Original Paper

Phonetica 2011;68:256–277
DOI: 10.1159/000335416

Received: November 15, 2011
Accepted: November 25, 2011

Perceptually Informed Quantification of Speech Rhythm in Pairwise Variability Indices

Ruth E. Cumming

Centre for Neuroscience in Education, Department of Experimental Psychology, University of Cambridge, Cambridge, UK

Abstract

Two previous experiments demonstrated that f0 and duration are interdependent in the perception of rhythmic groups in speech and sentence rhythmicality, and that the relative weighting of tonal and durational cues depends on listeners' native language. The listeners were native speakers of Swiss German, Swiss French, or Metropolitan French (i.e. from France). The experiment reported here investigates a means of applying this perceptual finding to production data from these three languages, to make a rhythm metric, the Pairwise Variability Index (PVI), perceptually informed. The relative weighting that an appropriate duration and an appropriate f0 contributed to listeners' rhythmicality judgements is calculated, and these language-specific weighting values are incorporated into combined durational-tonal PVIs, to quantify rhythm in the three languages. The results demonstrate that Swiss German and Swiss/Metropolitan French are distinct according to classic durational PVIs, but more similar according to PVIs which are acoustically multidimensional and language-specifically weighted. It is concluded that rhythm produced by speakers, when quantified to account for the acoustic multidimensionality and language-specificity of rhythm perceived by listeners, may be less cross-linguistically divergent than durational rhythm metrics suggest. An evaluation of these language-specifically weighted PVIs concludes that if rhythm metrics remain in use, they should link rhythm perception with rhythm production.

1. Introduction

1.1. Rhythm in Speech

The phenomenon of rhythm in speech has been researched for many decades [e.g. Steele, 1775; Classe, 1939] and has recently been the subject of several experiments (see the many references in the following sections). However, there is still no general consensus on exactly how speech rhythm is defined and indeed if speech is typically

KARGER

Fax +41 61 306 12 34
E-Mail karger@karger.ch
www.karger.com

0031–8388/11/0684–0256
$38.00/0
Accessible online at:
www.karger.com/pho

Ruth E. Cumming
Centre for Neuroscience in Education
Department of Experimental Psychology
Downing Street
Cambridge, CB2 3EB (UK)
Tel. +44 1223 767 509, E-Mail reg50@cam.ac.uk

a rhythmical activity. For the experiment reported in this article, the following definition of speech rhythm is proposed: the perceived regularity in an utterance, which is induced by the grouping of prominences (which involve various acoustic dimensions of the speech signal), and is influenced by the listener's native language [e.g. Arvaniti, 2009; Barry et al., 2009; Kohler, 2009a; Niebuhr, 2009; for recent similar ideas of what speech rhythm is]. Each part of this definition makes an important point.

First, speech rhythm is a perceptual phenomenon, which may not have an easy to capture correlate or set of correlates in speech production; it also involves a regular recurrence of some aspect of speech, though just how temporally exact 'regular' needs to be and what exactly it is that recurs are not simple questions. Second, there are many acoustic dimensions in the speech signal that may contribute to the percept of rhythm, e.g. the duration, tonal properties, amplitude (including amplitude-related properties such as spectral balance) and spectral properties (such as formant structure) of various units – segments, syllables and longer prosodic groups. Essentially, these acoustic properties are potential cues to prominence in a language, and the pattern of prominent and non-prominent sections of the speech signal, i.e. how these sections are grouped, is key in the percept of rhythm. Third, there is some evidence that the particular phonological properties of a listener's native language affect his/her perception of rhythm, by influencing which of the acoustic cues are more significant than others in contributing to the percept of rhythm that they hear [Cumming, 2011b].

However, rhythm may be more obvious in some languages than in others, and in some it could be argued that there is no perceptible rhythm. For example, it has been said that rhythm in English speech is perceived as the regular recurrence of prosodic groups that comprise one stressed (i.e. prominent) syllable and one or more unstressed syllables; whereas in Korean, prominence of sections of speech is much less obvious, and thus the concept that this language has a perceptible rhythm is more debatable [Nolan and Asu, 2009]. Indeed it has been suggested that speech might not actually be rhythmical at all, but some languages have given the impression of rhythmicality because they have prominences that the brain tries to sort into groups, which is a general tendency in perception [Pamies Bertrán, 1999; Nolan and Asu, 2009]. While this is a real possibility requiring further investigation, it is interesting that even non-expert listeners (i.e. with no phonetic training) were able to judge whether sentences had a natural rhythm, without being given a definition of rhythm, and were able to give their view of what rhythm in speech is, most of them naming which acoustic cue(s) they attended to [Cumming, 2011b]. Although these listeners were native speakers of only two different languages (Swiss German and French), it suggests that some regular pattern in speech was perceptible by people who had not particularly thought about it before; it would also be interesting to try such an experiment on speakers of e.g. Korean.

