The rhythm of text and the rhythm of utterances: from metrics to models

Daniel J. Hirst

The rhythm of text and the rhythm of utterances: from metrics to models.

Daniel Hirst

Laboratoire Parole et Langage, UMR6057, CNRS & Aix-Marseille Université
Aix-en-Provence, France
daniel.hirst@lpl-aix.fr

Abstract

The typological classification of languages as stress-timed, syllable-timed and mora-timed did not stand up to empirical investigation which found little or no evidence for the different types of isochrony which had been assumed to be the basis for the classification. In recent years, there has been a renewal of interest with the development of empirical metrics for measuring rhythm. In this paper it is shown that some of these metrics are more sensitive to the rhythm of the text than to the rhythm of the utterance itself. While a number of recent proposals have been made for improving these metrics it is proposed that what is needed is more detailed studies of large corpora in order to develop more sophisticated models of the way in which prosodic structure is realised in different languages. New data on British English is presented using the Aix-Marsec corpus.

Index Terms: speech prosody, rhythm, linguistic typology, metrics

1. Introduction

There has been considerable interest in the last decade in the modelling of rhythm both from a typological perspective (e.g. establishing objective criteria for classifying a language or dialect as stress timed, syllable timed or mora timed) and from a perspective of the evaluation of non standard or deviant varieties of speech such as that obtained from non-native speakers, from speakers with pathological disabilities or from automatic speech synthesis.

Pike [21] suggested that two types of rhythm are found in speech: **syllable-timed** rhythm where syllables give the impression of being of equivalent duration, and **stress-timed** rhythm, where it is the stressed syllables which are perceived as occurring at regular intervals, whatever the number of intervening unstressed syllables. The idea that accented syllables in a language like English are more or less equally spaced in time goes back at least as far as the 18th century [24].

Abercrombie [1] made the further claim that this corresponds to a universal typological distinction and that all languages can be classified into one of two rhythmic classes: syllable-timed languages (such as English, Russian and Arabic) and stress-timed languages (such as French, Telugu and Yoroba). Ladefoged [17] later proposed to add a third rhythmic class, that of *mora-timed* languages, for languages such as Japanese and Tamil where the rhythm is determined by units smaller than the syllable, known as **morae.**

Almost embarrassingly, however, this neat typological distinction of three rhythmic classes did not stand up very well to empirical investigation. A number of experimental studies ([23], [26], [4]) questioned the validity of the typology. Roach, for example, measured the syllable durations for about two minutes of spontaneous unscripted recordings by speakers of each of the six languages mentioned above and found no evidence that the languages classified as stress-timed exhibited any more variability of syllable duration than the languages classified as syllable-timed. He also measured the duration of inter-stress intervals for the six speakers (measuring the duration from the onset of each syllable which appeared to be stressed until the onset of the next one within the same intonation unit). This duration was expressed as a percentage of the duration of the whole intonation unit to compensate for any possible effects of change of *tempo*. The results of this, contrary to what would be predicted by the typological distinction, showed – if anything - a greater variability in the duration of the inter-stress intervals for the so-called stress-timed languages than for the so-called syllable-timed languages. There was, furthermore, no evidence that the duration of inter-stress intervals was any less correlated with the number of syllables which they contained for the stress-timed languages than for the syllable timed languages.

Dauer [8] looked at interstress intervals in data from five languages: English, Italian, Greek, Spanish and Thai. No significant differences in mean or standard deviations were found between the languages. She suggested that the impression of syllable-timing or stress-timing was the result of a combination of factors of phonological structure of the language: number of phonemes, presence of diphthongs, vowel reduction etc., rather than a genuine phonetic characteristic of utterances.

2. Measuring the rhythm of text and the rhythm of utterances

Just when it seemed that the status of the typological distinction was without any measurable empirical basis, work in the area of psycholinguistics brought the distinction back into discussion. Nazzi et al [20] had shown that young infants, including newborns, can discriminate between sentences taken from their own language and sentences taken from a language belonging to a different rhythmic class (stress-timed or syllable-timed), but not between sentences from languages of the same rhythmic class.

Building on this result, Ramus et al [22] identified what they termed "correlates of linguistic rhythm in the speech signal". Using recordings of five sentences spoken by 4 speakers for each language, they first segmented the sentences into "vocalic intervals" and "consonantal intervals", defined as portions of the speech signal containing respectively sequences of only vowels or only consonants.

