

Isochrony and Prosodic Structure in British English

Caroline Bouzon & Daniel Hirst

CNRS UMR 6057, Laboratoire Parole et Langage
Université de Provence, Aix-en-Provence, France
{bouzon; hirst}@lpl.univ-aix.fr

Abstract

This paper attempts to translate two phonological models of prosodic structure into quantitative predictions which can be empirically tested on a large corpus of spoken English. Specifically the Abercrombie/Halliday model of the stress-foot is compared to the Jassem model of (narrow) rhythm unit and anacrusis. The data analysed was a five and a half hour corpus of spoken English (Aix-Marsec). Preliminary results from this analysis suggest that the Jassem model is in nearly all cases superior to the Abercrombie/Halliday model, i.e. that it is the narrow rhythm unit and not the foot which is the essential component of the rhythm of British English. The data suggest furthermore that there is no specific lengthening for stressed syllables.

1. Introduction

One of the challenges in speech research is to account for the different rhythmic patterns observed in language by the formulation of theoretical models of phonological structure. Despite the abundant literature, there is today no general consensus on the subject.

One of the basic hypotheses behind rhythmic models is that of *isochrony*, i.e. the organisation of speech into portions perceived as being of equal or equivalent duration. There are two interpretations to this hypothesis: strict isochrony expects the different elements to be of exactly equal duration. Weak isochrony claims that there is a tendency for the different elements to have the same duration; hence, a constituent containing five sub-constituents, for example, will be less than five times as long as a constituent containing only one sub-constituent. Both involve a compression of the sub-constituents for the constituents to have similar duration, but less for weak isochrony.

The term *isochrony* has generally been reserved for the higher level constituents such as the syllable and the stress-group. It is, however, worth noting that the same principle can equally well be expected to apply at all levels. Thus if phones are grouped into syllables, we might well expect a syllable with only one phone to be shorter than a syllable with two phones, but not twice as short.

The principle of isochrony has led to the distinction of two types of rhythm [19]: in the first, the rhythm is created by the regular occurrence of the stressed syllables of an utterance, and in the second, it is the syllables themselves which create an impression of regularity. Pike coined the terms *stress-timed* and *syllable-timed* respectively for these two rhythmic types.

The most frequently used model of English rhythmic structure is one in which phones are grouped into syllables (or into the intermediate sub-syllabic constituents of onset, nucleus and coda), syllables into feet and feet into intonation units, even though the terminology varies from one author to the other. Between the phone and the foot, the syllable is generally considered the basic component of rhythmic structure. In this way, Campbell [6] proposed a

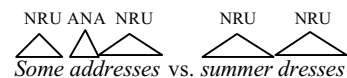
two-level model that predicted segmental duration by a top-down process of accommodation that first predicts the duration of the syllable from a small number of discrete phonological characteristics and then fits the phones into this duration.

The definition of the foot generally adopted is that of Abercrombie [1], i.e. a foot starts with a stressed syllable and includes all the following unstressed syllables up to, but not including, the next stressed syllable. Therefore, his definition of the foot does not take into account word boundaries. Stress feet are characterised by the principle of weak isochrony defined earlier in this introduction; in other words the duration of feet is not strictly proportional to the number of syllables they contain, instead syllables (and consequently phonemes) tend to be compressed when they are more numerous.

Several years before Abercrombie presented this model, Jassem [11] had claimed that the rhythmic organisation of English was based on two units; first, the *Narrow Rhythm Unit* (NRU), which consists of one stressed syllable and any number of following unstressed syllables *belonging to the same word*. Secondly, all the other unstressed syllables which are not part of the NRU belong to the *anacrusis* (ANA). Note that for many other authors, the term anacrusis is reserved for ANAS in initial position of intonation units; these initial units will be subsequently referred to here as *initial anacrusis*. The following example illustrates Jassem's model:



According to this model, NRUs tend to have a similar duration whatever the number of syllables there is. On the contrary, ANAs are uttered as quickly as possible and their duration therefore linearly increases in function of the number of syllables. This model allows Jassem to account for such minimal pairs as:



which are phonemically and accentually identical but rhythmically different. It is interesting to note that few researchers have explicitly adopted Jassem's model of rhythmic structure although authors such as Abercrombie [1] and O'Connor [18] were obviously strongly influenced by it.

