Measuring speech rhythm

Dafydd Gibbon, Ulrike Gut

Fakultät für Linguistik und Literaturwissenschaft
Universität Bielefeld, Germany
{gibbon,gut}@spectrum.uni-bielefeld.de

Abstract

We address the question of rhythm variation in typologically different languages (English, said to be a stress-timed language, Ibibio, said to be a syllable-timed language) and in different varieties of the same language (British and Nigerian English). Attempts to find correlates of different rhythm types in the acoustic signal have so far not been particularly successful. We examine a number of previous studies, in search of a promising measure of rhythm, and select a recently developed measure (the Pairwise Variability Index of Low & Grabe), with minor modifications and the addition of a binary classifier for focal and nonfocal components of rhythm units. The measure and the classifier are implemented as a software tool which takes esps/waves+ label files as input, and generates statistics on durations, duration differences, the rhythm measure, and a classification of the syllables in the labeled utterance. The results show distinct differences in stress-timing and syllable-timing between Ibibio and English.

1. Speech rhythm

The concept of speech rhythm is a much discussed issue in phonetics and phonology. Impressionistic accounts agree that the languages of the world differ in their rhythm (syllable-timing, stress-timing, mora-timing) but attempts to capture these differences acoustically have so far been unsatisfactory. The aim of the work reported in this paper is to find a reliable quantitative measure of rhythm for use in classifying typologically different languages, in this case English, Ibibio, (a Nigerian tone language), and the English of Nigerian speakers. We outline selected treatments of rhythm in the extensive literature on the topic, and discuss results based on a duration based heuristic measure.

To start with, we provide a general definition of rhythm as a basis for more detailed discussion:

> Rhythm is the recurrence of a perceivable temporal patterning of strongly marked (focal) values and weakly marked (non-focal) values of some parameter as constituents of a tendentially constant temporal domain (environment).

The temporal patterning can be binary alternation, or a more complex rhythm as found in metrical poetry and music. The terms 'focal' and 'nonfocal' rhythm constituent are assigned to value sequences such as $high \frown low$ pitch, $pitch-peak \frown pitch-trough$, $long \frown short$ syllable, $vowel \frown consonant$ segment, We therefore identify two factors in the temporal organisation of rhythm: the internal focal-nonfocal rhythmic pattern, and the external rhythmic environment. The rhythmic environment, whether the syllable, the foot, or some other unit, is sometimes called *rhythm unit*, *rhythmic unit*, etc.

1.1. Rhythm measurement methods

The rhythm of the languages of the world have been divided into stress-timed and syllable-timed at least since [1]. Stress-timing refers to regularly recurring beats or stresses as in English or German and syllable-timing to regularly recurring syllables as in French. A third category has also been postulated: mora-timed languages such as Japanese and Estonian, based on the regular recurrence of subsyllabic timing units.

It is not always clear what 'timing' refers to: is it related to internal durational properties of the rhythmic pattern, or to the duration of rhythmic environment (syllable, foot, etc.)? The term 'isochrony', which has been the subject of much controversial discussion (cf. [2], [3]), is used to refer to the equality of duration of instances of the rhythmic environment (rather than the internal structure of the rhythmic pattern, which may be based on other prosodic factors than duration alone).

After detailed discussion of the controversy, Campbell demonstrates in [4] that a complex hierarchical timing model with segmental, syllabic and higher level components is necessary, and that isolated consideration of a single timing parameter, or even of compensatory lengthening effects within and between syllables, is inadequate.

Many researchers who have tried to find acoustic correlates for stress-timing and syllable-timing have compared the absolute length of the foot in languages such as English, the foot being defined as the interval beginning with a stressed syllable up to but not including the next stressed syllable [3], [5] (the term 'foot-timing' may therefore be preferable to 'stress-timing' in this context). However all researchers found considerable variation in foot duration, and consequently the notion of isochrony in English speech rhythm has been rejected by most researchers. It was variously suggested that isochrony is only a tendency in production, or a perceptual category, or a syntactic and phonological construct.

