Is Speech Rhythm an Intrinsic Property of Language?

Jason Brown, Eden Matene

University of Auckland

jason.brown@auckland.ac.nz, emat730@aucklanduni.ac.nz

Abstract

Different languages have traditionally been classified into different rhythm types. Most studies of rhythm have either implicitly or explicitly accepted that rhythm is an inherent property of a language. This study aims to determine whether rhythm is an intrinsic property of languages, or whether rhythm is an epiphenomenal byproduct of the phonotactic structures of a given stimulus. The question that this project addresses is to what extent the phonological properties of a language can be correlated with rhythmic categories; for instance, whether a language has consonant clusters, makes use of contrastive tone, has complex syllables, exhibits vowel reduction, etc. and whether these can be linked to what kind of rhythmic profile a language fits into.

Index Terms: speech rhythm, duration, phonotactics

1. Introduction

Speech rhythm has been a hotly debated area in recent years. While it has been firmly established that different languages have different rhythmic profiles, or yield different rhythmic impressions, it is unclear what the source of these profiles or impressions is.

While there has not been a completely satisfactory or successful method of quantifying speech rhythm, the most widely accepted method has been through durational metrics. These are generally of two types: interval metrics and variability indices. Interval metrics [1] involve the following: \( \%V \) (the percentage of the total duration of a stretch of speech consisting of vocalic intervals), \( \Delta C \) (the standard deviation of the durations of all consonantal intervals in a speech stretch), and \( \Delta V \) (the standard deviation of the durations of all vocalic intervals in a speech stretch). It has been found that certain metrics were sensitive to speech rate [2]; thus, normalized versions of \( \Delta C \) and \( \Delta V \) were developed. These metrics measure the coefficient of variation of consonantal (VarcoC) and vocalic (VarCoV) intervals such that they are normalized for speech rate [3].

In contrast to these interval measures, variability indices have been proposed [4], [5]. The pairwise variability indices (PVIs) are intended to provide a measure of the pairwise comparison of any two subsequent intervals. In other words, PVIs are designed to capture patterns; that is, variations between intervals, as opposed to absolute variation across the entire speech stretch. For consonantal intervals this measure is typically a raw value (rPVI-C); for vocalic intervals this measure is normalized for speech rate (nPVI-V). These metrics will be utilized in the methodological aspects of this study. Precise formulas for each of these measures can be found in the original sources.

The point of these metrics is that they account for the durational fluctuations that characterize different speech rhythms. The global differences in consonantal and vocalic durations will have a direct effect in terms of metric scores, and the dynamics that are in part responsible for the surface configurations of consonants and vowels in a given string are the phonotactics of the language. The role that phonotactics plays in determining the rhythmic profile of a language has been proposed by in early work on rhythm [6], [7], [8], [9]. The purpose of the present study is to determine whether rhythm is an intrinsic aspect of a language, or simply a byproduct of phonotactics. Related to this is the question of whether rhythm is a byproduct of the stimuli selected for a given study, which this work also aims to address.

2. Methods

The experiment outlined here was designed to test whether the phonological structure of sentences (i.e. the complexity of syllable types in a given sentence) affects rhythm measures. This study aims to test this by altering the phonological properties of the stimuli. This is accomplished by observing two sets of stimuli from the same language (New Zealand English); one set with an intentionally simple phonotactics (to mimic syllable-timing), and one with a relatively more complex phonotactics (to mimic stress-timing). These sets can be contrasted with a ‘baseline’ set, the purpose of which is to provide a ‘controlled English’ which hasn’t been adjusted phonotactically. The design is influenced by recent approaches to rhythm [10], [11], [12], [13], and the idea is to determine whether the stimuli influence rhythm metric values. In all of these studies, the authors manipulated a control set of sentences and made them either more “syllable-timed” or more “stress-timed”. The intention is to replicate this with (i) a large number of speakers from a single language, and (ii) with a baseline set of sentences used in other comparable studies.

2.1. Participants

Participants included 32 speakers in total: 27 speakers of New Zealand English, 4 speakers of American English, and 1 speaker of British English. In order to avoid any possible intrinsic differences between dialects in terms of rhythmic profile [5], [14], the data presented here are from the same dialect (New Zealand English). Data from the other dialects was removed. Of the 27 speakers, 10 were male and 17 were female. None of the participants reported any speech or hearing problems.