1.2. Rhythm Metrics

Over a decade ago, two separate lines of research proposed what have since become known as 'rhythm metrics', which are essentially a method for quantifying from the acoustic signal various phonological properties that apparently characterise the rhythm of different languages. One proposal was the Pairwise Variability Index (PVI), originally set out by Low [1994, 1998], which calculated the difference in a

given acoustic property (e.g. duration, amplitude or spectral dispersion) between the members of each successive pair of vowels, and from this the mean pairwise difference across an utterance. Later research using the PVI, since Grabe and Low's [2002] study, has focussed on simply duration rather than several acoustic cues, though it has been extended to consonantal as well as vocalic variability. The other proposal, put forth by Ramus et al. [1999], was a set of three metrics: %V (the proportion of vocalic sounds in an utterance), ΔV and ΔC (the standard deviation of vocalic or consonantal interval duration).

Both these types of rhythm metric result in a summary index, a discrete value to represent rhythm. The main idea was that some languages (once called 'stress-timed', though this became controversial) showed high durational variability, i.e. high PVI, ΔV and ΔC values, whereas other languages (once called 'syllable-timed') showed low durational variability, i.e. low PVI, ΔV and ΔC values, and a high %V value. Two differences between these types of metric are: (i) the PVI takes into account the variability between successive units (vocalic or consonantal intervals), whereas %V, ΔV and ΔC do not, i.e. they calculate variability utterance-globally; (ii) the PVI includes a normalisation component (generally just used for vocalic and not consonantal variability), designed primarily to eliminate between-speaker differences in speech rate, whereas %V, ΔV and ΔC do not.

Since these original proposals, rhythm metrics have featured extensively in phonetic research on speech rhythm [e.g. Grabe et al., 1999; Low et al., 2000; Grabe and Low, 2002; Barry et al., 2003, 2009; Dellwo, 2006; White and Mattys, 2007; Bertinetto and Bertini, 2008; Nolan and Asu, 2009; Payne et al., 2009]. Some of these examples developed the original metrics further: Bertinetto and Bertini [2008] transformed the PVI into the Control/Compensation Index, which accounted for phonotactic complexity; Dellwo [2006] introduced VarcoC (ΔC divided by the mean consonantal interval duration, hence a normalisation); Nolan and Asu [2009] calculated PVIs using syllable and foot durations. (For recent overviews of the various metrics, including methodological information, see Wiget et al. [2010] and Fletcher [2010].) However, rhythm metrics have recently come under attack [Arvaniti, 2009; Barry et al., 2009; Kohler, 2009a]. There were two particular problems that provided motivation for the research reported in this article.

The first problem is that rhythm metrics have become predominantly durational. Exceptions to this are a few studies that have calculated non-durational PVIs [e.g. intensity: Ferragne, 2008; amplitude, spectral dispersion: Low, 1994], or have developed other similar metrics derived from measures of non-temporal properties, such as amplitude in frequency bands associated with vocalic energy [Tilsen and Johnson, 2008], or of automatically computed values based on f0, intensity and duration [Lee and Todd, 2004]. (Note that Lee and Todd [2004] discuss rhythm in the traditional 'stress-timing' versus 'syllable-timing' terms, whereas Tilsen and Johnson [2008] discuss rhythm in new terms.) All these properties may cue prominence in speech. It has long been recognised that the prominence of certain units in speech, as well as their timing (i.e. duration), both play an important role in speech rhythm [e.g. Kohler, 2009a]. Indeed in two perceptual experiments, which preceded the experiment reported later in this article, Cumming [2010b, 2011b] found that both duration and f0 are interdependent cues to perceived rhythmic groups in speech and to the perceived rhythmicality of sentences; the languages investigated were Swiss German and two varieties of French. However, most studies featuring rhythm metrics have regarded prominence as an abstract concept

(i.e. 'stress'/'accentuation'), and not investigated specific physical cues to prominence, but rather concentrated solely on timing by measuring durational properties.

The second problem is that rhythm metrics only capture the speech signal's physical nature, and do not account for whether the relative significance of each acoustic cue for perceived rhythm differs between languages. Cumming [2010b, 2011b] found not only that duration and f0 were interdependent cues to perceived rhythm, but also that the relative weighting or importance of each cue differed between Swiss German and French listeners. The following example may help to illustrate this problem with the metrics. Let a durational PVI of 68 and 42 represent the rhythm of languages A and B, respectively. Let us suppose that for speakers of A, durational variability creates a strong impression of perceived rhythmicality, but for speakers of B, tonal variability creates a stronger impression than durational variability does. Thus the acoustic durational variability of '42' is of little consequence to speakers of B, but that of '68' is highly significant to speakers of A, when they all perceive rhythm in their native language. If the duration-based metric is applied universally, the resulting figures (68, 42) do not have any meaningful interpretation in terms of cross-linguistic differences in perceived rhythm. This point was repeatedly made by Barry et al. [2009, p. 78], who stated that empirically grounded conclusions concerning the link between produced and perceived rhythm are rare in rhythm-metric studies, except those in which listeners discriminated (from segmentally degraded speech stimuli) languages that have different rhythms when quantified with metrics [e.g. Ramus et al., 2003; White et al., 2007]. These studies concerned how rhythm metrics could capture perceptually salient differences in the rhythm *produced* by speakers of different languages. However, they did not address the issue that speakers of different languages may *perceive* rhythm differently, because they may attach different importance to the various acoustic cues in the signal.