A number of studies comparing foot-timing in English showed methodological drawbacks. With the exception of [3] sentence-final feet were not excluded even though English is known to show final syllable lengthening. In general, studies also tended to disregard unstressed syllables before the first beat (but cf. [6]). With notable exception of [3], all studies use read speech. Claims for spontaneous speech, a distinct register from read speech, can therefore not be made. In most of the older studies, statistical analysis of the data is usually minimal, restricted to the calculation of means and occasionally variance of foot duration. In addition, the lack of normalization across speakers or speech tempi suggests that generalisation of findings over speech styles and speakers is not possible.

Roach [3] introduced a simple and precise method: Tone unit duration is divided by the number of feet in the tone unit, yielding a hypothetical ideal foot duration under the assumption of perfect isochrony. Actual measurements of foot duration were compared with the predicted value and the percentage deviation was calculated. Roach showed that the variance of the percentage deviation is higher in English than in French, Telugu and Yoruba, i.e. in languages classified as syllable-timed by [7], which is contradictory to expectations.

Other studies have concentrated on the relation between syllables in speech. Abercrombie [7] suggests that stress-timed rhythm shows considerable variation in syllable length, whereas syllable-timed rhythm implies syllables of roughly the same length, but there are few instrumental investigations of this claim. Roach [3] calculated the standard deviation of syllable durations in three stress-timed (English, Russian, Arabic) and three syllable-timed languages (French, Telugu, Arabic), and found no significant differences: English 86ms, Russian 77ms, Arabic 76ms, French 75.5ms, Telegu 66ms and Yoruba 81ms (see also [8]).

1.2. Rhythm measures

Two measures with attractive properties in the context of our study have been discussed in the literature.

Scott & al. [9] discuss a quantitative measure for 'rhythmic irregularity'. This is an open-ended, normalised measure which is calculated pairwise for all intervals in a sequence, i.e. globally over the whole sequence:

$$\Sigma_{i \neq j} \left| \log \frac{I_i}{I_j} \right|$$

The absolute value of the logarithm ensures that the correct ratio is found, independent of the order of division. We will refer to this measure as the RIM (Rhythmic Irregularity Measure). The more similar the durations of units are, the closer the RIM value gets to 0.

In a more recent approach by Low & Grabe (cf. [11], [12]), a *Pairwise Variability Index* (PVI) is proposed. The PVI is calculated pairwise for adjacent vowels, with normalised duration differences.

$$PVI = 100 \times \Sigma_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| / (m-1)$$

That is, for a sequence of units (e.g. syllables) of length m, the average of the absolute differences between adjacent units is calculated; the differences are normalised by dividing each difference by the average duration of the syllables in the pair. In this way, differences in tempo across the utterance, and between utterances, are reduced.

Unlike the RIM, the PVI is therefore locally, not globally normalised. Like the RIM, the PVI is lower when the length of vowels in adjacent syllables is close, and the more they they differ, the higher the index. The RIM is open ended as irregularity increases, but the PVI can be shown to range between the limits of 0 and 200:

- Case 1: the syllables in each pair have equal length. In this case, the difference between adjacent syllables is 0, the normalised difference is 0, the average multiplied by 100 is 0. This is the lower limit.
- Case 2: the syllables in each pair have very different length, with the length of one approaching zero to all intents and purposes, and the other being much longer. Then the difference will be approximately the same as the duration of the longer syllable, and the average duration will be approximately half this, so the normalised difference will be approximately 2 and the average multiplied by 100 is 200. This is the upper limit.