2.2. Materials

The data collected for the study included a set of read sentences along with a stretch of unscripted speech from each speaker. The read sentences consisted of three different sets: one set intending to reflect “syllable-timing”, one intended to reflect “stress-timing”, and a “baseline” set, which wasn’t intended to reflect any sort of timing bias. There were 5 sentences in each set. The “syllable-timed” set of stimuli were composed of only (C)V syllables. The “stress-timed” set consisted of many syllable codas and complex margins (both of which create consonant clusters). In order to make the “baseline” set comparable to other studies, the stimuli from
other well-known studies were used [15], [16], [17]. In order to make the syllable-timed and stress-timed stimuli as maximally different from the baseline stimuli as possible, a C/V ratio for each sentence was calculated, and a count of all consonant clusters was made. The aim was to achieve a C/V ratio as close to 1:1 as possible for the “syllable-timed” stimuli and a 2:1 C/V ratio for the “stress-timed” sentences (i.e. extremes maximally different from 1.5:1 – the approximate C/V ratio for the baseline stimuli), without resorting to lexical selections (i.e. those with only V syllables, or only those with multiple consonant clusters) that resulted in pragmatically odd or unusual sentences. Since the baseline sentences consisted of 3 sentences with 15 syllables each and 2 sentences with 17 syllables each, this was replicated for each of the manipulated sets such that each set consisted of the same number of syllables. Mean C/V ratios for the stress-timed set were 0.93, for syllable-timed 2.08, and for the baseline set, 1.51. In order to compare these numbers to an independently collected source of materials for New Zealand English, a C/V ratio was also computed for the passage of “The North Wind and Sun” [18]. The C/V ratio for this text was 1.52:1, which is comparable to the baseline ratio of 1.47:1. This is also comparable to C/V ratios and consonant cluster counts computed from another recent study [11], where the stress-timed stimuli had a mean C/V ratio of 1.89 (with a mean of 10.75 consonant clusters) the syllable timed stimuli had a mean ratio of 1.21 (with a mean of 4 consonant clusters), and the control stimuli had a ratio of 1.61 (with a mean of 7.8 consonant clusters).

Because of the difficulties in measuring approximant-vowel boundaries, the glides [j], [w] and liquids [l], [r] were avoided where possible, following recent standards [15], [16], [17]. Postvocalic /l/ is not present in New Zealand English [18]. Sentences were presented to participants in pseudo-random order. A short narrative of free speech was collected before any scripted speech. Each participant was asked to speak for 1-2 minutes on one of the selected topics, either parking on campus, or about public transportation [11]. Roughly 1 and a half minutes of speech were collected from each participant. These sentences served as instances of “uncontrolled” speech, which could be compared with respect to speech rate.

2.3. Measurements

Speakers were presented with the sentences preceding recording in order to become familiar with them. Speakers were instructed to read the sentences at a normal, comfortable rate. After the recording session, if any substantial pauses or disfluencies were detected during the reading of a given sentence, the speaker was instructed to re-read the sentence(s), and these were re-recorded. Recordings were made directly to disk at a sampling rate of 48 kHz. A high quality microphone was used, held approximately 10 centimeters from the participant’s mouth. Recordings were made in a quiet room in the University of Auckland Speech Research Lab. The second author identified boundaries for consonants, vowels, and pauses through inspection of a waveform and spectrogram of the speech signal, and these intervals were labelled appropriately. Measurement criteria employed for segments were based on standard assumptions [19]. Vocalic sequences identified as diphthongs (for New Zealand English [18]) were counted as a single vocalic interval; all other adjacent vowels were counted as separate vocalic intervals. Pauses were labeled as such. Utterance-final and pre-pausal intervals were counted; as has been argued [4], [15], [16], despite the probability of lengthening in these environments, this is not likely to artificially alter the rhythmic profile of the language (and, rather, would contribute to it).

For the unscripted speech samples: in order to make the rhythm measures comparable to the read sentences, the mean duration of the sentences for a given participant were calculated, and this duration was excised from the unscripted speech. This excised portion was taken from at least 15 seconds into the unscripted speech. The next string of sound uninterrupted by pauses or disfluencies was selected; for 2 speakers this was not possible, and so the portion was excised beginning at 15 seconds. Any intervals that were incomplete at the end of the portion were counted as pauses. Rhythm metrics were calculated using the Correlatore software [20].