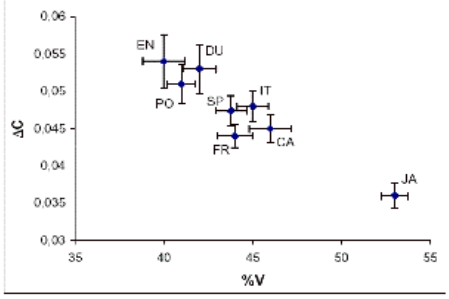

Figure 1. *Duration of vocalic intervals as percentage of total duration (%V) and standard deviation of consonantal intervals (ΔC) for 8 languages: English (EN), Dutch (DU) and Polish (PO), Spanish (SP), Italian (IT), French (FR) and Catalan (CA) and Japanese (JA), showing how the languages group into stress-timed, syllable timed and mora-timed rhythmic classes. Figure from Ramus et al. [22].*

They then calculated two parameters, **%V**: the sum of the durations of the vocalic intervals expressed as a percentage of the total duration of the sentence, and **ΔC** the standard deviation of the consonantal intervals within each sentence. They showed that these two parameters made it possible to separate the three rhythmic classes quite well, as can be seen in Figure 1, reproduced from Ramus et al. [22].

These metrics appear fairly robust - the following figure, for example, replicates the metrics applied to ten comparable sentences (the first two continuous passages from the Eurom1 corpus [7]) as read by one English, French and Japanese speaker.

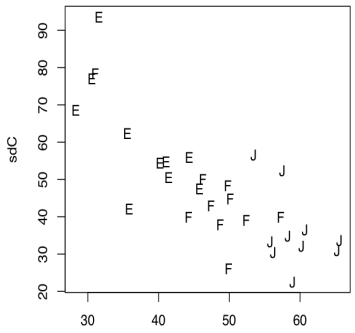

Figure 2. *Ramus' metrics applied to ten comparable sentences in English (E) French (F) and Japanese (J), showing a reasonably good separation of the three languages.*

The claim that this analysis reflects correlates of linguistic rhythm "in the speech signal" is, however, slightly misleading. The measurements rely on a prior linguistic segmentation of the signal into phonemes and a classification of the phonemes into consonants and vowels. Neither of these operations are purely acoustic. It might be wondered whether the parameters represented in Figure 2 are as much an image of the rhythm of the **utterances** themselves, as of the rhythm of the **text** of the utterances as had been suggested by Dauer [8].

Indeed, it can be shown that a similar discrimination can be obtained if, instead of the **duration** of each vocalic or consonantal interval, we measure simply the **number of phonemes** in that interval, as illustrated in figure 3, which shows the number of vowel phonemes expressed as a percentage of the total number of phonemes and the standard deviation of the number of consonant phonemes in each consonant interval for the same ten sentences in English, French and Japanese as in figure 2.

The reason for this is that there is, of course, a high correlation between the percentage of vowel phonemes and the percentage duration of vocalic vs consonantal intervals.

Since English, French and Japanese seem to be fairly well distinguished on the basis of the phonological structure of utterances without any reference to the speech signal, it seems possible, that a fairly large component of Ramus' metric is doing little more than counting the relative number of consonants and vowels in utterances in the different languages. More of the phonemes of French utterances are vowels than in English but less than in Japanese, and the number of consonants in consonantal intervals is less variable in French than in English but more so than in Japanese.

Of course Ramus et al.'s point is precisely that children derive such information about the phonological structure of their native language from the acoustic data. This does however, leave doubts as to the usefulness of this metric for characterising the rhythm of specific utterances in applications such as the evaluation of pathological, non-native or synthetic speech.

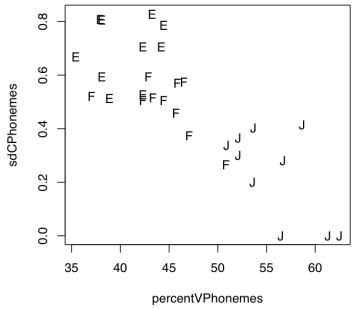

Figure 3: *The percentage of vowels compared to the total number of phonemes (percentVPhonemes) against the standard deviation of the number of consonants in the different intervocalic consonant clusters (sdVPhonemes) for the same ten sentences in English, French and Japanese as in Figure 2.*

Low [18], Low, Grabe & Nolan [19] proposed a "pairwise variability index" (PVI) for consonantal and vocalic intervals, which they showed to be correlated with the typological classification. This metric gave different results when applied to readings of the same text by speakers of different origins (native and non-native), which would not be the case for the phoneme based component of the Ramus metrics as discussed above.

Since these studies, there have been a number of proposals for improved metrics of rhythmicity such as using the coefficient of variation of vocalic and consonantal intervals rather than the standard deviation ([11][25]), relativizing the PVI measure to the number of segments composing each consonantal or vocalic interval (Bertinetto & Bertini's Control/Compensation Index [5]) or applying the PVI on the level of the foot as well as on the level of the syllable (Asu & Nolan [2]).