1.3. Empirical assumptions

This approach embodies a number of assumptions about properties of the domain being measured and about the relation embodied in the definition of rhythm measures:

1. Rhythmic durational differences can be associated with vowels rather than syllables.
   (a) This would appear to imply that other factors (such as phonemically contrastive conditioned length) which determine vowel length can be ignored, e.g. that *pretty Sally tickled Tim* would behave just like *tiny Davey fired Joan*.
   (b) It would also appear to imply that the length of consonant clusters is irrelevant, whether in the onset or the coda, and whether constant or arbitrarily variable.
2. Rhythm tends towards binary long-short (strong-weak) alternation. However, sequences like *these three large bears swam too soon* and *Jonathan Appleby merrily trundled along with a tune on his lips*, both with 7 stressed syllables (depending on speaking style) show that this is not the case. The RI would presumably come close to syllable timing in the first case and have an intermediate value in the second.
3. A single parameter is sufficient for the measurement of rhythm. This is not obvious, bearing in mind the definition of rhythm just given. Two measures are presumably required: first, a strong-weak measure (e.g. relative syllable duration), second, a measure for the temporal window. Otherwise we have a duration model, but not a rhythm model.

Recent approaches to rhythm measurement therefore include reference to the syllable structure fo the language [5] and to the consonantal intervals between vowels [13], [16].

2. Rhythm Ratio

We decided to use a modification of the PVI with a more conventional range between 0 and 100, which we refer to as *Rhythm Ratio* (RR):

$$RR = 100 \times \Sigma_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{d_k + d_{k+1}} \right| / (m-1)$$

There is no substantive difference between the PVI and the RR; in fact, RR = RVI/2.

In order to permit rapid comparisons of different measures, the RR algorithm was implemented as a program which takes an esps/waves+ or a Praat annotation file as input and outputs a range of statistics (min, max, range, sum, mean, median, mode, variance, standard deviation, standard error, confidence intervals) about durations, absolute differences between adjacent duration pairs, normalised differences between adjacent duration

pairs, as well as z-scores for durations and duration pair differences. In the implementation we used the following formulation of the RR:

$$RR = 100 * \Sigma_{k=1}^{m-1} \frac{d_i}{d_j}/(m-1)$$

where $dur_i = d_k$ & $d_j = d_{k+1}$ if $d_k < d_{k+1}$, else $dur_j = d_k$ & $d_i = d_{k+1}$
This RR was applied to both syllable and vowel durations. We also used a binary classifier for predicting typical focal and nonfocal units, which we refer to as the Focal-Nonfocal Measure (FNM). So far we have tested simple classifiers with boundaries based on a number of criteria such as mean syllable length, the mid value between minimum and maximum syllable length, and are planning to use further types.

3. Application of the method to corpus analysis

The RR and FNM defined above were applied to syllable duration in the comparison of two languages assumed to have different speech rhythm (stres-timing vs. syllable-timing), (British) English and Ibibio. Ibibio is a tone language of the Lower Cross family and is spoken in Nigeria. In addition, British English and Nigerian English were chosen for the inter-variety comparison, as Nigerian English has been suggested to be syllable- timed rather than stress-timed (cf. [14], [15]). In the comparison of the different varieties of English, the RR of both vowel duration and syllable duration was calculated.

The Ibibio corpus consists of two speakers reading 10 sentences of at least 12 syllables each. The speakers were recorded in Germany and in England. The Nigerian English corpus includes four speakers reading and three speakers retelling a story of 273 syllables. All Nigerian English speakers have a university degree and speak Standard Nigerian English. Three speakers were recorded in Germany and one in Nigeria. The British English data comprises one speaker reading and retelling the same story. He was recorded in Germany. Syllable length and vowel length (British and Nigerian English only) were measured using esps/waves+ and Praat. For Ibibio, the RR of syllable length was calculated for each sentence; for British English across the whole story or retelling. Final syllables, i.e. syllables occurring before a pause were excluded in the British and Nigerian English data in order to avoid artifacts due to final syllable lengthening.

4. Results

We discuss the results of three production experiments. In each case basic statistics for syllable durations are tabulated, and a number of specific further measures are given where relevant, including the syllable-based RR and vowel-based RR results, and the typical focal and nonfocal unit lengths (FNM).