1.3. Possible Solution: Motivation for This Research

Given these two problems, ultimately we may question whether applying durational metrics language-universally is justifiable, if they cannot capture the acoustic complexity and language-specific nature of perceived rhythm. Nevertheless, the experiment reported in this article proposes a possible solution to these two problems. In outline, the proposal is to combine duration and f0 in a PVI, and weight each cue's contribution according to its significance for perceived rhythm in each language investigated. This experiment was designed to explore the question of how the quantification of produced rhythm in different languages differs between such a perceptually informed language-specifically 'weighted' metric and a widely used durational language-universal metric. Moreover, PVIs derived from f0 measurements, but not involving duration or language-specific cue weighting, were also compared with durational PVIs and weighted PVIs.

The rationale behind this language-specific perceptual weighting of a rhythm metric is an attempt to capture in a quantification of rhythm more than simply timing, by taking into account the three important aspects of rhythm that were defined in section 1.1: it is a perceptual phenomenon, involving several acoustic cues, the relative significance of which depends on the listener's native language. It must be made clear that this article is not a full empirical study, but rather demonstrates a theoretical point regarding the perceptual weighting of rhythm metrics. That is, durational metrics alone

do not capture the phenomenon of rhythm, so (if it is to be claimed that they do quantify rhythm) they should be adapted to include more acoustic cues and a regard for listeners' perception.

The article introduces a concept that is so far limited to only two languages and two acoustic cues, and based on only one perceptual experiment with an arguably metalinguistic task; much more psychophysical testing involving many more languages and cues (e.g. amplitude and spectral properties, as discussed in section 1.1) would be needed to claim that the weighted PVIs presented here are anything more than a demonstration of a theoretical point. Therefore, although weighted PVIs provide a quantification of rhythm that is more perceptually informed than the traditional durational PVI, they are a first step which requires further investigation beyond the scope of this paper.

1.4. Languages Investigated

This experiment was the final one in a series of experiments concerning the language-specific integration of f0 and duration as rhythm cues [Cumming, 2010a, b, 2011a, b]. Throughout the series, the same languages were investigated: Swiss German (SG), Swiss French (SFr) and Metropolitan French (i.e. from France, Fr). A main reason for this choice was that the rhythm and intonation patterns produced in SG are very different from those produced in SFr and Fr. Therefore, if cross-linguistic differences in rhythm perception (involving duration and f0) were there to be found, they were likely to manifest themselves in the results. There has been little empirical investigation of the differences between Fr and SFr prosody [Miller, 2007]; the inclusion of both varieties of this language in this series of experiments was primarily intended to add to the scarce data comparing SFr and Fr prosody, and is not the focus of the experiment reported here.

It may seem that this choice of languages gives the paper a typological perspective (i.e. categorising different languages into rhythmic types including 'stress timing' and 'syllable timing'), since SG and SFr/Fr exemplify a different rhythm type in the traditional typology. However, this is not the intention; the point of the experiment is not to generate new figures that will better categorise these languages into rhythmic types, but rather to suggest a method for relating perceived rhythm to a quantification of produced rhythm. These languages are two which the author has much experience of. Their analysis needs to be supplemented by similar research in other languages.

1.5. Preliminary Details of the PVI

The formula for the normalised PVI is as follows:

$$\text{normalised PVI} = 100 \times \left[\sum_{k=2}^{n} \left| \frac{v_k - v_{k-1}}{(v_k + v_{k-1})/2} \right| / (n-1) \right]$$

where n is the number of intervals (vowel/syllable) and v is the value of property p (duration/f0 excursion) for the k-th interval.

This means that the difference in an acoustic property between the members of successive pairs of intervals is calculated, then normalised by taking each difference as a proportion of the mean value within the pair, and averaged across the total number of interval pairs in the speech analysed [Nolan and Asu, 2009, p. 65]. The PVI was chosen over other metrics for this experiment because its normalisation procedure and regard for successive units are an advantage over utterance-global metrics like ΔV and ΔC, and it has been popular in terms of the number of studies featuring it [Nolan and Asu, 2009].

However, the PVI (like the other metrics) has not been validated externally and independently from studies concerning the traditional rhythm typology. As Nolan and Asu [2009, p. 67] point out, 'there is no absolute benchmark for [the] performance [of the metrics], beyond correlation with impressionistic judgements about languages or dialects.' Recent evaluations of the (durational) PVI have been rather negative, concluding that: it does not measure rhythm but rather timing (which is only one part of rhythm) [Arvaniti, 2009]; it does not capture the temporal distribution of prominences [Barry et al., 2009; Kohler 2009b]; it does not include other cues to prominence such as f0, amplitude or vowel quality [Barry et al., 2009; Kohler, 2009b]; and it depends to a relatively large extent on individual speakers, the style of speech, the elicitation method (read versus spontaneous speech), and the specific text which is read [Arvaniti, 2009; Kohler, 2009b; Wiget et al., 2010]. Although the present experiment does not claim to solve all these issues, particularly not the last point concerning between-speaker and between-utterance/text variation, it is hoped that the weighted PVIs will start to address at least some of these.

