STOCHASTIC SUPRASEGMENTALS: RELATIONSHIPS BETWEEN REDUNDANCY, PROSODIC STRUCTURE AND SYLLABIC DURATION

Matthew Aylett
HCRC and Dept. Linguistics,
University of Edinburgh
email: matthewa@cogsci.ed.ac.uk

ABSTRACT

Prosodic Prominence and prosodic boundaries have been shown to effect syllabic durations. However another factor, redundancy, also appears to have a major impact. More common words and words you can easily predict from context (more redundant) tend to be articulated less clearly and so also have a tendency to have shortened syllabic durations.

This paper explores the relationship between measures of redundancy, prosodic structure and syllabic duration of a large corpus of spontaneous speech. Although 50% (r=0.71) of syllable variation is predictable from measures of accentedness, break index and other prosodic parameters, word frequency alone predicts 11% of the duration variation. Combining prosodic information and redundancy measurements improves prediction by 0.75% (r=0.72), suggesting that although redundancy measurements can offer a statistically independent contribution to predicting syllabic duration, prosodic structure implicitly represents most of the variation caused by redundancy.

1. INTRODUCTION

1.1. Motivation and Hypotheses

We appear to have two quite different factors controlling the care with which we articulate speech. On one hand we have a complex prosodic structure which predicts prominence and the chunking of speech and on the other we have complex interactions within the structure of language which make some sections of speech predictable and others less so. Understanding how these factors affect variation in articulation is of great importance for both engineers who wish to design effective speech recognition and synthesis software and also psycholinguistics and phoneticians who wish to understand the human language system. Potentially such an investigation can help refine theories of suprasegmental change and allow us to not only predict articulation variation in the speech stream but use this variation to explore the internal state of a speakers language system.

The central questions this paper will address are:

1. To what extent does a modern theory of prosodic structure account for such changes in the care of articulation, in contrast to some simple measures of redundancy?

2. How much interdependency exists between redundancy measurements and prosodic structure? Can concepts of predictability and prosodic structure be integrated together to offer a stronger predictive framework of changes in care of articulation?

1.2. Prosodic Structure

Theories of prosodic structure concentrate on three distinct though related phenomena:

1. **Prominence:** Some parts of the speech stream stand out more than other parts.
2. **Boundaries:** Speech is split up into chunks which are cued by changes in duration, f0, amplitude and voice quality.
3. **Information Giving:** Changes in prosodic structure can alter the meaning of the message, for example changing a statement into a question.

Laboratory phonetics has found that prominent syllables are more clearly articulated. That is, the segments tend to be longer, the spectral characteristics are more distinct, they are louder and often marked with pitch change. Words with such prominence also tend to be easier for human subjects to recognise when excerpted from context.

In general: Prominence = more care of articulation = more noticeable = easier to recognise

1.3. Redundancy

Prosodic structure clearly affects care of articulation however another factor, redundancy, also appears to have an impact. More common words and words easily predictable from context (more redundant) tend to be articulated less clearly. For example the 'nine' in the phrase 'a stitch in time saves nine' is less clearly articulated than the nine in 'I would like nine please' [10].

Lindblom [11] in his H&H theory suggests that we put only as much effort into articulation as required for the listener to understand. He argues that we tend to under articulate easily predictable (redundant) sections of speech and over articulate a difficult to predict (less redundant) sections of speech.

2. MATERIALS

This work is based on a large corpus of spontaneous task oriented dialogue collected by the HCRC at the University of Edinburgh - the HCRC Map Corpus [7]. The corpus is comprised of about 15 hours of spontaneous speech, 64 speakers and around 200,000 syllables.

2.1. Prosodic Coding

3190 words making up 679 full intonational phrases were coded using GlaToBI [12], a variant of the ToBI tone and break index coding system which was adapted for the Glaswegian accent. Automatic techniques were then used to assign nuclear accent placement as well as syllabic structure and lexical stress to these materials.

2.2. Automatic Coding

The entire HCRC map task is word segmented and transcribed. This allowed the automatic coding of word, syllable and phrase boundaries as well as the coding for lexical stress. The word boundaries were hand segmented. Syllable boundaries (for polysyllabic words) were determined using autosegmentation. A dictionary containing a canonical phonemic representation for each word was used to guess the probable segmental contents of each word. A hidden Markov model (HMM) speech recogniser with a model for each segment already trained from previous speech was used to posit the likely boundaries of each phoneme using the maximal onset principle. The syllabification as present in the CELEX dictionary lookup was then used to determine likely syllable boundaries as well as being used to assign lexical stress to syllables.

