Towards a perceptual model of speech rhythm: Integrating the influence of f0 on perceived duration

Robert Fuchs

1 Englisches Seminar, Westfälische Wilhelms-Universität Münster

robert.fuchs@uni-muenster.de

Abstract

Previous accounts of speech rhythm focus mainly on duration. For example, the normalised Pairwise Variability Index for vocalic intervals (nPVI-V) quantifies relative duration differences between successive vocalic intervals. Prototypical syllable-timing is characterised by small differences in duration, prototypical stress-timing by large differences. However, differences in f0 between vocalic intervals are thought to influence the perception of duration. This paper (1) quantifies the influence of differences in f0 on perceived duration in a perception experiment, and (2) suggests a modified PVI (nPVI-V(dur*f0)) that takes account of this influence. The new nPVI-V(dur*f0) is then applied to a speech corpus of (stress-timed) British English and (syllable-timed) Indian English. The results are compared to the application of the old nPVI-V, which takes into account duration only, to the same data set.

Index Terms: Speech rhythm, perceived duration, fundamental frequency, British English, Indian English, syllable-timing

1. Introduction

Recent research has criticised the idea of syllable- and stress-timed rhythm classes, and a gradient analysis is now preferred by many researchers. Accordingly, languages can be placed on a continuum between a prototypically stress- and a prototypically syllable-timed pole. For example, some (mainly more established) varieties such as British English (BrE) and American English are closer to the stress-timed pole. By contrast, many of the younger national varieties of English (such as Indian, Nigerian and Singapore English) are more syllable-timed than BrE. This is often explained with transfer from local syllable-timed languages [1, 2].

The gradient analysis of speech rhythm is supported by quantifications of based on the variability of durations of vocalic intervals, consonantal intervals and syllables, as suggested by [3–11]. One widely used rhythm metric is the normalised Pairwise Variability Index for vocalic intervals, or nPVI-V [5]. It is computed by calculating the mean of the differences between successive vocalic intervals divided by their sum, multiplied by 100:

\[ nPVI - V = 100 \times \frac{\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right|}{m - 1}; \]  

where \( m \) is the number of vocalic intervals

and \( d_k \) is the duration of the \( k^{th} \) vocalic interval

Comparisons of rhythm metrics by [12–14] suggested that speech rate normalised vocalic metrics are the most reliable measures of speech rhythm. Even these metrics are influenced by variation between transcribers, speakers, texts, implying that phonetically balanced or large samples have to be used together with clearly defined transcription rules. The importance of these suggestions are is demonstrated by [15], who failed to replicate the results of [12, 13], perhaps because samples from each language were used that were not representative, but consisted of sentences that differed in rhythm as much as possible.

However, the exclusive focus on durational variability may present an incomplete account of speech rhythm. Most studies on rhythm in the last 15 years have concentrated exclusively on duration, which means that they considered only one correlate of prominence and ignored a more broadly defined notion of prosodic prominence [16]. Instead of describing rhythm as the alternation of long and short intervals, it might be more useful to characterise it as the alternation of prominent and non-prominent intervals. [17–22] suggested rhythm metrics based on variability in intensity, \( f_0 \), and sonority. Together, these metrics may form a multidimensional account of rhythm, as demanded by [23–25]. A comparison of two languages might show that one has more variability in all these correlates of prominence and is therefore more stress-timed than the second language on all these dimensions, or that it has more variability in one but less in another correlate of rhythm. Any language might therefore have different co-existing rhythms [25], which contribute to a succession of elements relatively similar in prominence (syllable-timing) or relatively different in prominence (stress-timing). Apart from variability in duration, relatively little is known about how the various acoustic correlates of rhythm interact.