Most PVI studies have measured phonetically defined (vocalic/consonantal) intervals, perhaps influenced by the reasoning in two influential papers: (1) Ramus et al. [1999] developed rhythm metrics from the idea that infants perceiving rhythm have no phonological knowledge so rely on salient acoustic/phonetic properties of the signal. (2) Grabe and Low [2002], in a cross-linguistic PVI study, wanted to avoid subjective segmentation decisions based on phonological criteria for the languages in the sample which they did not speak. However, Nolan and Asu [2009] calculated syllable and foot PVIs, reasoning that 'if we assume that languages do have rhythm, it is surely reasonable to suppose that this is a property which can be informed by the phonological structure, part of which [. . .] is concerned with grouping smaller elements into larger units [e.g. syllables and feet]' [Nolan and Asu, 2009, p. 69]. These authors admitted that the syllable is often difficult to determine, but that it is central in phonological structure, which adult native speakers, whose rhythm is often under investigation, have learned.

In the present experiment, eight types of PVI are calculated: durational, tonal, combined, weighted, and each of these four types for vocalic intervals and syllables (to compare phonetic and phonological approaches). Combined PVIs incorporate durational and tonal variability, as do weighted PVIs, but the latter include language-/variety-specific weighting for each cue. Since f0 is lost during voiceless consonants, there is little point in calculating consonantal PVIs which involve f0 measurements.

2. Hypotheses

2.1. Language Groups

It is predicted that SG on the one hand and SFr and Fr on the other will have very different PVI scores, because the two languages differ in their durational and tonal

properties, and native speakers of each language attach different importance to duration and f0 as cues to rhythm. More details of the predicted difference between languages are discussed in the next section.

2.2. Types of PVI (Weighted, Durational, Tonal)

As weighted PVIs are a new concept, there are no previous weighted PVI scores on which to base predictions. However, it is possible to predict, from the previous perceptual experiment [Cumming, 2011b], the relative weighting of duration and f0 for these languages, and how the weighted PVI scores may compare with durational PVIs.

Cumming [2011b] found that an appropriate duration of prominent syllables contributed more to SG listeners' rhythmicality judgements than an appropriate f0 did, whereas an appropriate duration and f0 contributed to SFr and Fr listeners' judgements to similar extents. Therefore, it is predicted that in SG, duration has greater weighting than f0, whereas in SFr and Fr, the cues have more equal weighting. In this case, SG weighted PVIs should be similar to SG durational PVIs, whereas the SFr and Fr weighted PVIs should be somewhere between their durational and tonal PVIs.

Previous studies have reported a range of durational PVI values for various languages. It is predicted that the durational PVI is relatively high within this range for SG and relatively low for SFr and Fr. The rationale for this prediction comes from the PVI values reported in Galloway [2007] and Schmid [2001] for SG, Galloway [2007] for SFr, and Grabe and Low [2002], Lee and Todd [2004] and White and Mattys [2007] for Fr.

No previous studies have calculated tonal PVIs, so the predictions are based on general observations of each language's prosody. It is predicted that SG, SFr and Fr have tonal PVIs similar to their durational PVIs, because increased duration and substantial f0 movement co-occur in these languages. In SG, prominent syllables have phonologically long or short vowels, syllable structure often including consonant clusters, and most often a substantial f0 movement; these prominences contrast with non-prominent syllables, which do not receive large f0 excursions, and have reduced vowels and fewer consonant clusters [Häsler et al., 2005; Fleischer and Schmid, 2006]. Therefore, durational and tonal variability between prominent and non-prominent syllables is high. Fr has no phonological length contrasts or vowel reduction, and less complex syllable structure [Vaissière, 1991], hence the lower durational variability between syllables compared to SG. Yet, rhythmic-group-final syllables are lengthened, thus lowering SFr/Fr durational PVIs less than would be expected on account of the greater syllabic regularity, and this also applies to SFr/Fr tonal PVIs because of substantial rhythmic-group-final rises or falls [Di Cristo, 1999, 2000; Jun and Fougeron, 2000; Post, 2000]. If the durational and tonal PVIs for each language emerge as similar, the weighted PVIs will be similar to them.

These predictions and explanations for durational and tonal PVIs are based on properties observable in the production of each language. However, since the overall percept of rhythm is likely to be a complex interaction of various cues at various sections of speech, it may prove not to be reducible to simple observations of durational and tonal variability.