3. Results

The mean scores and standard deviations (in parentheses) for all metrics mentioned above for each set of sentences are reported in Table 1.

Table 1. Means and standard errors of sentences.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Syllable</th>
<th>Baseline</th>
<th>Stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>ΔV</td>
<td>61.05 (2.43)</td>
<td>46.17 (1.35)</td>
<td>39.31 (0.87)</td>
</tr>
<tr>
<td>ΔC</td>
<td>37.48 (1.22)</td>
<td>62.04 (1.68)</td>
<td>81.71 (1.92)</td>
</tr>
<tr>
<td>%V</td>
<td>56.53 (1.53)</td>
<td>42.11 (1.56)</td>
<td>34.79 (1.50)</td>
</tr>
<tr>
<td>VarcoV</td>
<td>57.82 (0.92)</td>
<td>53.64 (0.83)</td>
<td>45.42 (0.79)</td>
</tr>
<tr>
<td>VarcoC</td>
<td>45.54 (1.03)</td>
<td>55.92 (0.90)</td>
<td>54.82 (0.83)</td>
</tr>
<tr>
<td>nPVI-V</td>
<td>68.21 (1.50)</td>
<td>62.37 (1.17)</td>
<td>45.44 (1.00)</td>
</tr>
<tr>
<td>rPVI-C</td>
<td>46.39 (1.50)</td>
<td>76.79 (2.00)</td>
<td>96.84 (2.28)</td>
</tr>
</tbody>
</table>

A one-way ANOVA yielded significant variation among the three different conditions for all of the metrics: for ΔV, F(2,78) = 43.63; p < .0001; for ΔC, F(2,78) = 184.92; p < .0001; for %V, F(2,78) = 432.56; p < .0001; for VarcoV, F(2,78) = 56.66; p < .0001; for VarcoC, F(2,78) = 38.22; p < .0001; for nPVI-V, F(2,78) = 91.27; p < .0001; for rPVI-C, F(2,78) = 169.21; p < .0001. Post hoc comparisons of the Tukey HSD test indicated that the mean score for the syllable-timed condition was significantly different from the stress-timed condition for all metrics at the .05 level of significance, the syllable-timed condition was significantly different from the baseline condition for all metrics at the .05 level of significance, and that the stress-timed condition was significantly different from the baseline condition for all metrics at the .05 level of significance except for VarcoC, which yielded non-significant results. Taken together, these results suggest that the ratio of consonants to vowels in a sentence have an effect on the rhythmic profile of the utterance. Specifically, the results suggest that the higher the C/V ratio, the higher the scores for each rhythm metric.

The next relevant comparison to be made is between the baseline condition and the unscripted condition. Results from a one-way ANOVA indicate the difference between these categories is significant for ΔV (F(1,52) = 14.52; p < .0001), and %V (F(1,52) = 21.34; p < .0001; and however, the difference was not significant for any of the other metrics. A potential explanation for these vocalic interval metrics being different lies in the compressibility of vowels (as opposed to consonants) under increased rate. These results suggest that the utterances in the baseline condition are not significantly different from those found in unscripted speech; i.e. that the
baseline condition is reflective of the rhythmic properties of the language as a whole, if 'unscripted' is taken to be an accurate measure of language rhythm. An additional implication is that the significant differences that are found for \( \Delta V \) are relatable to the higher compressibility of vowels (as compared to consonants) under higher rates of speech [21].

In order to place these scores into a wider cross-linguistic context, the comparable recordings from a published study [15], [16] (corresponding to our ‘baseline’ data set). This included data from two canonical stress-timed languages, British English and Dutch, and two canonical syllable-timed languages, Spanish and French. Graphically presenting the New Zealand English results along with the results from these languages will help to put the results in context. For consistency and to avoid inter-rater differences [17], measurements were conducted independently over this data set. The means and standard errors for this data are presented in Table 2.