2.3. Measuring Duration

A duration model using a combined log distribution for each phonemic segment (as in [4]) was used to produce a normalised duration measurement. It assumed that a change in the duration of a syllable was divided equally among the segments of that word in terms of z-scores for duration. Therefore, the change between a syllable's predicted duration and actual duration could be measured in terms of a single z-score calculated for all of a syllable's segments. This value, called here the 'k-score', was used as a measure of how much a syllable had been 'stretched' or 'compressed' from a citation form.

The predicted duration, d, of any word may be expressed as:

$$d = \sum_{i=1}^{n} \exp^{(\mu(i)+k\sigma(i))} \quad (1)$$

where:

n = the number of phonemes in a word,

k = a constant function of average segment length,

μ = the mean log duration of a segment,

σ = the standard deviation of the log distribution of a segment's duration

One log distribution was used (μ=-2.7478 (64ms) σ=0.5702 (-1 sd=36ms, +1 sd=113ms)) for **all** phonemes, so that there was effectively no differentiation between phonemes. Expected syllable durations therefore depended on how many segments there were in any given syllable (For more detail on this and other duration models based on this approach see [1]).

2.4. Measuring Redundancy

The predictability of a syllable in running speech is dependent on many factors. Without understanding all the dependencies between semantics, syntax, pragmatics and the structure of language any measure or redundancy is an approximation. In this work three measurements were taken:

1. **Log of Word Frequency.** More frequent words should be more easy to predict and thus be more redundant. Each syllable was associated with the COBUILD word frequency of the word it was part of.
2. **Syllabic Trigram Measurement.** Using the BNC national corpus the transition probability of guessing a third syllable on the basis of the first two was calculated. The CMU-Cambridge toolkit was used to calculate trigram probability using good turing and backoff [5]. This measurement will give some idea of predictability produced by frequent sequences of words and the redundancy in later syllables in polysyllabic words. Together with word frequency this measurement gives a more interword sense of redundancy.
3. **Givenness.** Both word frequency and trigram measurements can be regarded as low level measures of redundancy, in that they take no consideration of the meaning in language or of the flow of meaning in a stream of speech. In contrast givenness is related to the introduction of a referent in a dialogue. The more this referent is mentioned the more 'given' it becomes. This final measurement of redundancy measures how many times a referent (in this case a landmark on a map e.g. 'white mountain' 'east lake') has been mentioned.

3. RESULTS

A number of linear regressions were carried out to investigate the extent to which prosodic factors and redundancy factors predicted change in syllabic duration.

3.1. Scope of materials

The number of materials available for different analyses varied depending on the factors considered. All syllables in the corpus are coded for word frequency and trigram probability together with a syllabic duration measurement. Of these 3698 syllables are prosodically coded using GlaToBI. Of these 1553 are also coded for givenness.

3.2. Does Prosodic structure account for duration variation?

Laboratory research has shown that many prosodic features have an effect on the duration of syllables. The factors included here are not exhaustive but do represent the major findings.[1]

1. **Boundary effects:** Phrase final lengthening is a well documented effect in speech. Wightman et al [15] showed that, moreover, these effects extended into other prosodic boundaries such as word and intermediate phrase boundaries. Using a Break Index coding system [13] a strong relationship was found between such an index and the duration of the rhyme of the syllable preceding the boundary. The break index coding used here is the ToBI modified version and is as follows:
   - 0 = No boundary (within word/cliticised)
   - 1 = Word boundary
   - 2/3 = two 'strengths' of intermediate intonational phrase Boundary
   - 4 = full intonational phrase boundary
2. **Prominence:** Prominence is the extent that a sound or syllable stands out from others in its environment. A number of

[1] Effects of syllabic structure, for example the total number of syllables in the word, are not reported here. Including these factors (which have an effect on duration) was complicated by being confounded by frequency effects. A more complex linear regression analysis grouping by number of syllables resulted in similar results.

different factors [9] [6] have been put forward as contributing to prominence and these factors affect duration.