4.1. Ibibio vs. British English

Table 1 compares the syllable RR for Ibibio and British English. Since the Ibibio corpus consists only of read speech, values for the semi-spontaneous British English speech, the retelling, were not included. The RR for Ibibio is 16 for both speakers (mean value across all sentences) and 26 for the British English speaker. The range of syllable length is 0.242 sec and 0.169 sec for the Ibibio speakers and 0.631 sec for the British English speaker. The standard deviation of syllable durations is 77 ms and 57 ms for the Ibibio speakers and 93 ms for the British English speaker. Clearly, syllable durations in Ibibio are more similar than in British English. The range of syllable durations in Ibibio is about a third of the range in British English.

Table 1: Ibibio vs. British English, read speech, with syllable numbers, duration means and ranges, Focal-Nonfocal Measure for duration classes (FNM), standard deviation, standard error and Rhythm Ratio (RR).

	n	mean	range	FNM	SD	SE	RR
BE	273	0.175	0.631	0.131;0.22	0.093	0.005	26
IE 1	129	0.165	0.242	0.175;0.225	0.077	0.015	16
IE 2	129	0.234	0.169	0.138;0.204	0.057	0.025	16

4.2. British English vs. Nigerian English

Table 2 lists the RR for syllables and vowels, and the standard deviation of syllable duration for British English and Nigerian English speakers in the reading condition.

Table 2: British English vs. Nigerian English, read speech, with syllable numbers, means, standard deviations, standard errors and RR for syllables and vowels.

	n	mean	SD	SE	RR	vowel RR
BE	273	0.175	0.093	0.005	26	30
NE B	304	0.237	0.137	0.007	26	18
NE E	268	0.192	0.104	0.006	28	32
NE I	293	0.213	0.128	0.007	32	30
NE G	290	0.181	0.107	0.006	32	28

Syllable duration: Successive syllable duration is very similar in British English and Nigerian English, with some variation between speakers. The standard deviation of syllable durations even tends to be higher for the Nigerian English speakers than for the British English speaker.
Vowel duration: Comparing the ratio of successive vowels (the Low & Grabe method), speaker B shows a smaller ratio (18) than all other speakers, whose ratio is around 30. Comparing the ratio of successive vowel durations, differences between the Nigerian English speakers become apparent.

Whereas speakers E, I and G show a similar vowel RR to the English speaker, Nigerian speaker B's vowel RR is considerably smaller.

4.3. Read speech vs. semi-spontaneous speech

Comparing semi-spontaneous speech with read speech, differences between the British English speaker and the Nigerian English speakers can be found (Table 3).

Table 3: British vs. Nigerian English, semi-spontaneous speech.

	n	mean	SD	SE	RR	vowel RR
BE	224	0.178	0.094	0.006	24	32
NE E	108	0.195	0.131	0.012	30	22
NE I	293	0.206	0.126	0.007	32	28
NE G	174	0.172	0.089	0.006	26	24

Syllable duration: Whereas the British English speaker shows a slightly smaller RR between successive syllables in semi-spontaneous speech than in read speech, the Nigerian English speaker E shows a higher RR. In other words, the British English speaker produces a greater difference between syllable

durations in read speech than in semi-spontaneous speech, for speaker E it works the other way round.
Vowel duration: Looking at the vowel duration differences, calculated as the ratio, however, the Nigerian speakers E, I and G show a smaller ratio in semi-spontaneous speech than in read speech, whereas the British speaker shows a slightly increased difference in semi-spontaneous compared to read speech. This means that for speakers E, I and G the proportion of vowel duration within the syllable changes. In semi-spontaneous speech, successive vowels are more similar than in read speech, for speaker E even despite the fact that syllable durations become less similar.

5. Discussion and prospects

Our measurement of RR demonstrated that Ibibio rhythm is indeed more syllable-timed than British English speech rhythm. The ratios between adjacent syllable durations in Ibibio were smaller than those in British English. Equally, the range of syllable durations and their standard deviation was greater in British English than in Ibibio.