The usefulness of these various metrics is likely to be dependent on the task for which they are employed but in a sense they can only give us a crude intuition into the way in which rhythmic structures are realised in different languages.

A more profound knowledge about this structure can only come from an explicit **model** of prosodic structure, and of the various factors which influence segmental duration, a problem I have referred to elsewhere ([13]) as "Klatt's unsolved problem":

> *One of the unsolved problems in the development of rule systems for speech timing is the size of the unit (segment, onset/rhyme, syllable, word) best employed to capture various timing phenomena*. Klatt [16] p 760.

3. Modelling rhythmic structures

We saw above that the simplest model of prosodic structure, namely that in a so-called stress-timed language, each interstress interval or stress-foot has the same duration, is clearly in contradiction with empirical evidence.

Faure et al. [10] recorded 11 sentences read by two speakers containing a total of 114 interstress intervals and found an almost perfect linear correlation between the number of syllables and the duration of the interval. They concluded:

> *It is simply not true that stressed syllables are separated by even 'roughly equal' intervals of time. (p 73).*

Since the isochronous stress-foot hypothesis is not valid, an alternative model would be to assume (as these authors did) that accented syllables in English are simply longer than unaccented syllables with a fixed mean duration of e.g. 220 ms for stressed syllables and 140 ms for unstressed syllables.

Eriksson [9] compared published data on interstress intervals for English, Swedish and Icelandic on the one hand and Spanish, Greek and Italian on the other. He found a very

high linear correlation between the duration of interstress intervals and the number of syllables in all these languages. Linear regression showed similar slopes for all the languages analysed of about 100ms per unstressed syllable. But he also found an intercept of about 200 ms for 'stress-timed' languages and of about 100 ms for 'syllable-timed' ones. (p 43)

Eriksson demonstrated, however, that a linear increase in duration of the interstress interval does not necessarily imply a constant duration of stressed and unstressed syllables, since it is also compatible with a non-linear decrease in the duration of both accented and unaccented syllables.

But of course this sounds like a very strange explanation. Why should two non-linear functions combine to produce a linear function?

Most of the results I have cited so far depend on rather small quantities of data. In the rest of this section, in order to try to answer this question, I look at data from the Aix-Marsec corpus ([3]). The data from this corpus consists of 5.5 hours of recordings of spoken British English. The data was automatically aligned with a phonematic transcription within manually defined intonation units. Manual inspection showed that alignment errors generally resulted in very large errors with phoneme values either very large (over 1 second) or very small (typically 19 ms). For this reason our analysis was carried out only on phonemes with a duration between 25 and 500 ms, the others being considered as errors. A small number of feet containing more than 6 syllables were also eliminated from the analysis.

Inspection of the mean duration of feet as a function of the number of syllables reveals a strikingly linear correlation as can be seen in Figure 4 where the continuous line represents the linear regression line.

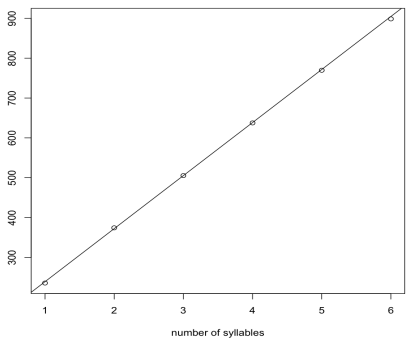

Figure 4: *Mean duration of stress feet in the Aix-Marsec corpus as a function of the number of syllables in the foot. Continuous line corresponds to the linear regression on the same data.*

As Eriksson observed for his data, the mean duration of the stressed and unstressed syllables decreases non-linearly with the number of syllables in the foot (Figure 5).

Abercrombie [1] in his account of the stress foot as the basic rhythmic unit in spoken English made two strong and, in fact, quite controversial claims about the foot.

Firstly, he stated specifically that the foot does not take into account word-boundaries – a phrase like "It's almost impossible" would be grouped into feet as follows

/ ɪts|ɔːlməʊstɪm|pɒsɪbɬ /

where (|) corresponds to the foot boundary – the initial phonemes /ɪts/, under this analysis, do not belong to a foot.