2.3. Intervals Measured (Vowel, Syllable)

No previous studies have reported syllable PVIs for SG, SFr or Fr. It is possible that syllable variability differs from vowel variability, depending on the language, as may be indicated by other studies with different languages: the syllable (durational) PVI for English reported by Nolan and Asu [2009] was much lower than the vowel PVI for English reported by White and Mattys [2007] (though similar to that reported by Grabe and Low [2002]), whereas the syllable PVI for Spanish reported by Nolan and Asu [2009] was similar to the vowel PVI for Spanish reported by White and Mattys [2007] (though the difference in PVIs here may also be due to methodological differences between studies, such as the text used to elicit the speech).

It is predicted that the vowel PVI and syllable PVI (both durational) will be more similar for both SFr and Fr than they will be for SG, in which syllable variability will be higher than vowel variability. This is because syllable variability includes consonantal as well as vocalic variability; SFr and Fr have mainly CV syllables and few consonant clusters, whereas SG syllables can range from long (e.g. a complex consonant cluster plus a diphthong) to short (e.g. a single consonant plus a reduced vowel), and various other combinations in between. How any differences in vowel and syllable variability will manifest themselves in tonal and weighted PVIs is difficult to predict, given that the hypotheses for these different PVI types are already difficult to propose since they have not previously been calculated.

3. Method

3.1. Reading Text

The text read by subjects was 'The North Wind and the Sun' in their native language (SG 'De Biiswind und d Sune' [Fleischer and Schmid, 2006]; SFr and Fr 'La bise et le soleil' [Handbook of the IPA, 1999]). This text was used in previous rhythm metric studies [e.g. Grabe and Low, 2002]. No claim is made that the findings of the present experiment are generalisable to spontaneous speech [Arvaniti, 2009].

3.2. Subjects and Procedure

Most subjects who participated in the perceptual experiment reported in Cumming [2011b] agreed to be recorded reading the text, after the listening task, when they were still naïve to the purpose. The speech data of 10 subjects (5 male, 5 female) per language were selected for acoustic analysis. The mean age of the selected SG, SFr and Fr speakers was 26.2, 21.6 and 23.9 years, respectively. All were monolingual, i.e. had not learned another language before obligatory second-language classes starting around 9–11 years old. The SG speakers were all originally from the city of Zürich, the SFr speakers were all from the Neuchâtel canton (mostly the town of Neuchâtel, though some from one other nearby town), and the Fr speakers were all from greater Paris.

The SG speakers and the Fr speakers were recorded in the sound-attenuated studios of Zürich and Cambridge University Phonetics Laboratories, respectively. SFr speakers were recorded in a quiet room in Neuchâtel University. For all recordings, the mode was 16 bit linear PCM, with a 44.1-kHz sample rate. Before the recordings, subjects were given time to read the text, and then for the recording they were instructed to read the text 4 times, speaking at a rate and in a style which was normal and comfortable for them, and resting between each repetition. Each subject's final repetition was selected for analysis, unless it contained disfluencies, in which case the most fluent of the previous ones was selected.

3.3. Analysis

3.3.1. Acoustic Measurements

The software used to analyse the recordings was Praat [Boersma and Weenik, 2009]. The criteria for vowel-consonant segmentation were those of Peterson and Lehiste [1960]. All start and end points of intervals were located at the start or end of a glottal period and at the zero-crossing on the waveform. Vowels followed or preceded by plosives, nasals, fricatives or laterals were generally unproblematic, as discrete landmarks occurred in the signal.

The French approximants /ɥ j w/ were more problematic. Phonologists regard them as consonants [e.g. Tranel, 1987]; phonetically, a smooth transition of formants usually occurs from approximant to vowel and vice versa. The boundary between approximant and vowel was determined by examining the waveform amplitude, spectrogram intensity and formant structure. The point where a fairly sudden change in amplitude/intensity occurred during the formant transition was marked as the boundary; the author's auditory perception also played a role. The French text contained two adjacent vowels, which were mostly marked as one interval, since speakers produced a diphthong-like rapid transition. In some speakers, a clear separation was evident with glottalised periods; these vowels were marked as separate (since they were more likely to be rhythmically relevant as perceived separate units), with the glottalised periods belonging to the vowel whose formant structure matched that of the glottalised periods.

Syllabic segmentation requires phonological and phonetic consideration. A phonetic and a phonological syllable have been differentiated in discussions over the concept [e.g. Fudge, 1969; Blevins, 1995]. As noted above, few rhythm-metric studies have measured syllables. In calculating syllable PVIs, Deterding [2001] took account of both phonological rules and phonetic realisations in his syllabification method for British and Singapore English. Nolan and Asu [2009] used Deterding's [2001] strategy for English, and they followed accepted phonological rules for Estonian and Spanish syllabification which they found less controversial than in English. Likewise the present experiment considered accepted syllabification rules from phonological descriptions and speakers' variable phonetic realisations, as follows:

Phonologists maintain that French prefers open syllables (i.e. syllable breaks generally occur post-vocalically/pre-consonantally) and that resyllabification also occurs across some word boundaries to maintain a (roughly) CV.CV.CV. . .. structure [e.g. Post, 2000, p. 97; Walker, 2001, p. 27; for intricacies and exceptions to these generalisations, see Walker, 2001]. Theoretically, French does not generally have consonant clusters larger than two; optional schwas are realised and epenthetic schwas are inserted to maintain this situation [Walker, 2001]. However, the (phonetically variable) pronunciation or suppression of schwas depends on many interacting phonological, morphosyntactic and stylistic factors, e.g. in informal and faster speech schwas are pronounced less often, sometimes resulting in larger consonant clusters [Walker, 2001]. The present experiment syllabified according to what was produced, e.g. different speakers pronounced *arriverait* as [a.ʁi.və.ʁɛ] or [a.ʁi.vʁɛ].

