Implications for the processing of spoken nonsense words", Language and Speech **50**, 47-62.

Whiteside, S.P. & Varley, R.A. (1999). "Verbo-motor priming in the phonetic encoding of real and non-words", Proceedings of Eurospeech'99, Budapest, 1919-1922.

Zue, V.W. (1985). "The use of speech knowledge in automatic speech recognition", Proceedings of IEEE **73**, 1602-1616.