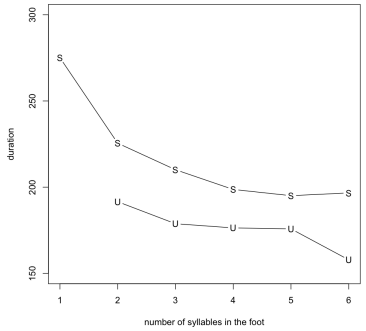

Figure 5: *Mean duration of stressed and unstressed syllables as a function of the number of syllables in the foot.*

The second claim made by Abercrombie was that the relative quantity of syllables in a foot in English is not directly dependent on stress. This seems in complete disagreement with work on speech synthesis in a number of different languages. Klatt [16], for example, included among his 11 rules for predicting segmental duration a rule stating that unstressed segments are considerably shorter and more compressible than stressed segments. Similar claims have been implemented into a considerable number of speech synthesis systems for different languages.

Other writers have claimed that word boundaries play an important role in the determination of the rhythmic structure of utterances. Jassem [14][15], for example, proposed a **Narrow Rhythm Unit** for English which, just like the foot, starts with a stress, but, unlike the foot, ends at the following word-boundary. Any syllables not part of a narrow-rhythm unit, form an **anacrusis** which, according to Jassem, is pronounced "*as quickly as possible*". The phrase "It's almost impossible" in Jassem's model has a structure like:

/ (ɪts)|ɔːlməʊst(ɪm)|pɒsɪbɬ /

where the phonemes in brackets are those of the anacrusis.

Like Abercrombie, Jassem suggested that the duration of phones is not directly affected by the stressed/unstressed nature of the syllable but that the duration of the unit is spread out more or less equally among the phones, with a certain degree of compression so that a unit containing six phonemes, for example, would be longer than one with only three phones but not twice as long.

Hirst & Bouzon [13] found that, as predicted by Jassem, word boundaries **do** play an important role in the rhythmic structure of English. Their data further suggested that there is no specific compression at the level of the syllable. Their most surprising result was the confirmation of Jassem's prediction that once we know whether a given phone belongs to an anacrusis or to a narrow rhythm unit, and in the latter case once we know the number of phones in that unit, the fact that the phone occurs in a stressed or an unstressed syllable has no specific effect on its duration.

Hirst & Auran [12] proposed an algorithm which takes into account a complete utterance and calculates a value of tempo for the utterance giving a scalar weight to each rhythm unit in the utterance, following the results found by Eriksson. The duration of the rhythm unit in this model is a function of the number of phonemes in the unit plus a lengthening factor which is independent of the size of the rhythm unit. This model did not, however, investigate the way in which this lengthening factor is spread out among the different phonemes.

Figure 6 shows that the duration of the Narrow Rhythm Unit in our corpus is indeed highly correlated with the number of phonemes it contains.

One of the major factors influencing phoneme duration is the identity of the phoneme itself. To neutralise this, following

Campbell [6], the z-score of the phoneme duration was used instead of the raw duration.

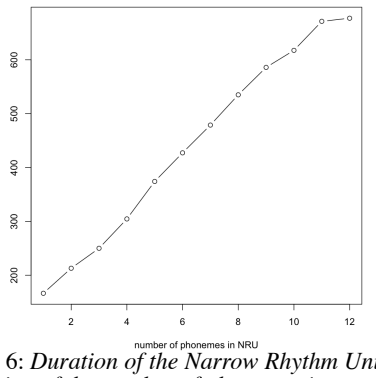

Figure 6: *Duration of the Narrow Rhythm Unit as a function of the number of phonemes it contains*.

A second well known effect is that of final lengthening, which in [13] was shown to effect in particular the final 3 phonemes of an intonation unit. In the following, to neutralise final lengthening, the last three phonemes of each intonation unit were excluded from the analysis.

Analysis of variance revealed a highly significant effect of both number of phonemes in the NRU and position of the phoneme within the NRU. When phonemes were coded as NRU Initial, Medial and Final, analysis of variance once again revealed highly significant differences ($p<2.2e-16$) between the three positions, with mean values of z-score:

initial 0.245 *medial* -0.118 *final* 0.073

For NRU initial phonemes, the size of the NRU was non-significant ($F(1, 23309) = 2.5$ $p = 0.1522$). For NRU medial phonemes, also, analysis of variance on index of phoneme within the NRU was also non-significant.

4. Conclusions

The model suggested by this data is simply a lengthening of the intial and final phoneme of each Narrow Rhythm Unit. Since each Narrow Rhythm Unit contains one initial and one final phoneme this could explain why, as Eriksson had noted, the lengthening appears to be uniform across the NRU regardless of the number of phonemes it contains. It remains to be seen the way in which this NRU intial and final lengthening interacts with the quantal lengthening proposed in [12], as well as with the final lengthening observed in Intonation Unit final position.

