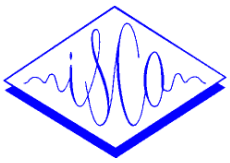

6th International Conference on Spoken Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

# AN ACOUSTIC PROFILE OF SPEECH EFFICIENCY

*R.J.J.H. van Son, Barbertje M. Streefkerk, and Louis C.W. Pols*

University of Amsterdam, Institute of Phonetic Sciences/IFOTT
Herengracht 338, 1016 CG Amsterdam, The Netherlands
tel: +31 20 5252183; fax: +31 20 5252197; email: Rob.van.Son@hum.uva.nl

## ABSTRACT

Speaking is generally considered efficient, in that less effort is spent articulating more redundant items. Two possible mechanisms for this optimization are tested. The use of prosodic structure, i.e., lexical stress and sentence accent, to (de-)emphasize (un-)important words, and the facilitation of syllable-articulation by retrieving often used motor-programs from memory. Such a "stored versus computed" principle in syllable articulation would implicitly result in efficient speech because of the correlations between syllable- and word-frequencies. These mechanisms are tested for Dutch speech by means of a hand labeled single-speaker corpus of spontaneous and matched read speech, and an automatically labeled multi-speaker corpus of read telephone speech. It is concluded that the use of lexical stress and sentence accent/prominence cannot explain all of the frequency-of-occurrence effects found in speech. Furthermore, at least in unstressed syllables, syllable-frequency effects proved to be more important than word-frequency effects, leaving room for an articulatory "stored versus computed" mechanism in the optimization of speaking effort.

## 1. INTRODUCTION

Speech can be seen as an efficient communication channel: less speaking effort is spent on redundant than on informative items. Previous studies showed that listeners tend to identify redundant tokens better and that speakers compensate for this by better articulating unpredictable items [1-5][7][9][16][19][20]. For example, *nine* is pronounced more reduced in the proverb *A stitch in time saves nine* than in the carrier sentence *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 [11]. 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 frequency effects only.

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. The question remains whether there is an effect of frequency in addition to these prosodic enhancements [1].

A further problem is the extent to which the optimization of speaking effort is a high-level process, involving syntactic, semantic, and pragmatic knowledge, instead of being the (low-level) consequence of the way articulation is organized and controlled [1]. There are suggestions that "assembling" the articulation of syllables is a time limiting process in speech [22]. Articulation speed would, therefore, benefit from using stored "motor-programs" for common syllables. Syllables whose articulations are retrieved from memory are pronounced faster, and most likely, more reduced than syllables that have to be assembled "from scratch". There is a strong correlation between syllable- and word-frequency: rare syllables must be part of rare words and are also most likely to carry lexical stress. Therefore, the emphasizing of rare syllables as a consequence of their rule based assembly will automatically emphasize the stressed parts of unpredictable words, as required for efficient speech.

In this paper we investigate these two questions simultaneously. First, is there a frequency of occurrence effect beyond lexical stress and sentence (pitch) accent or prominence? Second, could syllable familiarity determine reduction in pronunciation independent of word frequency?

## 2. MATERIALS

For this study we selected recordings of spontaneous Dutch speech and a read transcription of it from a single male speaker [16][17][19]. 791 Pairs of corresponding VCV realizations were selected for further study (see Table 1, and [16][17][19]). Accented words were marked by one of us. Monosyllabic function words are considered unstressed. Word medial consonants are considered to carry maximal stress (i.e., they are stressed when either neighboring vowel carries lexical stress). Other stress assignment methods gave comparable but less consistent results. The VCV pairs were randomly selected to cover all consonants present and both stress conditions (except for /h/, primary lexical syllable stress only [16][17][19]). 22 Native speakers of Dutch were asked to identify these 1582 intervocalic consonant realizations in their original VCV context [16][19]. For each token, the $\log_2$(Perplexity) of the 22 responses i.e., the entropy, was calculated and used as a measure of confusion [18]. Perplexity as a measure of confusion is expected to anti-correlate with all other parameters used in this study ($R < 0$). For convenience, all correlations with perplexity will be reversed to obtain positive correlations, i.e., $-R$ is used.

A second body of speech was selected from the Dutch Polyphone corpus [6][15]. This corpus contains 25000 newspaper sentences read by 5000 different speakers over the telephone [6][15]. A subset of 1244 sentences from this Polyphone corpus was used for this study (273 speakers, 13092 words, and 22496 syllables). All sentences were automatically labeled using an HMM recognizer developed by [21]. All vowel data used in this paper are based on this labeling (see Table 1). In contrast to our single speaker corpus, all monosyllabic function words carried syllable stress in this corpus.