Table 2. Means and standard errors of “baseline” sets

<table>
<thead>
<tr>
<th>Metric</th>
<th>English</th>
<th>Dutch</th>
<th>French</th>
<th>Spanish</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \Delta V )</td>
<td>50.5 (1.6)</td>
<td>47.5 (4)</td>
<td>39.5 (2.3)</td>
<td>35.3 (3.8)</td>
</tr>
<tr>
<td>( \Delta C )</td>
<td>59.8 (1.2)</td>
<td>60.2 (6.8)</td>
<td>48.6 (1)</td>
<td>40.2 (9)</td>
</tr>
<tr>
<td>%V</td>
<td>41.1 (0.6)</td>
<td>41.4 (2.5)</td>
<td>46.5 (1.5)</td>
<td>51.3 (1.6)</td>
</tr>
<tr>
<td>VarcoV</td>
<td>59.7 (2.4)</td>
<td>56.4 (4.4)</td>
<td>43.4 (2.3)</td>
<td>42.4 (3)</td>
</tr>
<tr>
<td>VarcoC</td>
<td>51.7 (2.0)</td>
<td>51.2 (3.9)</td>
<td>45.1 (1.5)</td>
<td>49.8 (4.1)</td>
</tr>
<tr>
<td>nPVI-V</td>
<td>66.9 (2.7)</td>
<td>67.9 (5.7)</td>
<td>44.7 (3.3)</td>
<td>35.6 (1)</td>
</tr>
<tr>
<td>rPVI-C</td>
<td>73.7 (2.5)</td>
<td>63.3 (8.6)</td>
<td>53.8 (1.6)</td>
<td>44.0 (1.3)</td>
</tr>
</tbody>
</table>

The only significant difference between the “baseline” New Zealand English set and British English was VarcoV, \( F(1,31) = 7.51; p < .05 \), and VarcoC approached significance, \( F(1,31) = 4.15; p = 0.0503 \); otherwise, there were no other significant comparisons: %V, \( F(1,31) = 0.65; p = 0.426 \); \( \Delta V \), \( F(1,31) = 2.15; p = 0.153 \); \( \Delta C \), \( F(1,31) = 0.37; p = 0.547 \); VarcoV, \( F(1,31) = 7.51; p < .05 \); nPVI, \( F(1,31) = 2.85; p = 0.1014 \); rPVI, \( F(1,31) = 0.5; p = 0.485 \).

In order to put these values into context, the values for various metrics were plotted against each other in two-dimensional space. %V and \( \Delta C \) are plotted against each other. The values for %V and \( \Delta C \) [1] are presented in Figure 1. As can be seen, the “syllable-timed” data pattern very closely with the other syllable-timed languages, French and Spanish, while the “baseline” data pattern much like British English, Dutch, and unscripted New Zealand English. The “stress-timed” data are substantially removed from this clustering, though in the opposite direction of the syllable-timed set. In observing the array that the Varco scores yield (Figure 2), there are some immediate anomalies. The low VarcoC score puts the “syllable-timed” set near the French and Spanish sets, though the VarcoV score also puts the stress-timed set in range of French and Spanish. Turning to the pairwise variability indices (Figure 3), we can see that the stress-timed and syllable-timed data are nearly equidistant from the baseline (and British English) data points, a result that is expected if rhythm (as reflected by rhythm metrics) is simply the result of the overall prosodic complexity of a given sentence.

A previous study [13] found that metric scores were generally sensitive to stimuli with different phonotactic properties, with vocalic indices more reflective of what they consider to be inherent rhythmic differences, and consonantal indices more sensitive to phonotactic differences. For instance, for %V, they found that the simpler the syllable structure, the higher the metric score. However, in addition to this monolithic effect, they found that both Catalan and Spanish cluster together (and differently from English) for CVC and mixed (roughly equivalent to our ‘baseline’) stimuli, which is interpreted as a cross-linguistic difference that is also encoded in %V. They posit that the reason rhythm differences still show up when syllable structure is controlled for is due to how other phonological properties are controlled; these cross-linguistic differences include the durational

![Figure 1: Means of sentences in %V/\( \Delta C \) rhythmic space](image)

1695
marking of stress and pre-final lengthening. In another study [11], [12], there was a significant difference across sentence types for all metrics. Within languages, the findings were consistent with the findings here: that stress-timed sentences have a significantly higher (lower for %V) score than the baseline condition. Across languages, many of the pairwise comparisons between languages which were significant break down when sentence types are pulled apart.

4. Conclusions

It was shown that rhythm, as reflected in the various metrics, is a byproduct of syllable structure; or, more specifically, the net consonants and vowels available in a language/speech utterance. There can be great variation not only across languages, but also between sentences within a language. The implication of these findings is that speech rhythm is not an intrinsic property of languages, but rather is an emergent one.

5. Acknowledgements

The authors wish to acknowledge Sven Mattys and Laurence White who kindly shared their recordings with us.
6. References


