The Influence of Prosodic Factors on the Duration of Words in British English

Caroline Bouzon & Daniel Hirst

CNRS, Laboratoire Parole et Langage
Université de Provence, Aix-en-Provence, France
caro.bouzon@wanadoo.fr, daniel.hirst@lpl.univ-aix.fr

Abstract

English, like all languages, typically involves a great variability in the duration of the structural units taken into account in the observation (phonemes, syllables, feet, words, etc.). In this paper some prosodic factors likely to influence the duration were tested for their influence in two analyses. In the first analysis, the predominant factor turned out to be the final position in the intonation unit. The second analysis takes a closer look at the influence of stress and accent and of the position in the intonation unit on the relative and absolute duration of each word.

1. Introduction

An important area of research in linguistics is the way in which langage works in real time. Extensive research has been carried out on duration. Klatt's model [10] used eleven rules to account for the variability of phonemic duration in speech. The model takes into account inherent segment duration and applies a percentage increase or decrease defined by the eleven rules. Campbell [4][6] proposed a timing model, described as a multi-level process. This model first predicts the duration of a syllable, and then predicts the duration of each segment belonging to that syllable, respecting the accomodation and elasticity rules. The temporal organisation is therefore located at a higher level than the segment. In the same way, ProSynth [11][13] was developed as a linguistic model for speech synthesis which takes a "rich linguistic structure as central to the generation of natural-sounding speech". In terms of duration, the process is based on joining syllables overlaying one over another (thus involving ambisyllabicity) and on the compression of syllables (or "squish") proportional to their complexity and position in the foot.

There has in general been a tendency to move towards higher-level constituents in a search for increased naturalness of speech synthesis. The global quality of speech synthesis, however, is still considered unnatural and unsatisfactory. Hawkins et al. [7] explain that "the rhythm, intonation and fine phonetic details reflecting coarticulatory patterns are poor". Zellner-Keller [16] notes that, as long as duration is not better investigated, the output of speech synthesis will not be more natural, and this because most authors assign a secondary role to timing. Consequently, further research still needs to be done in the area of the temporal organisation of speech.

Informal experiments synthesising English texts with the word-durations measured from natural recordings convinced us that using prosodic units of a still higher-level than the syllable is likely to result in a considerable gain in the perceived naturalness of synthetic speech. Besides, there are a number of corpora available which have been manually aligned for word boundary – equivalent corpora for syllable and phoneme boundaries are much harder to find.

The data used for this study is taken from the MARSEC corpus (MAchine Readable Spoken English Corpus) [14]. This contains over six hours of speech, divided into eleven different speech styles. It is prosodically transcribed, using a system of tonetic stress marks [15]. The orthographic transcription of MARSEC has been manually alligned with the recordings.

We first converted the MARSEC label-files into "TextGrid" format for use with the PRAAT software [1] by means of a Perl script. The word labels for each file were then checked manually with the acoustic signal, incorporating corrections for those labels which appeared to be clearly misaligned, and abandonning a few files in which the alignment was not consistent with the signal, or in which a portion of the signal was not labeled at all. A few other files were also not used because of overlaps of the different speakers or because of noise.

In the rest of this paper we look at the influence of some prosodic factors on the duration of words in the Marsec corpus. We describe two statistical analyses of the data. Both aim at observing the influence of prosodic factors on the duration of words and their usefulness in predicting these durations. The first analysis (described in detail in [3]) aimed at examining the relative influence of as many prosodic factors as possible by means of a classification and regression tree. In the second analysis the influence of the most important factors were examined in greater detail by means of an analysis of variance.

2. Analysis I: CART

2.1. Prosodic factors

One factor which obviously has considerable influence on the duration of a word is its phonemic content. In order to factor out this influence we calculated for each word the predicted duration obtained by the sum of the mean values of each phoneme in the word (using values from [4]). We then used the absolute error (in ms) with respect to the predicted duration as dependent variable for the analysis. In the rest of this paper we refer to this variable as the 'lengthening' of the word.

We next determined eight prosodic factors which have been described in the literature as influencing duration and which we could derive automaticallly from the prosodic labels of the corpus.

These factors (analysed into features following [8]) were

- presence of a stress, of a pitch-accent or of a pitch-glide
- direction, complexity and width of the pitch-glide when present
- position (initial, medial or final) of the word in the intonation unit
- strength of the minor or major intonation boundary after the word when present

We included in the analysis the prosodic characteristics not only of each word but also of the preceding and the following words. Finally, we added the number of syllables in the word, providing a total of 25 factors.