Both models claim that all the syllables in a rhythmic unit (foot or NRU) tend to have the same duration, whether they are stressed or unstressed. However, most other authors like [6] and [13] have claimed that there is a significant and independent lengthening of stressed syllables.

In this paper, we attempt to translate these different phonological hypotheses into quantitative predictions which can then be tested statistically with a large prosodically annotated corpus of spoken British English. Our aim is not to make actual predictions of durations but to find evidence for phonological structure.

The following section is devoted to the description of the experimental evidence for rhythmic structure in British English in the literature. The third section then details the different statistical analyses used to test the two rhythmic models. Finally, the fourth section presents and comments the results of the statistical analyses.

2. Experimental evidence for rhythmic structure

2.1. The syllable

Hill *et al.* [10] observed a certain degree of compression on the level of both the foot and the rhythm unit, but no such compression was observed on the level of the syllable. Syllables consisting of just one segment had systematically greater mean segment duration but this is due to the fact that such segments were necessarily syllabic nuclei and were consequently intrinsically longer.

Campbell [6] noted that with the exception of syllables containing just one segment, there seemed to be a positive correlation between mean segment duration and the number of segments in the syllable. He showed, however, that this apparently paradoxical effect was an artefact due to the fact that longer syllables were more likely to be stressed. Once the factor of stress was taken into account, a systematic effect of compression was found within the syllable.

2.2. The foot

Much literature has been devoted to attempting to provide experimental evidence of isochrony in speech at the level of the foot [18]; however, little physical evidence in favour of strict isochrony has been found ([22][16][9]). Nevertheless, despite the lack of experimental evidence in favour of such regularity, there remains an intuition that the duration of inter-stress intervals is equivalent, if not equal; this led some authors to interpret this as meaning that isochrony is essentially a perceptual phenomenon ([2][17]).

Faure *et al.* [9] showed that the duration of feet in English increases linearly in function of the number of syllable components, thus clearly contradicting the principle of strict isochrony as well as that of weak isochrony. They therefore concluded that there is no syllable compression and that the impression of regularity comes from the difference of duration between stressed and unstressed syllables as well as from the occurrence of pitch accents on stressed syllables.

Eriksson [8], however, pointed out that a linear increase in the duration of feet is not necessarily contradictory with the idea of syllabic compression: he produced evidence that both stressed and unstressed syllables adjust to match the duration of the inter-stress interval to which they belong. This is corroborated by data from [6], as both stressed and unstressed syllables show a tendency to shorten in longer feet.

The distinction between languages on the basis of their stress-timed or syllable-timed characteristics has largely been contested. Roach [21] showed that Abercrombie's binary classification does not hold and that languages are not either stress-timed or syllable-timed but rather share timing characteristics with both classes. Instead of a categorical distinction, [5] and [7] showed that languages share characteristics of stress-timed or of syllable-timed languages but to different degrees.

2.3. Rhythm unit and anacrusis

As was mentioned in the introduction, Jassem's model claims that NRUS tend to be isochronous, as opposed to ANAS which are uttered as quickly as physiologically possible. Jassem *et al.* [14] confirm the hypotheses of

Jassem's model as they conclude that there is a significant tendency towards approximately isochronous NRUS, but did not observe such an effect in anacruses.

3. Statistical analyses

3.1. Data

The data used in this study is the Aix-Marsec corpus ([3][4]), which is a five and a half hour corpus of natural sounding British English. It has been automatically phonetically transcribed and automatically aligned at the phoneme, sub-syllabic constituent, syllable, foot, NRU/ANA, word and minor and major intonation unit levels. In order to avoid erroneous phoneme durations due to automatic alignment errors, it was decided that phoneme durations less than 15ms. and greater than 500ms were excluded from the analyses.

In order to avoid bias due to the effect of the identity of phones, we follow [6] in using as dependent variable for this study not the raw phone duration, but the normalised duration (z_d) obtained by subtracting the mean duration for the phoneme from the raw value, and then dividing the result by the standard deviation for the phoneme (formula 1).

$$z_d = (d - \text{mean}_d) / \text{sd}_d \quad (1)$$

In order to carry out the z -transform of the data we calculated the mean and standard deviation for each phoneme from the whole corpus and then calculated the z -transform using formula (1).

It is possible that this procedure introduces some bias into the results since it may be that the means and standard deviations vary somewhat from one speaker to another. As a safeguard, we decided to perform all the analyses using both the z -transformed data and the raw data. When results are similar for both analyses, we may reasonably conclude that they are sufficiently robust to be reliable.