2. Aims

This paper aims to contribute to recent research that investigates the interaction of different acoustic correlates of speech rhythm by focussing on the interaction of the variability in duration and fundamental frequency (\( f_0 \)). Specifically, it aims to

• Quantify the influence of differences in mean \( f_0 \) on perceived duration (in contrast to acoustic duration)

• Integrate this influence into a quantitative measure of speech rhythm, by modifying the nPVI-V formula

• Apply the new nPVI-V(dur*f0) to data from (stress-timed) BrE and (more syllable-timed) Indian English (IndE), in order to determine whether the effect of differences in mean \( f_0 \) on perceived duration contributes to rhythm differences between the varieties

The following sections discuss the influence of variability in \( f_0 \) on speech rhythm and its interaction with variability in duration (sections 3.1-3.2). Next, the results of a perception experiment...
are reported which aimed to quantify the effect of differences in mean $f_0$ on perceived duration (sections 3.3-3.5). These results are then integrated into the nPVI-V formula (section 4), which will be applied to data from BrE and IndE (section 5).

### 3. The influence of fundamental frequency on speech rhythm

When considering the interaction of variability in duration and fundamental frequency ($f_0$) in the emergence of the rhythm of an utterance, two types of influence may be distinguished: (1) Both duration and $f_0$ are, independently from each other, prominence-lending (duration by alternating short and long intervals, and $f_0$ by alternating low and high frequency or small and wide rises/falls); (2) In addition, $f_0$ may have some influence on perceived duration. For example, given two syllables of the same duration, a syllable with a high tone may be perceived as longer than one with a low tone, or a syllable with a large rise as longer than one with a small rise.

In a multi-dimensional model of speech rhythm, both types (1) (discussed in section 3.1) and (2) (section 3.2) should be taken into account in order to quantify how duration and $f_0$ interact when giving rise to succession of more and less prominent beats [26].

#### 3.1. Direct contribution of duration and $f_0$ to prominence

One way to account for the independent contribution of duration and $f_0$ is to calculate different PVIs for duration and $f_0$ slope size, and take their mean for complete utterances. [19, 21] applied this method to Swiss German, Swiss French and Metropolitan French, using language-specific weightings derived from perception experiments.

This global approach (taking the whole utterance as its domain) can be contrasted with a local approach (taking the vocalic interval as its domain): For each vocalic interval, it needs to be determined in how far it is marked as prominent by increased duration or increased $f_0$ slope. This would take account of whether vocalic intervals are usually marked as prominent by both duration and $f_0$ (thus reinforcing each other in the generation of prominence), or whether duration and $f_0$ increases tend to occur independently from each other, and are in a trading relationship (this method was suggested by [22] to quantify how duration and loudness interact in the generation of prominence).

#### 3.2. Influence of $f_0$ on perceived duration

The influence of $f_0$ on perceived duration can be integrated directly into the PVI formula. We can distinguish the influence of tone height (vocalic intervals with higher $f_0$ are perceived as longer) and the influence of slope size (vocalic intervals with a more extreme rise or drop in $f_0$ are perceived as longer) on perceived duration [27–34]. The present paper focusses on the influence of tone height. Assuming that there is a linear relationship between tone height and perceived duration, perceived duration can be calculated by estimating a factor $p$ in $\frac{d_{p}}{d}$. $p$ expresses the relative increase in perceived duration due to differences in tone height.

The perceived duration $d'$ can then be calculated by taking the acoustic duration $d$, and adding to it the increase in perceived duration, which depends on the difference in $f_0$ between successive vocalic intervals (see formula 2 below). If the mean $f_0$ of the preceding vocalic interval is smaller than that of the following interval ($f_k < f_{k+1}$), then the perceived duration of the $k^{th}$ vocalic interval ($d'_k$) is smaller than its acoustic duration. If $f_k > f_{k+1}$, it is greater. By contrast, the perceived duration of the $k+1^{th}$ vocalic interval ($d'_{k+1}$) is greater than its acoustic duration, if $f_k < f_{k+1}$.

$$d'_k = d_k + d_k \times p \times (f_k - f_{k+1}); \quad (2)$$

$$d'_{k+1} = d_{k+1} + d_{k+1} \times p \times (f_{k+1} - f_k); \quad (2)$$

where $d_k$ is the acoustic and $d'_k$ the perceived duration, and $f_k$ the mean fundamental frequency of the $k^{th}$ vocalic interval, and $p$ a constant that adjusts for the increase in perceived duration