2.2. Statistical analysis

The aim of this first analysis was to try to predict the lengthening of each word according to the different prosodic factors given above, and thus to determine the most influencing factors. For this, we used a classification and regression tree, using the CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation) program [9]. Since CRUISE is designed to predict discrete classes rather than continuous variables, we converted the degree of lengthening of each word into a scale from one to five.

We then ran CRUISE with each of the eleven parts of the corpus, the output being a regression tree. Each node of the tree represents one prosodic feature, each branch determines whether the feature applies or not, and the terminal nodes of the tree give the predicted lengthening.

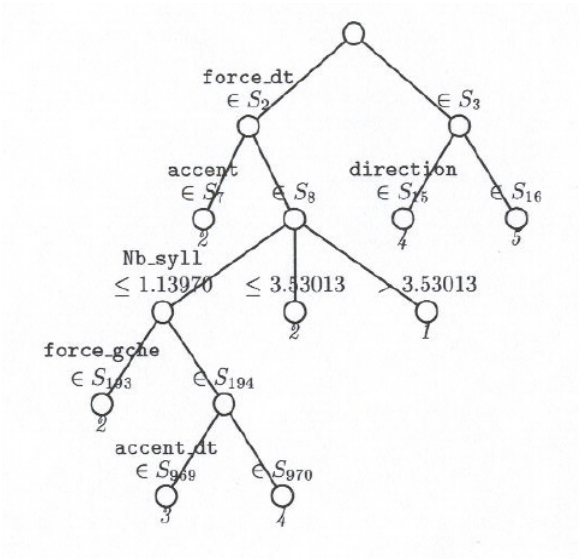

Figure 1: Example of a regression tree (from [3]). force_dt refers to the presence/strength of a boundary after the word, accent is stress, direction the direction of the glide, Nb_syll, the number of syllables in the word, force_gche the presence/strength of a boundary before the word and accent_dt the presence of a stressed word imediately after it.

2.3. Results

In the sample regression tree shown in figure 1, the most important factor was the presence of a boundary after the word, but there is, in this case, no difference between a minor and a major boundary, both being grouped into the second part of the tree. The second most important factor with non-final words is the presence of a stress, this node being divided into two branches (presence or absence of stress), etc. This tree enables us first of all to set up a hierarchy of factors according to their influence, and secondly to obtain a prediction of the lengthening for each final node in such a context.

Each of the eleven speech styles was analysed in the same way. For nine of them, the most important factor was the final lengthening with the presence of a major or minor boundary after the word. This was the most important prosodic factor in all our results. With the exception of one file, however, there was no difference between the influence of a minor and a major boundary. Another important factor which stands out from the results is the presence of stress: this appears as the major factor in two cases (corresponding to poetical and lithurgical readings) where we can assume that rhythmic organisation is more important than in less formal styles.

The number of syllables in the word was also an important factor. Monosyllabic words were lengthened more than bi- or tri-syllabic words which in turn were lengthened more than polysyllabic words. Other factors apparent in the results (but not systematically predominant), were the lengthening of a word when immediately followed by a stressed word (as noted by [2]), the shortening of words after an intonation boundary, and the direction and width of the glide. The complexity of the glide did not appear in our results as being important, whereas we had expected that it would at least be as important as the direction and the width of the glide. Campbell, in a study of the timing of syllables in the SEC (the original version of MARSEC) [5] also reported no significant difference between syllables with simple pitch movement and syllables with complex pitch movement. Finally, we noted that the prosodic characteristics of adjacent words appear to have no significant influence on lengthening.

These results thus confirm the importance of final lengthening, before a major or a minor intonation boundary, and enable us to obtain a hierarchy of the various factors, and to visualise their influence as a first step to predicting the word durations.

In our second analysis we used this data to further examine the influence of the most important factors, that is stress/accent and the position of the word in the intonation unit.

3. Analysis II - Anova

This second analysis took a closer look at the specific influence of stress, accent and the position in the Intonation Unit. We looked particularly for any significant difference between minor and major boundaries, since this factor did not appear as determinant in the CART analysis. We also investigated the relation between unstressed, stressed and accented words in terms of their duration and lengthening, and their interaction with position in the intonation unit.

3.1. Data

We first reorganised the data from the first analysis. We decided to test the effect of stress/accent (words were coded as either *unstressed*, *stressed* or *accented*) and the influence of the position in the intonation unit in terms of the presence of a minor or major boundary before or after the word in question (coded as 5 levels L(eft)-major, L-minor, none, R(right)-minor R-major).