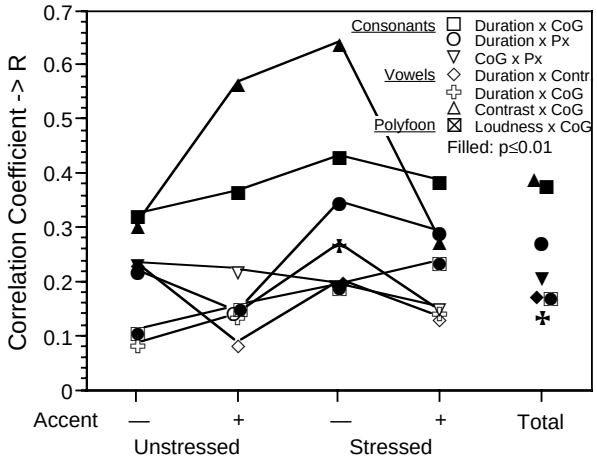


**Figure 1:** Consistency of measurements. Correlation coefficients between individual factors of the three token sets, Consonants and Vowels from the single speaker corpus and Vowels from the Polyphone corpus. CoG: Spectral Center of Gravity (in semitones), Px: $\log_2$(Perplexity) plotted is –R. Loudness in sones. Filled symbols represent statistically significant correlation coefficients ($p \leq 0.01$)

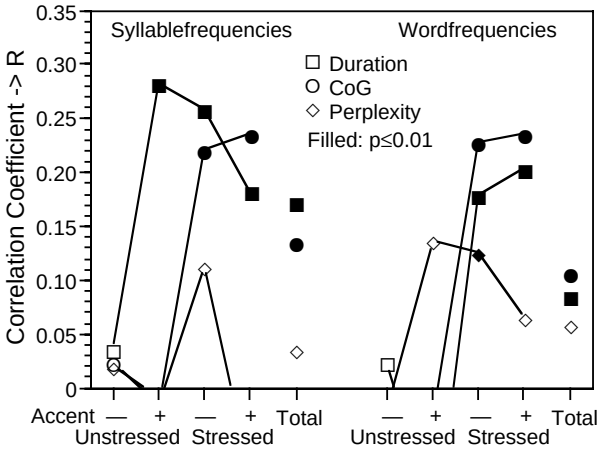


**Figure 2:** Correlation between consonant reduction and frequency of occurrence of syllables and words. All tokens from the single speaker corpus. CoG: Spectral Center of Gravity in semitones, Perplexity: $\log_2$(Perplexity), plotted is –R. Filled symbols represent statistically significant correlations ($p \leq 0.01$). CELEX syllable and word frequencies in $-\log_2$(relative frequency), both were correlated ($R = 0.230$, $p \leq 0.01$).

Frequency of occurrence of all words and syllables was determined from a CELEX word list based on a 30 million word, Dutch text corpus. Frequency of occurrence was always used as $-\log_2$(relative frequency). Duration and two measures of spectral reduction are used as acoustic measures of reduction for our single-speaker corpus: maximal Spectral Center of Gravity (CoG, i.e., the "mean" frequency in semi-tones, weighted by spectral power) and Vowel contrast, the $F_1/F_2$ distance to the center of vowel reduction (300, 1450 Hz vowels only) in semitones. These measures have been shown to be related to both reduction as used here and intelligibility [12][13][14][17][18]. The $\log_2$ perplexity of the identification responses to single consonant tokens is used as a measure of unintelligibility, i.e., confusion [18]. The automatically labeled vowels of the multi-speaker Polyphone corpus didn't lend themselves to reliable duration and formant measurements. However, reasonably consistent measurements of the mean Spectral Center of Gravity and Loudness (in sone) could be obtained. Both of these are related to reduction [13][14][15][17].

# 3. RESULTS

To compensate for the large systematic variation in values between our phonemes, we calculated the correlation coefficients after subtracting the individual mean values from each quasi-homogeneous group of phoneme realizations (homogeneous with respect to phoneme identity, speaking style when relevant, syllable stress, and word accent or prominence). The degrees of freedom in the statistical tests were reduced accordingly to compensate for this procedure. A level of significance of $p \leq 0.01$ was chosen to compensate for the number of conditions tested (2 stress levels times 2 accent levels)

**Table 1:** Number of realizations from the single speaker corpus (Consonants/Vowels) and vowels from the Polyphone corpus (Polyphone). Accent: Sentence accent/prominence, Stressed/Unstressed: Lexical stress (see text).

| | Unstressed | | Stressed | | |
|---|---|---|---|---|---|
| Accent | – | + | – | + | Total |
| Consonants | 550 | 180 | 569 | 283 | 1582 |
| Vowels | 812 | 461 | 528 | 224 | 2025 |
| Polyphone | 4435 | 4942 | 9603 | 3516 | 22496 |

We use several measures to assess the amount of reduction in consonant and vowel segments. An important consideration is how well these measures agree. This consistency is expressed in terms of the cross-correlation between measures (Figure 1). It is clear that the CoG and Vowel Contrast to a large extent measure the same aspects of vowel pronunciation. The same can more or less be said of Duration and CoG in consonants. In all other cases, we do see statistically significant correlations between the measured parameters, but their explanatory power seems to be low. As expected, the perplexity (note the use of –R), correlates reasonably well with both consonant duration and CoG. This indicates that these two acoustic measures of reduction are indeed indicative for segmental intelligibility.