Our results thus stand in contrast to those of Roach [3], who failed to find a significant difference between English and Yoruba (also a Nigerian tone language, but not closely related to Ibibio). We replicated his measurement of a standard deviation of syllable durations of 77 ms and 57ms for Ibibio (he found 77ms for Yoruba), but whereas he found a standard deviation of 86 ms in English, we found a standard deviation of 93 ms. Clearly, our data are preliminary and need to be confirmed with more speakers and other speaking styles such as semi-spontaneous and spontaneous speech.

The comparison between British English and Nigerian English speech rhythm showed that, in read speech, the ratios of successive syllable durations between the two varieties are very similar. Equally, standard deviation of syllable duration was very similar in British English and Nigerian English. One speaker's tendency towards syllable-timing might be indicated in her smaller RR for successive vowels.

Comparing semi-spontaneous with read speech, differences between British English and Nigerian English were found. Whereas in British English, the ratios of syllable duration decreased, for one Nigerian English speaker the ratios increased. Moreover, the ratios of vowel differences decreased for all Nigerian English speakers, which means that, in semi-spontaneous and read speech, the temporal relationship between consonants and vowels is different, the effect observed for speaker B in the read condition. Whereas syllable duration does not change, vowel differences become smaller. This finding replicates Ramus, Nespor & Mehler's [13], Dauer's [5] and Grabe & Low's [16] findings that both the the vocalic and consonantal proportions determine speech rhythm. Despite the normalisation factor in the RR formula, differences between individual speakers became apparent. This suggests that speech rhythm is not exclusively determined by the language spoken but shows individual strategies, too.

The results show that use of finely tunable rhythm measures turn up results which, on the one hand, correspond better to well-established intuitions about the temporal organisation of speech than previous studies have led us to believe. On the other hand, use of such measures also turns up surprising details, for example in connection with individual variation. Apart from the intrinsic interest of results of this kind for typological lingusitics and cognitive psychology, we anticipate applications both in training programs and language identification applications. Ongoing work based on the work reported here is concerned with broadening both the quantitative basis for the analyses in terms of statistical analyses and their qualitative basis in terms of further speech styles and speakers.

6. References

[1] Pike, Kenneth L. (1945). *The Intonation of American English*. Ann Arbor: U Michigan Press.

[2] Lehiste, Ilse (1977). Isochrony reconsidered. *Journal of Phonetics* 1977,5:253-263.

[3] Roach, Peter (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal, ed., *Linguistic Controversies*. London: Edward Arnold pp. 73-79.

[4] Campbell, W. Nicholas (1992). *Multi-level speech timing control*. Ph.D. thesis, U Sussex.

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

[6] Hill, D. R., Wiktor Jassem & I. H. Witten (1978). A statistical approach to the problem of isochrony in spoken British English. Calgary, Alberta: Computer Science Technical Reports 1978-27-6.

[7] Abercrombie, David (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.

[8] Hoequist, Charles (1983). Durational correlates of linguistic rhythm categories. *Phonetica* 40:19-31.

[9] On the measurement of rhythmic irregularity: a reply to Benguerel. *Journal of Phonetics* 14:327-330.

[10] Allen, George D. (1975). Speech rhythm: its relation to performance universals and articulatory timing. *Journal of Phonetics* 3:75-86.

[11] Low, E. & Grabe, E. (1995). Prosodic patterns in Singapore English. Proceedings of the International Congress of Phonetic Sciences, Stockholm, 3, 636-639.

[12] Watson, Ian, Esther Grabe & Brechtje Post (1998). The Acquisition of Prosody in Speech Production: English and French. Project Report B(RE)9714.

[13] Ramus, Franck, Marina Nespor, & Jacques Mehler (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73,3: 265-292.

[14] Jibril, Munzali (1986). Sociolinguistic variation in Nigerian English. *English World-Wide* 7,1: 47-74.

[15] Bamgbose, Ayo (1971). The English Language in Nigeria. In John Spencer, ed., *The English Language in West Africa*. London: Longman, pp. 35-48.

[16] Grabe, E. & Low, E.-L. (to appear). Durational Variability in Speech and the Rhythm Class Hypothesis. To appear in Papers in Laboratory Phonology 7.