As we noted above the number of syllables in a word seems to be an important factor in determining its degree of lengthening. Attempts to include this factor in the analysis of variance failed however since, despite the very large number of words analysed, there were insufficient data to cover all combinations of the different factors. One way to take into account the number of syllables more indirectly would be to work with the relative (%) lengthening of a word rather than its absolute lengthening as we used in analysis I. Comparing the F-scores for Anovas for relative and absolute lengthening (table 1) showed that the significance of both accent and boundary was considerably greater for both factors although the significance of the interaction between the two factors was marginally better for relative lengthening.

Table 1. F-scores for the effect of accent and boundary on relative and absolute lengthening of word durations.

	Relative lengthening	Absolute lengthening
accent	F=87.420	F=248.374
boundary	F=733.415	F=1138.866
accent*boundary	F=16.143	F=10.338

3.2. Results

We did several ANOVA analyses on the data to test the effects of the various factors enumerated above, and their interaction. We first obtained the simple effects of each factor on the duration and the lengthening of words, and then the interaction of the ryhtmic and boundary factors in terms of duration and lengthening.

As expected, for both types of data (duration and lengthening), the influence of the factors "accent" and "boundary" was highly significant ($p<0.0001$). In the same way, the difference between "unstressed", "stressed" and "accented" on the one hand, and between all sorts of boundary on the other hand was highly significant ($p<0.0001$).

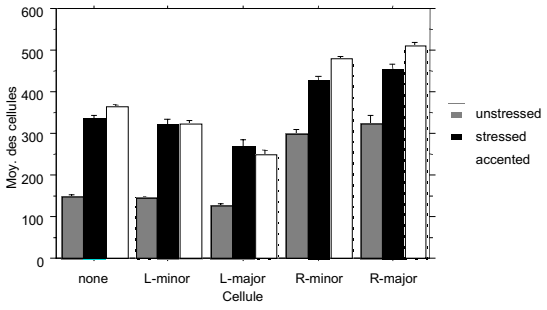

Figure 2: Results for the duration of words (ms)

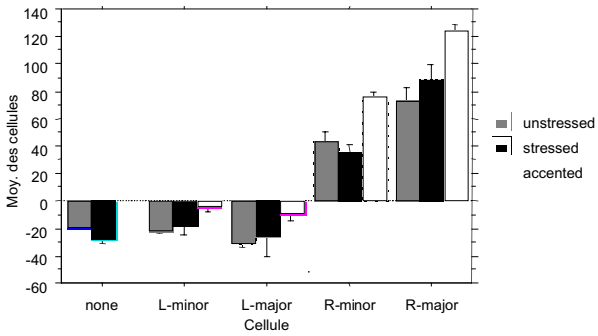

Figure 3: Results for absolute lengthening (ms)

It stands out clearly from the results on figures 2 and 3 that words fall into two groups: initial and medial words on the one hand, final words on the other hand. Both their duration and lengthening are influenced by these two categories of position.

3.2.1. Initial and medial words

First of all, initial and medial words, whether they are unstressed, stressed and/or accented, are considerably shorter than corresponding final words. For example, if we look at *stressed* words, their average duration in initial or medial position is 300ms (lengthening: –24ms), and their average duration in final position is 450ms (lengthening: +62ms).

Both the duration and the lengthening of initial and medial words follow the scale:

none > *L-minor* > *L-major*

That is medial words are longer, and therefore more lengthened, than initial words. This is true for *unstressed* and *accented* words. For stressed words the lengthening is as follows:

L-minor > *L-major* > *none*

with very little difference between *none* and *L-major*. Stressed words in medial and L-major positions are shorter than stressed words in L-minor position.

The difference between L-minor and L-major is highly significant ($p<0.0001$), and indeed, there is an obvious influence of the type of initial boundary, the duration of a word is shorter after a major boundary. Similar results are obtained for the lengthening data: shortening is greater with L-major than with L-minor.

3.2.2. Final words

The duration of final words is much longer than that of initial and medial words, and so is their lengthening. Both duration and lengthening follow the scale:

R-major > *R-minor*

this difference being much stronger for lengthening.

The duration of final words increases according to the scale:

accented > *stressed* > *unstressed*

It might appear from figure 1 that there is hardly any difference between the respective duration of unstressed, stressed and accented words, whether in R-minor or in R-major context. The difference is, however, highly significant ($p<0.0001$), and if we compare these with the lengthening of the categories in the same position, we can see that there is actually a considerable influence of the type of final boundary. As opposed to initial and medial positions, the difference between R-minor and R-major is maximised in terms of

lengthening. It is interesting to see that this difference is highly significant, although it did not appear as one of the significant categories in the CART analysis.

3.2.3. Unstressed, stressed and accented words

Once again, we find two sub-categories among words, but the results are rather different for duration and lengthening. The *duration* of stressed words is closer to that of accented words than to that of unstressed words. The *lengthening* of stressed words is closer to that of unstressed words than to that of accented words.