## 3.1. Single speaker data

The results for our single speaker data are represented in the figures 2 and 3. Figure 2 shows the correlation for consonants between, on the one hand, Duration, Spectral Center of Gravity, and Perplexity (note that –R is plotted) and on the other hand, syllable-frequency (left) and word-frequency (right). Figure 3 shows the correlation for vowels between, on the one hand, Duration, Spectral Center of Gravity, and Contrast ($F_1/F_2$ distance) and on the other hand, again, syllable-frequency (left) and words-frequency (right). From Figures 2 and 3 it is clear that there is indeed a statistically significant effect of syllable frequency and word frequency even after accounting for phoneme identity, speaking style, lexical stress and sentence accent. However, the variance explained is small, less than 10% for even the strongest factor (Vowel Contrast). Overall, it shows that syllable frequency is more important than word frequency for determining the measured values in unstressed syllables ($p \leq$

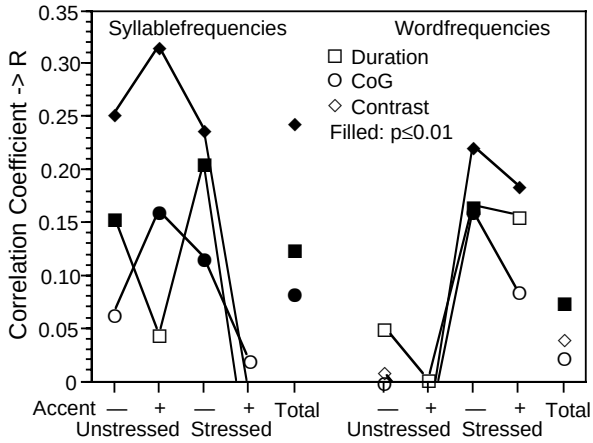


**Figure 3:** As Figure 2 but now for the vowels of the single speaker corpus. Contrast: $F_1/F_2$ distance to the center of reduction (300, 1450). CoG: Spectral Center of Gravity. Both are expressed in semitones. Syllable- and word frequencies were correlated (R = 0.280, p≤0.01).

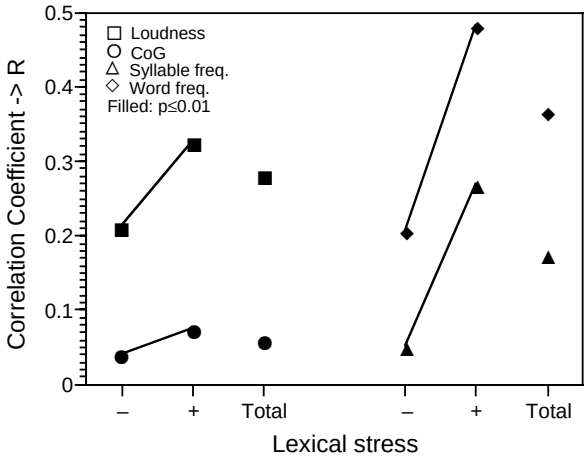


**Figure 4:** Correlation between prominence judgements for words from the Polyphone corpus and each of the acoustic measurements, syllable and word-frequencies. Loudness in sone, Center of Gravity (CoG) in semitones.

0.01, two tailed Sign test on the values for unstressed realizations in Figures 2 and 3). No such difference could be found for the stressed syllables. The correlation between Vowel Contrast and *syllable*-frequency was larger than between Vowel Contrast and *word*-frequency (Figure 3, $p \leq 0.01$, Chi-square test on two correlation coefficients). None of the other differences were statistically significant.

It is clear that, while the ($\log_2$) perplexity of the consonants is strongly correlated with duration and spectral CoG (Figure 1), it cannot really be seen as sensitive to either syllable- or word-frequency. Only for stressed consonants from unaccented words could a relation with word-frequency be shown (Figure 2).

## 3.2. Multi-speaker data

For the multi-speaker data set from the Polyphone corpus, 10 independent judges marked prominent words [15]. These prominence judgements were correlated with both the acoustic measures (Figure 4, left) and word and syllable frequency (Figure 4, right). These correlations were statistically significant for both stressed and unstressed syllables. The strong correlation between word-frequency and prominence judgments can be attributed to the high-frequency mono-syllabic function words with low prominence. The results displayed in Figure 4 again indicate that both Loudness, and to a lesser extent CoG, measure parameters relevant to the perception of reduction and prosody [12][13[14].