For the analyses where the dependent variable was the duration of higher level constituents than the phone, we used the raw duration in ms. of the constituent.

3.2. Hypotheses

Throughout this section, the hypotheses discussed concern the effects of different prosodic constituents on the duration of the phone, the syllable, the rhythm unit and the foot. Thus for example we shall be looking at the hypothesis that the duration of a particular unit is (or is not) affected by the number of phones or syllables contained in that unit. In order to avoid unnecessary repetition, the term *complexity* will be used as a cover term to be glossed as 'the number of phones or syllables in the unit in question'.

3.2.1. Syllable and foot structure

According to the principle of strict isochrony, feet should be of equal duration, meaning that the complexity of feet should not have any effect on their duration. According to the principle of weak isochrony, the foot as a phonological unit does effect the duration of its constituents which tend to be shorter when they are more numerous; we, therefore, expect a significant negative effect of the complexity of a foot on the duration of its constituents.

Isochrony at the level of the foot does not exclude (strong or weak) isochrony at the level of the syllable and the phoneme. In case of strong isochrony, the duration of the constituent should not be influenced by its complexity. On the other hand, in case of weak isochrony, we should expect to find a significant negative effect of the complexity of a constituent on the duration of its sub-constituents.

3.2.2. Rhythm unit and anacrusis

Jassem's hypothesis is that the NRU is the domain of isochrony, not the foot as claimed by Abercrombie. The NRU being part of the foot, we might expect any effect in the NRU to be found in the foot as well, only weaker. On the contrary, phones in the ANA are uttered as quickly as possible, which means that they should not be affected by the complexity of the foot or the ANA.

3.2.3. Prosodic structure and stress

In both Jassem's and Abercrombie's models, the length of the phones should not be related to the stressed or unstressed nature of the syllable to which they belong. But this contrasts with the assumption that stressed syllables - and therefore phones - are longer than their unstressed counterparts.

4. Results

The statistical tools we use in this study are linear regression, and correlation and analysis of variance, all statistical analyses being carried out using the R-project statistics package [20].

4.1. Strict isochrony

Table 1 presents the results of the linear regression analyses testing the hypothesis of strict isochrony. The dependent variable (rows) is the duration in ms. of the different prosodic constituents: syllable, foot, NRU, ANA, initial ANA, word and intonation unit (IU), and the independent variable (columns) is the complexity of the unit as measured by the number of phones, syllables and feet in the constituent, where appropriate.

Table 1: Results of the linear regression for the strict isochrony hypothesis. Dependent variable (rows) was duration of the constituent in ms. and independent variable (columns) the number of sub-constituents. The figures correspond to the adjusted R^2 (regular typeface) and the regression slope coefficient (italic).

	phones	syllables	feet
syllable	0.289 <i>69.773</i>	-	-
foot	0.427 <i>51.974</i>	0.352 <i>110.919</i>	-
NRU	0.336 <i>51.654</i>	0.228 <i>104.687</i>	-
ANA	0.446 <i>53.960</i>	0.384 <i>122.082</i>	-
initial ANA	0.404 <i>51.605</i>	0.367 <i>119.883</i>	-
word	0.596 <i>69.492</i>	0.483 <i>163.014</i>	-
IU	0.589 <i>68.115</i>	0.575 <i>165.123</i>	0.553 <i>367.600</i>

In all the analyses, the results of the linear regression were highly significant ($p < 0.001$). In all of the analyses the correlation was positive, meaning that at every level the duration of a constituent is proportional to its complexity, i.e. to the number of its sub-constituents. This clearly refutes the hypothesis of strict isochrony at every level of prosodic constituency.

4.2. Weak isochrony

Table 2 shows the results of the linear regression analyses testing the hypothesis of weak isochrony. The dependent variable (columns) is the duration of the subconstituents measured both in ms. and also in z-transformed duration (two consecutive figures), while the independent variable (rows) is the complexity of the constituent as measured by the number of phones, syllables or feet it contains. All of the analyses were highly significant ($p < 0.001$).