- Vowel Type: Reduced vowels (e.g. schwa) appear less prominent than full vowels. Reduced vowels and their syllables tend to be shorter than syllables with full vowels [8]. For example the /i/ in /sIti/ is more prominent than the /@/ in /Aft@/ although neither are lexically stressed
- Lexical Stress: Lexical stress affects duration independent of pitch accents, which are often associated with lexically stressed syllables [3]. Syllables with secondary stress such as 'mul' in 'multiplication' are treated as stressed.
- Pitch Accents and Spillover: Pitch accents are marked by changes in pitch and a strong impression of prominence. Lengthening occurs on the syllable accented as well as, in some cases, syllables adjacent to the accented syllable (Spillover - my term) [14]. Spillover is affected by word boundary and occurs more strongly to the right of the accented syllable. In this analysis syllables within a word to the left of a pitch accented syllable are given a spillover of 0.04, to the right 0.2 and to the right across a word boundary 0.05. These values are in line with Turk and White's results of a 4,20,5% increase in syllable duration in these contexts.
- Nuclear Accents: Nuclear accents (or primary phrasal stress or sentential accent) are a subset of pitch accents that occur, in English, before an intonational phrase boundary. Although there is is evidence that nuclear pitch accents are perceived as more prominent than other pitch accents, it is unclear whether these accents have different effects on segment duration. However for completeness the nuclear/non-nuclear distinction is retained in this analysis.

A linear regression including these factors predicts 51% of the variance in the normalised duration score (r=0.715). Table 1 shows the independent contribution of these factors, and the independent significance (Maximum Likelihood) that each factor has in this model.

Factor	Var	$p <$
Break Index	17.97%	0.001
Full/Reduced Vowel	00.30%	0.001
Lexical Stress	04.35%	0.001
Pitch Accent	03.32%	0.001
Spillover	02.16%	0.001
Nuclear Pitch Accent	00.01%	NS

Table 1 Contributions to predicting duration change. %Var - The independent contribution to predicting the variance, $p <$ - Significance of the maximum likelihood ratio test.

From Table 1 we can see that most of these factors are deeply interrelated. The strongest contribution by far is from the break index representing the boundary strength following the syllable. All factors have a significant effect on duration variation except the distinction between nuclear and non-nuclear pitch accents.

3.3. Does Redundancy Account for Duration Variation?

Context has been shown to affect articulation. When more contextual information is available, making speech easier for a listener to recognise, talkers have been shown to produce more reduced speech ([10], [11]). The more predictable a word the more reduced it tends to become. This study examines three different measures of redundancy: Word frequency, trigram syllabic predictability and mention of a reference within a dialogue.

3.3.1. Low Level Measurements of Redundancy: Word Frequency and Trigram Syllabic Predictability Low level measurements of redundancy do not take into account the syntactic and semantic structure of an utterance. They are blunt instruments which give an indication of predictability. Word frequency is an example of such a low level factor and it has been proposed as a factor in overall word shortening [2].

Other factors, such as neighbourhood density in the lexicon, together with word frequency can produce a more robust measure of redundancy and its effect on reduction [16]. However in this work we are establishing whether redundancy has an effect on reduction outwith prosodic factors and if so by how much. For this reason, although more complicated measurements of word redundancy have been presented in other work, we will use the less robust but simpler and more theory-independent log word frequency measurement. To augment this measurement and take into account word internal and some phrasal predictability, a trigram syllable measurement will also be examined. This value is the probability of a syllable occurring given the two proceeding syllables. The trigram measurement adds:

- **Within word redundancy:** Word initial syllables have a lower trigram probability than the following syllables.
- **Between word redundancy:** For example, the predictability of 'have' following 'do you' is higher than might be predicted by the word frequency alone.

A linear regression taking into consideration the log of word frequency and trigram measurement predicts 12% of the variation in duration (r=0.35). Both factors are significant ($p<0.001$) but the log of word frequency accounts for most of the predictive power (11%).

3.3.2. The Independent Contribution of Prosodic and Redundancy Factors. A maximum likelihood analysis of prosodic factors against redundancy factors shows that although redundancy factors make a significant contribution to predicting syllabic duration change ($p<0.001$) the contribution is very small (0.75%). Prosodic factors implicitly represent most of the effect of redundancy on syllabic duration change.

3.3.3. High Level Measures of Redundancy: Givenness. In contrast to log word frequency and syllabic trigram measurements givenness represents a higher level of redundancy. In the HCRC Map Task Corpus all mentions of landmarks on the maps used are coded for mention. The more a landmark is mentioned the more given it generally becomes. However only a subset of the materials described above are coded for mention and a large proportion of these will be accented nouns and adjectives (e.g. 'white mountain', 'telephone box') and thus these materials are more homogeneous

in terms of prosodic factors.

Redundancy parameters account for 19% of duration change in these materials with mention contributing a small (0.3%) but significant ($p<0.01$) independent contribution to the predictive power. Across these materials prosodic factors account for 58% of the variance (r=0.76) with redundancy variables contributing an independent 1.5% to the model.