The perceived durations $d'_k$ and $d'_{k+1}$ can then replace the acoustic durations in the PVI formula, which now takes into account the effect of tone height ($f_0$) on the perception of differences in duration (hence $nPVI - V(dur + f0)$):

$$nPVI - V(dur + f0) = 100 \times \frac{\sum_{k=1}^{m-1} d'_k - d'_{k+1}}{m - 1} \quad (3)$$

where $m$ is the number of vocalic intervals

and $d'_k$ is the perceived duration of the $k^{th}$ vocalic interval

### 3.3. Determining the influence of tone height on perceived duration

[32] found that vowels with higher level $f_0$ are perceived as longer than those with lower $f_0$. However, in order to integrate the influence of $f_0$ on perceived duration into the measurement of speech rhythm, it must first be quantified. This was attempted by [35], who used 500 Hz and 4000 Hz tones in perception experiments. Notwithstanding these results, 4000 Hz is far beyond the human vocal range, and lower frequencies are needed in order to quantify the influence of differences in tone height on perceived duration. This is why, for the present paper, a perception experiment with stimuli differing in duration and $f_0$ was conducted.

#### 3.4. Perception experiment

##### 3.4.1. Stimuli

In a two-alternative forced choice (2AFC) experiment, participants were presented with utterances consisting of the syllable /pap/, generated with the en1 MBROLA voice (see appendix for an example) [36]. The vowel of the first syllable always had a duration of 200 ms, and the duration of the vowel of the second syllable was varied in 18 steps from 40 to 300 ms. $f_0$ was varied in both syllables on three levels (85, 115, and 145 Hz), resulting in 6 $f_0$ combinations x 18 durations = 108 different stimulus pairs.

The experiment was presented over headphones on a computer in a quiet room using the Praat MFC environment [37]. Participants were instructed that they would hear two different sounds in every round of the experiment, and had to decide which of the two is longer (replay was allowed). After a short

---

1 Closer to 200 ms, duration was varied in 10 ms steps, and otherwise in 20 ms steps (40, 80, 100, 120, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 260, 280, 300 ms).

2 In addition, 144 pairs with different durations and rising $f_0$ were presented. For reasons of space, these results are not reported here, and the 169 additional stimulus pairs can be regarded as distractors.
training session, all stimuli were presented once in random order with a 0.8 s inter-stimulus interval. Participants were offered a short break after every 50 stimulus pairs. 31 native German speakers (21 f., 10 m., median age 24, range 19–62) with no reported history of hearing problems participated voluntarily and without pay.

3.4.2. Results
A logistic regression was used to test whether the perception of the second syllable as longer or shorter than the first was influenced by DIFFERENCE_DUR and DIFFERENCE_F0 (ranging over -60, -30, 30, 60 Hz; e.g. syll. 1 85 Hz, syll. 2 145 Hz means DIFFERENCE_F0 = 145 - 60 = 60 Hz). Both DIFFERENCE_DUR and DIFFERENCE_F0 (both p<0.0001) had a significant effect on the perception of the second syllable as longer or shorter. Post-hoc Tukey tests showed that most difference contrasts were significant (60 vs. -60 Hz and 60 vs. -30 Hz p<0.01, 30 vs. -60 Hz and 30 vs. -30 Hz p<0.05, -30 vs. -60 Hz and 30 vs. 60 Hz n.s.).

When the second syllable had the same duration as the first and had an f0 that was 60 Hz higher, it had a chance of 59.3% to be classified as longer than the first. A syllable with an f0 that was 60 Hz lower had a chance of only 41.6% to be classified as longer than the first. Given the coefficients estimated by the logistic regression, it is possible to transpose the equation in such a way that we can derive the effect on perceived duration of any given f0 difference. On average, an f0 difference of 60 Hz (second syllable 60 Hz higher) adds 8 ms (4 %) in perceived duration. If the second syllable is 60 Hz lower than the first, it is perceived as 8 ms/4 % shorter than the first.