For SG, syllabification [following Fleischer and Schmid, 2006; Reese, 2007] was straightforward. Like in Fr, across some word boundaries in SG, a CVC(C) syllable followed by a VC syllable is resyllabified into CV.C(C)VC. Deterding's [2001] reason for measuring syllable (not vowel) duration was that many syllables in the conversational British English sample lacked vowels, i.e. had syllabic consonants. In the present experiment, where a syllabic [l̩] occurred (some SG speakers occasionally produced this rather than [əl] in 'Mantel') it was not treated as a separate syllable, because f0 movement was only calculated during vowels, so this syllable would have had no vocalic f0 value. The [l̩] was syllabified with the preceding [t] and following vowel ([ɒ] or [æ]), e.g. [mɒn.tlɒb.t͡so.gə].

It has been pointed out that there is an inconsistency here. On the one hand, measuring syllable properties stems from the idea that rhythm involves the phonological structure of a language, a key part of which is the syllable, which is perceptually salient for adult (as opposed to infant) listeners. On the other hand, SG syllabic consonants were not counted as a syllable (i.e. by measuring the duration and f0 of the consonant as if it were a vowel), even though an adult SG listener presumably would count them as a syllable, whereas French approximants (phonetically more vocalic than consonantal) were counted as consonants, like an adult French listener presumably would. The effect of this inconsistency in the present data set is negligible, since only a few SG speakers produced a syllabic consonant just once in the whole text.

Vowel durations (ms), syllable durations (ms), and vowel f0 excursions (semitones) were calculated. Syllable f0 excursions were not calculated, because perceptually most relevant f0 movements

occur during steady-state vowels rather than consonants [House, 1990], and f0 would be lost during many consonants due to voicelessness and transitory perturbations. The vowel f0 excursions were checked manually by visual inspection of the pitch trace on the spectrogram, and some discrepancies due to non-modal voice were corrected. Occasionally vowels were so laryngealised that the f0 excursion (and duration) of these were excluded from analysis, i.e. non-adjacent intervals were used for measurement; this is not ideal, but laryngealisation is an inherent and difficult to resolve problem in measuring pitch, and as this only occurred occasionally with these speakers, the overall difference that this would make to their PVI scores is negligible.

Excursion was calculated as the absolute difference between the minimum and maximum f0, i.e. rises and falls were not differentiated, because the PVI deals with absolute differences, and there is a lack of data bearing on whether rises are perceptually more/less prominent than adjacent equally sized falls; f0 was measured in semitones, because pitch perception does not correspond linearly to absolute decreases/increases in Hertz, and there is evidence that the best psychoacoustic scale which captures this non-linear relationship is the essentially logarithmic semitone scale [Nolan, 2003].

3.3.2. Weighting Values

The values used to weight the relative significance of duration and f0 in the PVIs for SG, SFr and Fr were derived from the results of the previously conducted perceptual experiment reported in Cumming [2011b], which found that f0 and duration are interdependent cues to the perceived rhythmicality of sentences, and that listeners' native language significantly affected the weighting of each cue.

In each trial, listeners were presented with nine stimuli; these were lexically identical sentences, but the duration and/or f0 excursion was manipulated on one of the prominent syllables. Listeners had to judge which stimulus sentence out of the nine had the most natural rhythm. The duration and f0 manipulations were designed to test whether the duration and f0 excursion of prominent syllables must both be appropriate (i.e. non-deviant) for a sentence to have a natural-sounding rhythm, and whether a deviant duration results in a less natural-sounding rhythm than a deviant f0 excursion, or vice versa. The listeners' responses showed that the most natural-sounding rhythm was most often perceived when both cues were non-deviant; however, the listeners did tolerate some deviance for a sentence to still have a natural-sounding rhythm – SG listeners often tolerated tonal deviance but hardly ever tolerated durational deviance, whereas SFr and Fr listeners often tolerated some deviance in both duration (shorter) and f0 excursion (lower).

From these listeners' responses, it was possible to find out, for each language, the relative weight that the variables 'non-deviant duration' and 'non-deviant f0 excursion' (of stimuli) contributed to whether a stimulus was perceived as having the most natural-sounding rhythm. To do this, three separate logistic regression analyses (generalised linear mixed models) were run using the *R* software environment, one on each language group's (SG, SFr, Fr) perceptual task responses.