All words judged prominent by more than half the judges are considered in-focus (accented). All others are considered out-focus (unaccented). Figure 5 shows the correlation between the acoustic parameters, Loudness and mean CoG, and syllable- and word-frequency. It is clear that, for this multi-speaker corpus too, there are small but statistically significant correlations between acoustic parameters and frequency of occurrence after accounting for vowel identity, lexical stress and word prominence. None of the differences found for the single-speaker data between using syllable- versus word-frequencies (Figures 2 and 3) were statistically significant for the multi-speaker data (Figure 5).

## 4. DISCUSSION AND CONCLUSIONS

Although the correlations found in our data are generally statistically significant, they are also quite weak ($R^2 < 0.1$). Part of this weakness can be attributed to large measuring errors in determining the relevant parameters and the use of inherently noisy automatic labeling. The limited amount of speech available also prevented us from controlling for many of the factors that are known to affect the acoustics of speech, like coarticulation, constituent boundaries, word-length and position in the word. Neglecting these factors is at least partly responsible for the high level of "noise" in our data.

Consistent correlations between acoustic parameters of reduction and word- and syllable-frequency were found for both consonants and vowels after controlling for lexical stress and sentence accent/prominence. For our single-speaker corpus,

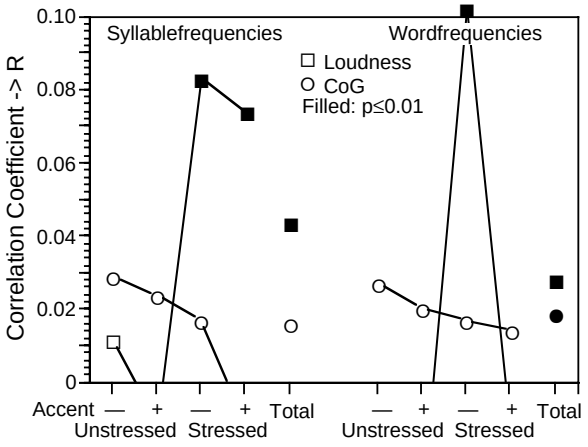


**Figure 5:** As Figure 3 but now for the vowels of the Polyphone corpus. Loudness in sone, Center of Gravity (CoG) in semitones. CELEX syllable and word frequencies in $-\log_2$(relative frequency), both were correlated (R = 0.316, $p \leq 0.01$).

syllable-frequency showed to be a better predictor of acoustic reduction than word-frequency, at least for unstressed syllables.

We can conclude that the efficiency of speech cannot be explained completely by the differential assignment of lexical stress and sentence accent. Frequency of occurrence and possibly predictability in larger constituents are needed to fully understand vowel and consonant reduction. Furthermore, the strength of the syllable-frequency effect in unstressed syllables might be explained by a frequency sensitive "stored versus computed" principle in articulation that supports efficient speaking more or less independent of word-level mechanisms [22].

## 5. ACKNOWLEDGMENTS

We would like to thank dr. Xue Wang for kindly supplying the automatic segmentation software used in this paper. This research was made possible by grants 300-173-029 and 355-75-001 of the Netherlands Organization of Research.

## 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. Damhuis, M., Boogaart, T., in 't Veld, C., Versteijlen, M., Schelvis, W., Bos, L. and Boves, L. Creation and analysis of the Dutch Polyphone corpus. Proc. ICSLP-94, Yokohama, 1803-1806, 1994.
7. Fowler, C.A. "Differential shortening of repeated content words in various communicative contexts", Language and Speech 31, 307-319, 1988.
8. Hunnicutt, S. "Intelligibility versus redundancy - conditions of dependency", Language and Speech 28, 47-56, 1985.
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. 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.
12. Sluyter, A.M.C. and Van Heuven, V.J. "Spectral balance as an acoustic correlate of linguistic stress", J. Acoust. Soc. Am. 100, 2471-2485, 1996.
13. Sluyter, A.M.C., Van Heuven, V.J., and Pacilly, J.J.A. "Spectral balance as a cue in the perception of linguistic stress", J. Acoust. Soc. Am. 101, 503-513, 1997.
14. Sluyter, A.M.C. *Phonetic correlates of stress and accent*, HIL dissertations 15, PhD thesis, University of Leiden, 188 pp, 1995.
15. Streefkerk, B.M., Pols, L.C.W., and ten Bosch, L.F. M. "Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's", Proc. EUROSPEECH'99, Budapest, 551-554, 1999.
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. 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.
21. Wang, X. *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, Ph.D. Thesis, University of Amsterdam, 190 pp, 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.