3.5. Discussion
In the 2AFC listening experiment, syllables of varying duration and f0 were presented. Participants had to judge whether the second syllable was longer or shorter than the first. Transposing the logistic regression formula, a 60 Hz difference can be estimated to add 8 ms to a 200 ms syllable, i.e. a 4 % increase in perceived over acoustic duration.

Generalisation of the results might be limited by the conditions used in the experiment. Only the second syllable was varied in duration, and only level f0 contours were used. Future research should determine whether these factors influence the effect of f0 on perceived duration. However, given the exploratory nature of the present purposes it will be assumed that a +60 Hz difference in mean f0 will lead to a 4 % increase in perceived duration, and a -60 Hz difference in mean f0 to a 4 % decrease in perceived duration.

4. Integrating the influence of f0 on perceived duration into the PVI
Based on the results of the perception experiment, factor p in formula 2 above can be estimated. The experiment indicated that for every 60 Hz in f0 difference, there is a 4 % increase in perceived duration in the second syllable/vocalic interval.

In the PVI formula, there are always two durations being compared at a time, which is why formula 2 adjusts the durations of the kth and the following interval (d_k and d_k+1). The 4 % increase in perceived duration arrived at in the experiment must therefore be halved (2 %), and divided by 60 to arrive at the factor p (in \(\sqrt{2}\)) to be used in the formula (i.e. \(p = 0.000333\)). In the following, the new nPVI-V(dur*f0) (formula 3) and the nPVI-V(dur) (formula 1) will be applied to data from educated IndE and BrE.

5. Case study: British English and Indian English speech rhythm

5.1. Indian English
IndE is a postcolonial variety of English used mainly in public contexts. Around 23% of the population of India have at least basic knowledge of English, and 4% are fluent [38], which means that there are around 50 million fluent speakers.

Standard IndE is not yet fully codified and still lacks full official recognition, but is the de facto standard taught in schools and universities [39, 40]. The phonology of IndE differs in several respects from that of BrE [41]. Despite phonological differences between Indian languages, the phonology of educated IndE is relatively homogeneous, independently of the first languages used by particular speakers (or the language families they belong to) [42].

IndE has been suggested to be more syllable-timed than BrE, and acoustic evidence based on a wide variety of acoustic correlates of speech rhythm has confirmed this [22, 26]. In particular, educated IndE has less variability in vocalic durations than BrE (as measured by nPVI-V).

5.2. Research questions
The aim of the case study is to determine whether the influence of f0 on perceived duration contributes to rhythm differences between IndE and BrE. This could be the case, if, in BrE, vocalic intervals with a long acoustic duration are also marked by high f0, further increasing their perceived duration. By contrast, vocalic intervals with a short acoustic duration might tend to have low f0, so that their perceived duration is even shorter than their acoustic duration.

This seems likely because focussed syllables, which have a longer acoustic duration than unfocussed syllables, are frequently marked by an H*(L) pitch accent in BrE [43]. Speakers of IndE, on the other hand, tend to use both L*(H) and H*(L) pitch accents for focussed syllables [44, 45]. This suggests that the perceived duration of accented syllables is sometimes longer and sometimes shorter than their acoustic duration.

It is therefore hypothesised that, in BrE, variability in perceived duration is higher than in acoustic duration, while it is smaller in IndE. This would cause the difference in speech rhythm between the varieties to be greater when measured by perceived duration (nPVI-v(dur*f0)) than by acoustic duration (nPVI-V(dur)).

In the case study, results from the perception experiment with German-speaking listeners will be applied to data from BrE and IndE. While it would have been possible to have either of these two groups participate in the perception experiment, it might be interesting to determine how an independent, third group of listeners perceives the rhythm difference between BrE and IndE. This is in keeping with recent research on English as a Lingua Franca (see, e.g., [46]), which recognises that native speakers of English (such as Britons) are outnumbered by speakers of non-native varieties (such as Germans) and post-colonial varieties of English (such as Indians).