The data input to each model were the dependent variable *response* (1 = the stimulus judged the most rhythmically natural; 0 = any of the remaining eight out of nine stimuli not judged the most rhythmically natural), and the predictor variables X_{dur} [the difference in duration (ms) between the to-be-manipulated syllable in the originally recorded sentence and this manipulated syllable in the stimulus] and X_{f0} [the difference in f0 excursion (semitones, st) between the to-be-manipulated syllable in the originally recorded sentence and this manipulated syllable in the stimulus]. The random effect *subject* was also included in the model.

Table 1 shows the b-coefficients outputted from each language's model, and the X-standardised b-coefficients calculated from these. Non-standardised b-coefficients indicate how much a 1-ms or 1-st deviation contributes to the judgement of whether the stimulus had the most natural-sounding rhythm. However, a 1-ms difference between stimuli is less perceptible than a 1-st difference. Therefore, standardisation was necessary to find the contribution of each variable relative to the other, regardless of whether duration happened to be measured in milliseconds (or centiseconds, seconds) and f0 in semitones (or Hertz, ERB rate).

All the X-standardised b-coefficients are negative, i.e. if duration or f0 excursion were deviant in a stimulus, the *less* likely it was that the stimulus was judged as having the most natural-sounding rhythm. Between languages, it is informative to compare the ratio of the duration and f0 b-coefficients for each language, rather than the absolute values of these b-coefficients. Within languages, there are differential effects of durational and tonal deviance: the more negative b-coefficient marks either duration or f0 as the more important non-deviant cue for a sentence to sound rhythmically natural. Thus

Table 1. For each language and each variable: b output from the regression model, and standardised b calculated from this

Language	Variable (X)	b coefficient (b)	sdX	X-standardised b coefficient (sdX × b)
SG	dur	–0.002**	91.04 ms	–0.150
	f0	–0.006*	2.45 st	–0.015
SFr	dur	–0.004**	55.57 ms	–0.231
	f0	–0.109**	2.45 st	–0.268
Fr	dur	–0.006**	55.57 ms	–0.346
	f0	–0.124**	2.45 st	–0.304

sdX = Standard deviation of X. **p < 0.0001, *p = 0.59.

non-deviant duration is far more important than non-deviant f0 for SG listeners (–0.150 vs. –0.015), whereas non-deviant duration and non-deviant f0 are almost equally important for SFr and Fr listeners, though duration relatively less for SFr listeners (–0.231 vs. –0.268) and relatively more for Fr listeners (–0.346 vs. –0.304). These X-standardised b-coefficients, after conversion to proportions of 1 in each language, were the weighting values used in the weighted PVIs (table 2).

3.3.3. PVI Calculations

The normalised PVI formula as presented in section 1.5 was used as the basis for all PVI calculations. As in previous studies [e.g. Grabe and Low, 2002; White and Mattys, 2007], some pairwise comparisons were over a pause. Table 3 shows that the acoustic property (p) was: duration (in milliseconds) for durational PVIs; f0 excursion (in semitones) for tonal PVIs; duration and f0 excursion in equal proportions (50:50) for combined PVIs; duration and f0 excursion weighted according to the language-specific weighting values in table 2 for weighted PVIs. (Excursion, as opposed to any other f0 measurement, was necessary because the measurement had to correspond to that from which the f0 weighting values were derived, i.e. f0 excursion manipulations in the perceptual experiment's stimuli reported in Cumming [2011b].) Table 3 also shows how the normalised PVI formula was adapted to include the weighting values for combined and weighted PVIs.

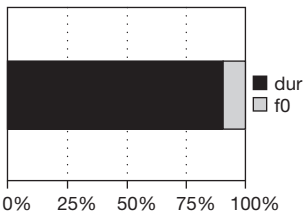
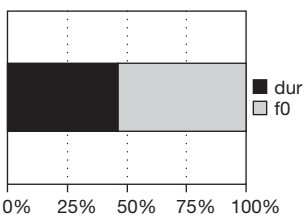
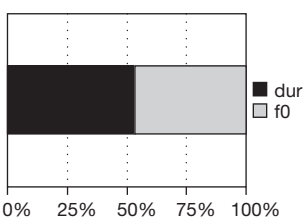
4. Results and Discussion

Since the PVIs were all normalised, the difference in measurement units between milliseconds and semitones is irrelevant when comparing the different types of PVI (i.e. they are dimensionless). A mixed-measures ANOVA was run in SPSS on the PVI scores (means plotted in fig. 1), with the factors *interval* (vowel, syllable), *PVI type* (durational, tonal, combined, weighted) and *language* (SG, SFr, Fr). All main effects and interactions were significant, the most interesting of which will be presented and discussed in the following sections, which relate to each section of the hypotheses (section 2).