The explanation for this difference would seem to be that a major factor in difference of duration between stressed and unstressed words is the phonemic content of the words: unstressed words tend to have fewer phonemes and shorter vowels than stressed words.

This is interesting since it means that when we neutralise the external factors of the number of syllables and phonemes, the relation between unstressed, stressed and accented words is different. We indeed notice that, unlike for word duration, stressed words are always more shortened (or less lengthened) than accented words. However, the relationship between unstressed and stressed words is less regular in terms of lengthening. Indeed, unstressed words tend to be shorter than stressed words, except for words in initial and R-minor position. These remarks reflect the opposite differenciation between unstressed, stressed and accented words.

4. Conclusions

We used two sorts of statistical analyses, CART and ANOVA. The CART analysis enabled us to formulate a hierarchy of the degree of influence of some of the factors. However, regression trees tend to leave minor factors aside. While both types of analysis are capable of handling considerable quantities of data, CART analyses have the advantage that they can cope with large numbers of parameters (25 in our first analysis) whereas for ANOVA these need to be reduced to a smaller number of factors (just two in our second analysis). The CART analysis was consequently useful as a data-exploration tool allowing us to reformulate more precise hypotheses which can then be tested using ANOVA.

The results from this analysis convince us that despite limitations inherent in the fact of working with duration at the level of the word, rather than the foot, syllable or phoneme, a obvious advantage of being able to work with larger quantities of data might well offset these. A quantitative estimate of the quality of prediction of duration on the basis of the word, together with a comparative evaluation of the perceived quality of synthetic speech predicted in this way will need to be addressed in future research. We limited our study to the information from boundaries and accentuation, but this study could naturally be extended to other parameters with the same methodology.

References

[1] Boersma, P.; Weenink, D., 1992-2001. Praat. A system for doing phonetics by computer. http://www.praat.org

[2] Bolinger, D., 1963. Length, vowel, juncture. Linguistics 1, 5-29.

[3] Bouzon, C., 2001. Influence des facteurs prosodiques sur la durée des mots en anglais britiannique contemporain. Mémoire de DEA, Université de Provence.

[4] Campbell, N., 1992. *Multi-Level Timing in Speech*. Ph.D. thesis, University of Sussex (Exp. Psychol.).

[5] Campbell, N., 1996. Speech Timing in the SEC. In *Working with Speech. Perspectives on Research into the Lancaster/IBM Spoken English Corpus*, G. Knowles; A. Wichmann; P. Alderson (eds) :. London: Longman, 214-233.

[6] Campbell, N., 2000. Timing in Speech: A Multi-level Process. In *Prosody: Theory and Experiment*, M. Horne (ed.). Dordrecht: Kluwer, 281-334.

[7] Hawkins, S.; House, J.; Huckvale, M.; Local, J.; Ogden, R., 1998. ProSynth: An integrated prosodic approach to device-independent, natural-sounding speech synthesis. *Proceedings of International Conference on Spoken Language Proceeding, 1707-1710*.

[8] Hirst, D.J., 1977. Intonative features. A syntactic approach to English intonation. The Hague: Mouton Publishers.

[9] Kim, H.; Loh, W-Y., 2001, CRUISE, User Manual, *Technical Report 989*, March 3, 1998, revised November 10, 2001, Department of Statistics, University of Wisconsin, Madison. http://www.wpi.edu/~hkin/cruise/

[10] Klatt, D., 1987. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America* 82, 737-793.

[11] Local, J.; Ogden, R., 1997. A Model of Timing for Nonsegmental Phonological Structure. In *Progress in Speech Synthesis*, J.P.H. Van Santen; R.W. Sproat; J.P. Olive; J. Hirschberg (eds). Springer, 109-121.

[12] *MARSEC.* http://midwich.reading.co.uk/research/speechlab/marsec/marsec.html

[13] Ogden, R.; Local, J.; Carter, P., 1999. Temporal Interpretation in ProSynth, a Prosodic Speech Synthesis System. *Proc. XIVth International Congress of Phonetic Sciences,* 1059-1062.

[14] Roach, P.; Knowles, G.; Varadi, T.; Arnfield, S., 1993. MARSEC: A machine readable spoken English corpus. *Journal of the International Phonetic Association*, 23 (2), 47-53.

[15] Roach, P., 1994. Conversion between prosodic transcription systems: "Standard British" and ToBI. *Speech Communication 15*, 91-99.

[16] Zellner-Keller, B., 1996. Structures temporelles et structures prosodiques en français lu. Revue française de linguistique appliquée, vol. 1, 7-23.