The correlation was negative for all of the analyses except one. The correlation between the duration of the phones and the number of phones in the syllable was negative for the duration measured in ms. but positive for the zscore. The apparent compression effect (negative coefficient) on the level of the syllable thus seems to be essentially due to the difference between syllables containing only one phoneme (necessarily a vowel or a syllabic consonant which are on average longer than non-vocalic phones) and syllables containing more than one phoneme. According to the results with the z-score, there is no compression of phonemes in function of the complexity of the syllable (this confirms the findings of [10]).

Table 2: Results of the linear regression analysis for the weak isochrony hypothesis. Dependent variable (columns) was duration of the sub-constituents in both ms.(left) and z-score(right) and independent variable (rows) the number of subconstituents. The figures correspond to the adjusted R^2 (regular typeface) and the regression slope coefficient (italic).

	phones	syllables	feet
syllable	0,000/0,000 <i>-0.632/0.018</i>	-	-
foot	0.020/0.010 <i>-3.139/-0.036</i>	0.019/0.011 <i>-7.223/-0.092</i>	-
NRU	0.034/0.014 <i>-5.773/-0.059</i>	0.024/0.012 <i>-11.745/-0.134</i>	-
ANA	0.009/0.007 <i>-2.769/-0.044</i>	0.003/0.002 <i>-5.084/-0.093</i>	-
initial ANA	0.005/0.002 <i>-1.642/-0.018</i>	0.003/0.002 <i>-3.133/-0.047</i>	-
word	0.005/0.003 <i>-1.624/-0.020</i>	0.004/0.003 <i>-3.380/-0.048</i>	-
IU	0.008/0.007 <i>-0485/-0,008</i>	0.007/0.007 <i>-1.144/-0.019</i>	0,003/0.004 <i>-1.667/-0.033</i>

All the other correlations between the duration of phones and the complexity of the prosodic constituents are negative (foot, NRU, ANA and initial ANA, word and IU) and they can be interpreted as evidence for some degree of weak isochrony (i.e. some compression) at all these prosodic levels.

The largest effects both for R^2 and for the regression slope were observed for the NRU. This seems to imply that the NRU, not the foot, is the essential component of rhythmic structure in English as claimed by Jassem. Jassem's hypothesis about the ANA is not, however, entirely supported by our results as there is some compression taking place in the ANA (even more so in initial ANAs), although there is far less than in the NRU.

4.3. Isochrony and stress

An analysis of variance with zscore as dependent variable and with presence/absence of stress and number of phonemes in the narrow rhythm unit as independent variables was carried out for all phonemes within the narrow rhythm unit to test whether

stress has an effect on phone duration orthogonal to that of the number of phonemes in the NRU. The results showed an extremely significant effect for both stress: $F(1,123747) = 34.135$ and number of phonemes in the NRU: $F(11,123747) = 183.141$. The interaction between the two factors was also highly significant: $F(10,123747) = 8.636$ ($p \ll 0.001$ in all three cases).

The overall mean z-score for phones in stressed syllables was globally much higher (0.121) than that for phones in unstressed syllables (-0.016). The interaction between the factors, however, shows that this is essentially due to the difference of z-scores between phones in stressed syllables containing only one phoneme and the others. For the other syllables the mean z-score for phones in unstressed syllables is as often higher as lower than that for phones in stressed syllables as can be seen from figure 1.

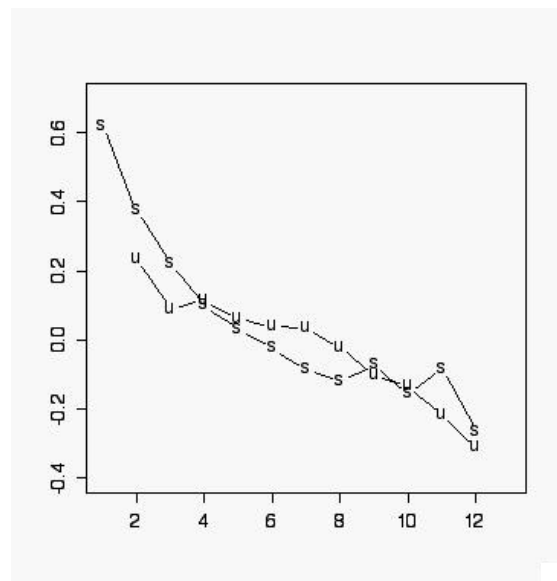


Figure 1. Zscore for phones in stressed (s) and unstressed (u) syllables, as function of the number of phonemes in the narrow rhythm unit (RU).