In future work we intend to explore these interactions in more detail and to investigate how far the model of prosodic structure we are unfolding for British English can apply to other languages, in particular to French, for which we plan to investigate comparable data.

5. References

[1] Abercrombie, D. 1967. *Elements of General Phonetics.* Edinburgh: Edinburgh University Press.

[2] Asu, E.L.; Nolan, F. (2006) Estonian and English rhythm: a two-dimensional quantification based on syllables and feet. In *Proceedings of the Third International Conference on Speech Prosody*, Dresden, Germany

[3] Auran, C.; Bouzon, C.; Hirst, D.J. 2004. The Aix-MARSEC project: an evolutive database of spoken English. In Proceedings of the Second International Conference on Speech Prosody, Nara, Japan, 561-564

[4] Bertinetto, P. M. 1989. Reflections on the dichotomy 'stress' vs. 'syllable-timing'. *Revue de Phonétique Appliquée* 91–92–93: 99–130.

[5] Bertinetto, P. M.; Bertini, C. 2008. On modeling the rhythm of natural languages. *Proceedings of the Fourth International Conference on Speech Prosody*, Campinas, Brazil.

[6] Campbell, N. 1992. *Multi-level Timing in Speech*. Ph.D. thesis, University of Sussex.

[7] Chan, D.; Fourcin, A.; Gibbon, D.; et al. 1995. EUROM - A Spoken Language Resource For The EU. *ESCA. EUROSPEECH'95.* Madrid, September 1995.

[8] Dauer, R. M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* **11**: 51–62.

[9] Eriksson, A. 1991. *Aspects of Swedish Speech Rhythm*. Gothenburg Monographs in Linguistics, 9. Gothenburg University: Department of Linguistics.

[10] Faure, G.; Hirst, D.J.; Chafcouloff, M. 1981 Rhythm in English: isochronism, pitch and perceived stress. in L.R. Waugh & C.H. van Schooneveld (eds) T*he Melody of Language*, Balstimore: University Park Press, 71-79.

[11] Fernandez-Cruz 2000. *Analyse phonologique et phonétique du portugais brésilien parlé par les communautés noires d'Amazonie*. Doctoral thesis, Université de Provence, France.

[12] Hirst, D.J.; Auran, C. 2005. Analysis by synthesis of speech prosody: the ProZed environment. *Proceedings of Interspeech/Eurospeech 05*, Lisbon. 3225-3228.

[13] Hirst, D.J.; Bouzon, C. 2005. The effect of stress and boundaries on segmental duration in a corpus of authentic speech (British English). *Proceedings of Interspeech/Eurospeech 05.*, Lisbon. 29-32.

[14] Jassem W. 1950. Indication of speech rhythm in the transcription of educated Southern English. le Maitre Phonétique. III/92, 22-24.

[15] Jassem, W. 1952. *Intonation in Conversational English*.Warsaw, Polish Academy of Science.

[16] Klatt, D.H. 1987. Review of text-to-speech conversion for English. *JASA* 82 (3): 737-793.

[17] Ladefoged, P. 1975 A Course in Phonetics. New York: Harcourt Brace Jovanovitch.

[18] Low, E.L., 1998. *Prosodic Prominence in Singapore English.* Doctoral dissertation. University of Cambridge.

[19] Low, E.L.; Grabe, E.; Nolan, F., 2000. Quantitative characterisations of speech rhythm: syllable-timing in Singapore English. *Language and Speech* 43, 377-401.

[20] Nazzi, T.; Bertoncini, J.; Mehler, J.: 1998 Language discrimination by newborns: Toward an understanding of the role of rhythm. *J. exp. Psychol. hum. Perception Performance* 24: 756–766

[21] Pike, K.N. 1945. *The Intonation of American English.* Ann Arbor, University of Michigan Press.

[22] Ramus, F.; Nespor, M.; Mehler, J. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 72, 1-28.

[23] Roach, Peter 1982. On the distinction between 'stress-timed' and 'syllable-timed' languages. in Crystal (ed.) *Linguistic controversies. Essays in linguistic theory and practice*. Edward Arnold, London.

[24] Steele, J. 1779. *Prosodia Rationalis.* London: J. Nichols, Reprinted Georg Olms Verlag, Hildesheim, 1971.

[25] Wagner, P.; Dellwo, V. 2004. Introducing YARD (Yet Another Rhythm Determination) And Re-Introducing Isochrony to Rhythm Research. In: *Proceedings of the Second International Conference on Speech Prosody*, Nara, Japan.

[26] Wenk, B.J. & Wioland, F. 1982. Is French really syllable-timed? *Journal of Phonetics* 10: 193-216.