5.3. Data and Methods
Recordings of a text read by 10 speakers of Standard Southern BrE and 20 speakers of IndE were used. The BrE data was
taken from the DyViS database [47]. The IndE speakers were recorded by the present author reading the same text [26]. All speakers were university students at the time of recording. The IndE speakers were equally divided between four different L1 groups, and had either Hindi or Bengali (both Indo-Aryan languages), or Telugu or Malayalam (both Dravidian languages) as first languages (L1). Speakers were between 20 and 28 years of age, and, with the exception of one speaker each, had exclusively attended English-medium schools and universities, and had not resided outside of India.

392 words of the reading passage were segmented following [12, 13]. Duration and mean $f_0$ were extracted with the help of Praat, and nPVI-V(dur) and nPVI-V(dur*$f_0$) were computed individually for utterances comprising at least three vocalic intervals, excluding the final interval.

5.4. Results

While IndE has a slightly lower variability in perceived duration ($nPVI$-$V(dur)=55.85$) than in acoustic duration ($nPVI$-$V(dur*$f_0$)=55.92, paired t-test $p<0.05$), BrE has significantly more variability in perceived duration ($63.09$) than in acoustic duration ($62.71; paired t-test p<0.0001$). Hence, the difference in variability between IndE and BrE is higher in perceived duration (difference between means $7.24$) than in acoustic duration (difference between means $6.79$), as Fig. 1 shows. The difference in variability in perceived duration is therefore $6.6\%$ higher than the difference in acoustic duration.

5.5. Discussion

The results confirm the hypothesis that, in BrE, variability in perceived duration (when taking into account the influence of $f_0$) is greater than variability in acoustic duration, while in IndE variability in perceived duration is smaller than in acoustic duration. This suggests that differences in speech rhythm between the two varieties are further strengthened by perceptual effects, which cause an increase of $6.6\%$ in the rhythmic difference between IndE and BrE.

The differential influence of differences in $f_0$ on variability in perceived duration are likely to be caused by differences between the intonation of IndE and BrE. In IndE, pitch accents often bear a $L^*(H)$ or $H^*(L)$ tone. By contrast, in BrE mainly $H^*(L)$ tones are used [43–45]. In consequence, syllables (and their vowel nuclei) with a pitch accent might more often than not have lower mean $f_0$ than adjacent syllables in IndE, whereas in BrE they usually have higher mean $f_0$ than adjacent syllables. Because the vowels of accented syllables are usually lengthened, the sometimes lower mean $f_0$ of accented syllables in IndE causes variability in perceived duration to be slightly lower than variability in acoustic duration. By contrast, in BrE the usually higher mean $f_0$ of accented syllables amplifies duration contrasts, and causes variability in perceived duration to be higher than variability in acoustic duration.

6. Conclusions

This paper set out to quantify (by means of a perception experiment) the influence of mean $f_0$ on perceived duration in order to integrate it into the measurement of speech rhythm. Using data from BrE and IndE, a case study showed how the new nPVI-V($dur*$f0), taking this influence into account, can be applied.

Result indicate that the stress-timed rhythm of BrE is further reinforced by these perceptual effects because accented syllables (and their vowels) are usually lengthened and higher in $f_0$ than adjacent syllables. By contrast, the more syllable-timed rhythm of IndE is also slightly reinforced because accented syllables (and their vowels) are sometimes lower in $f_0$ than adjacent syllables.

The analysis supports a description of speech rhythm as an, at least partially, perceptual phenomenon [48]. Because vocalic intervals with higher $f_0$ are perceived as longer than those with lower $f_0$, the influence of $f_0$ on perceived duration needs to be taken into account in the measurement of speech rhythm. Future research should investigate how other factors, such as the influence of dynamic $f_0$ being perceived as longer than level $f_0$, contribute to the perception of speech rhythm. This would further contribute to a multi-dimensional model of speech rhythm that takes into account various acoustic correlates of rhythm as well as their interaction.

7. Appendix

MBROLA .pho file to generate a /pap/ syllable with a 110 ms [p], a 200 ms [a], and a 110 ms [p] phoneme (preceded and followed by silence), with a level $f_0$ of 145 Hz.

- 110
  p 110 53 145
A: 200 100 145
p 110 53 145
- 11

8. References