5. Conclusions

The majority of the preliminary evidence we have examined so far seems to clearly confirm the superiority of Jassem's model over that of Abercrombie and Halliday. The amount of phonemic compression observed within the NRU is clearly far greater than that observed within the anacrusis. Since the stress-foot combines the two units the degree of compression observed there is intermediate between that of the NRU and the anacrusis.

The apparent independent effect of stress on segmental duration appears, finally, to be limited to the case of NRUs consisting of from one to three phonemes. In nearly all other syllables in fact the average zscore appears to be as often higher in unstressed syllables as in stressed syllables. This suggests, then, that Jassem was essentially correct in concluding that it is not stress but the fact of belonging to an NRU which is the essential factor influencing the duration of phonemes.

6. References

- [1] Abercrombie, D., 1967. *Elements of General Phonetics*. Edinburgh : Edinburgh University Press.
- [2] Allen, J.S., 1975. Speech rhythm: its relation to performance universals and articulatory timing. *Journal of Phonetics*, 3, 75-86.
- [3] Auran, C.; Bouzon, C., in press. Phonotactique prédictive et alignement automatique : application au corpus MARSEC et perspective. *TIPA*, 22, 13-44.
- [4] Auran, C.; Bouzon, C.; Hirst, D.J., 2004. The Aix-MARSEC project: an evolutive database of spoken British English. *Speech Prosody 2004*, Nara, March 23-26.
- [5] Bertinetto, P.M., 1989. Reflections on the dichotomy "stress" vs. "syllable-timing". *Revue de Phonétique Appliquée*, vol. 91-93, 99-130.
- [6] Campbell, N., 1992. *Multi-level Timing in Speech*. PhD Thesis, University of Sussex.
- [7] Dauer, R.M., 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11, 51-62.
- [8] Eriksson, A., 1991. *Aspects of Swedish Speech Rhythm. Doctoral dissertation*, University of Göteborg: Sweden.
- [9] Faure, G.; Hirst, D.J.; Chafcouloff, M., 1980. Rhythm in English: Isochronism, Pitch and Perceived Stress. In Waugh, L.R.; van Schooneveld, C.H. (eds): *The Melody of Language. Intonation and Prosody*, 71-80.
- [10] Hill, D.R.; Witten, I.H. & Jassem, W. 1984, *Some results from a preliminary study of British English speech rhythm*. Research Report 78/26/5 University of Calgary, Department of Computer Science.
- [11] Jassem, W., 1952. *Intonation in Conversational English*. Polish Academy of Science, Warsaw.
- [12] Jassem, W.; Hill, D.R., Witten, I.H., 1984. Isochrony in English Speech: its Statistical Validity and Linguistic Relevance. In Gibbon, D.; Richter, H. (eds): *Intonation, Accent and Rhythm. Studies in Discourse Phonology*, 203-225.
- [13] Klatt, D.H., 1987. Review of text-to-speech conversion for English. *Journal of the Acoustical Society of America*, 82, 737-793.
- [14] Knowles, G.; Wichmann, A.; Alderson, P., 1996. *Working with Speech: perspectives on research into the Lancaster/IBM Spoken English Corpus*. London: Longman.
- [15] Knowles, G.; Williams, B.; Taylor, L., 1996. *A Corpus of Formal British English Speech*. London: Longman.
- [16] Lea, W.A., 1974. *Prosodic aids to speech recognition: IV. A general strategy for prosodically-guided speech understanding*. Univac Report No. PX10791. St Paul, Minnesota: Sperry Univac, DSD.
- [17] Lehiste, I., 1977. Isochrony reconsidered. *Journal of Phonetics*, 5, 253-263.
- [18] O'Connor, J.D., 1967. *Better English Pronunciation*. Cambridge: Cambridge University Press.
- [19] Pike, K.L., 1945. *The Intonation of American English*. USA: The University of Michigan Press.
- [20] R-Project: <http://cran.r-project.org/>
- [21] Roach, P., 1982. On the Distinction between "Stress Timed" and "Syllable Timed" Languages. In Crystal, D. (ed.): *Linguistic Controversies. Essays in Linguistic Theory and Practice*, In Honour Of F.R. Palmer, 73-79.
- [22] Shen, Y. ; Peterson, G.G., 1962. Isochronism in English. *University of Buffalo Studies in Linguistics, Occasional Papers*, 9, 1-36.