4. DISCUSSION

Just over 50% of the variance in duration is predicted by prosodic factors. This leaves a lot of variance unexplained. Some of this unexplained variance is due to the limitations of the duration model. Segmental identity and phonemic context are not included in the duration model even though these factors are known to affect duration. It was found that making use of segmental identity is complicated by being confounded with lexical structure. For example, nearly all examples of 'th' /D/ occur in 'the' /D@/ [1]. It is hoped that a more sophisticated duration model could be used in future research reducing the noise in the duration measurement and improving the predictive power of the prosodic factors.

It is also important to note that the prosodic factors used are quite sophisticated and have been developed over years of research into this area while the redundancy factors I examined were very simple. Although the independent contribution made by redundancy factors is very small there is a consistently significant effect. There is much scope for developing more complex redundancy measurements both at a low level, for example by including lexical access factors, and at a higher level, by including more sophisticated models of dialogue structure. It is possible that more sophisticated measures of redundancy will contribute a greater independent contribution to modelling duration change. To what extent current prosodic theory can represent such redundancy factors remains to be seen.

5. CONCLUSION

There are benefits in modifying articulation to match predictability in language. It is an efficient use of effort and it lowers the chance that a crucial part of the message is obliterated by noise. However maintaining statistical language models at all levels and calculating redundancy from them as we speak would be a resource intensive (and a possibly intractable) exercise. The results from this analysis suggest that a large proportion of redundancy information is implicitly coded in prosodic structure. This structure appears to act firstly at the level of the lexicon in terms of lexical stress and vowel type and at the phrase level in terms of accenting and break index.

Although prosodic structure does appear to explain most of the duration change predicted by redundancy there is a small but significant independent effect. This implies that prosodic theory, as it stands, may not explain all redundancy effects. If this is the case then either the production system is using redundancy information directly or prosodic theory should be modified so that it does account for these effects. For example, lexical stress could be modified to take into account word frequency so that syllables in rare words were regarded as having stronger lexical stress than stressed syllables in common words. Perhaps these stronger stressed syllables could be regarded as more desirable sites for accent placement than their more common neighbours. In this way suprasegmentals could be connected to stochastic information and be used to produce the redundancy effects we have observed.

REFERENCES

[1] M. Aylett and M. Bull. The automatic marking of prominence in spontaneous speech using duration and part of speech information. In *Proceedings of ICSLP-98.*, pages 2123–6, 1998.

[2] D.A. Balota, J.E. Boland, and L.W. Shields. Priming in pronunciation: Beyond pattern recognition and onset latency. *Journal of Memory and Language*, 28:14–36.

[3] W. N. Campbell. Automatic detection of prosodic boundaries in speech. *Speech Communication*, 13:343–54, 1993.

[4] W. N. Campbell and S. D. Isard. Segment durations in a syllable frame. *Journal of Phonetics*, 19:37–47, 1991.

[5] Philip Clarkson and Ronald Rosenfeld. Statistical language modeling using the CMU-Cambridge toolkit. In *Proceedings of Eurospeech 97*, pages 2707–10, 1997.

[6] Alan Cruttenden. *Intonation.* Cambridge University Press, Cambridge, 1986.

[7] Anne H. Anderson et al. The HCRC Map Task Corpus. *Language and Speech*, 34(4):351–366, 1991.

[8] D.H. Klatt. Linguistic uses of segmental duration in english: Acoustic and perceptual evidence. *The Journal of the Acoustical Society of America*, 59:1208–20, 1976.

[9] Peter Ladefoged. *A Course in Phonetics.* Harcourt, Brace, Jovanovich, New York, second edition, 1982.

[10] P. Lieberman. Some effects of semantic and grammatical context on the production and perception of speech. *Language and Speech*, 6:172–187, 1963.

[11] Björn Lindblom. Explaining phonetic variation: a sketch of the H & H theory. In William J. Hardcastle and Alain Marchal, editors, *Speech Production and Speech Modelling*, pages 403–439. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[12] C. Mayo, M. Aylett, and D. Ladd. Prosodic transcription of Glasgow English: An evaluation study of GlaToBI. In Kouroupetroglou G. & Carayiannis G. Botinis, A., editor, *Proceedings of the ESCA Intonation Workshop*, pages 231–234. ESCA, 1997.

[13] P.J. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. *The Journal of the Acoustical Society of America*, 90:2956–70, 1991.

[14] A. Turk and L. White. Structural influences on accentual lengthening in english. *Journal of Phonetics*, To Appear.

[15] C.W. Wightman, S. Shattuck-Hufnagel, M. Ostendorf, and P.J. Price. Segmental durations in the vicinity of prosodic phrase boundaries. *The Journal of the Acoustical Society of America*, 91:1707–17, 1992.

[16] Richard Wright. Lexical competition and reduction in speech: A preliminary report. *Progress Report, Indiana University*, 21:471–485, 1997.