4.1. Language Groups

First, the main effect of *language* [$F(2, 27) = 30.59$, $p < 0.0001$] along with post-hoc tests (Tukey HSD) showed a significant difference between SG and SFr, and SG and Fr ($p < 0.0001$), but not between SFr and Fr ($p > 0.05$). The between-language

Table 2. Weighting values used in weighted PVIs (third column from left)

Language	$\frac{b_{dur}}{b_{f0}}$	$\frac{b_{dur}}{b_{f0}}$ (if $b_{dur}+b_{f0} = 1$)	Relative weighting of $b_{dur} : b_{f0}$
SG	$\frac{-0.150}{-0.015}$	$\frac{0.908}{0.092}$	
SFr	$\frac{-0.231}{-0.268}$	$\frac{0.463}{0.537}$	
Fr	$\frac{-0.346}{-0.304}$	$\frac{0.533}{0.467}$	

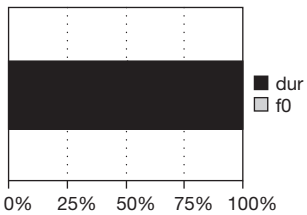
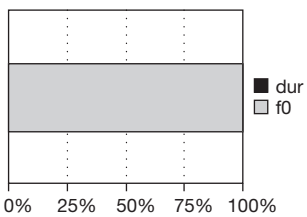
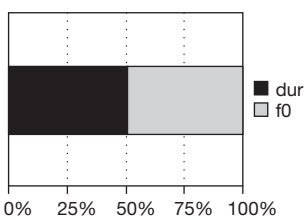
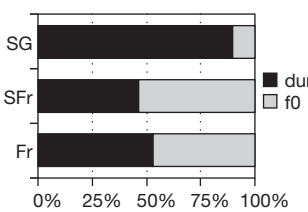
result is as predicted and is of course not surprising given how different the languages' rhythmic properties are, so this is not discussed further here. The between-variety result shows that the subtle phonetic differences between SFr and Fr were not captured by these rhythm metrics [Cumming, 2010a, for further analysis and discussion of the SFr and Fr PVIs].

4.2. Intervals Measured (Vowel, Syllable)

There was a main effect of *interval* [$F(1, 27) = 100.06$, $p < 0.0001$]: vowel PVIs were significantly higher than syllable PVIs. There was a significant interaction of *interval* × *PVI type* [$F(1.23, 33.10) = 91.80$, $p < 0.0001$]: the difference between durational PVIs and each of the other PVI types was significantly greater for syllable than vowel variability. There was also a significant interaction of *interval* × *language* [$F(2, 27) = 5.37$, $p < 0.05$]: the vowel and syllable PVIs were more different from each other for SG than for SFr and Fr. This was partly as predicted, though it was thought that SG (durational) syllable variability would be higher than vowel variability.

In terms of duration, a possible explanation for the greater variability of vowels than syllables is the differential effect on vowels and consonants of reduction in

Table 3. For four PVI types: the relative weighting of duration and f0 excursion, and how the normalised PVI formula was adapted to incorporate these weightings

PVI type	Graphical representation of weighting values	Normalised PVI formula
Durational		$100 \times \left[\sum_{k=2}^{n} \left| \frac{d_k - d_{k-1}}{(d_k + d_{k-1})/2} \right| /(n-1) \right]$ (d = duration)
Tonal		$100 \times \left[\sum_{k=2}^{n} \left| \frac{f_k - f_{k-1}}{(f_k + f_{k-1})/2} \right| /(n-1) \right]$ (f = f0 excursion)
Combined		$100 \times \left[\sum_{k=2}^{n} [c_k] /(n-1) \right]$ where $c_k = \left[\left| \frac{d_k - d_{k-1}}{(d_k + d_{k-1})/2} \right| + \left| \frac{f_k - f_{k-1}}{(f_k + f_{k-1})/2} \right| \right] /2$
Weighted		$100 \times \left[\sum_{k=2}^{n} [w_k] /(n-1) \right]$ where e.g. b_{dur}=0.8, b_{f0}=0.2 (note that these example b-coefficients add up to 1), and $w_k = \left[\left(0.8 \left| \frac{d_k - d_{k-1}}{(d_k + d_{k-1})/2} \right| \right) + \left(0.2 \left| \frac{f_k - f_{k-1}}{(f_k + f_{k-1})/2} \right| \right) \right]$

connected speech. In SG, all non-prominent (non-loanword) syllables contain [ə] or [i] which are reduced and central(ised) [Fleischer and Schmid, 2006; Reese, 2007]. In SFr and Fr, optional schwas are produced to maintain clusters of maximally two consonants, though larger clusters may occur if optional schwas are suppressed, often in faster speech (section 3.3.1). In this experiment, many speakers (both languages) sometimes produced very short [ə] or [i] (two or three glottal periods in the acoustic record) in non-prominent syllables, e.g.: SFr and Fr [sə.di.spy.tɛ] (‘were arguing’), [ki.sa.vɑ̃.sɛ] (‘who was approaching’), i.e. not quite full suppression of schwa and [i] in function words; SG [ˈtːi.k͡χə] (‘thick’), [ˈb̥ri.ŋi] (‘to bring’). The consonants in these syllables, [s] [k] [k͡χ] [ŋ], were not markedly reduced compared to the vowels. Logically then, vowel variability is greater than syllable variability,

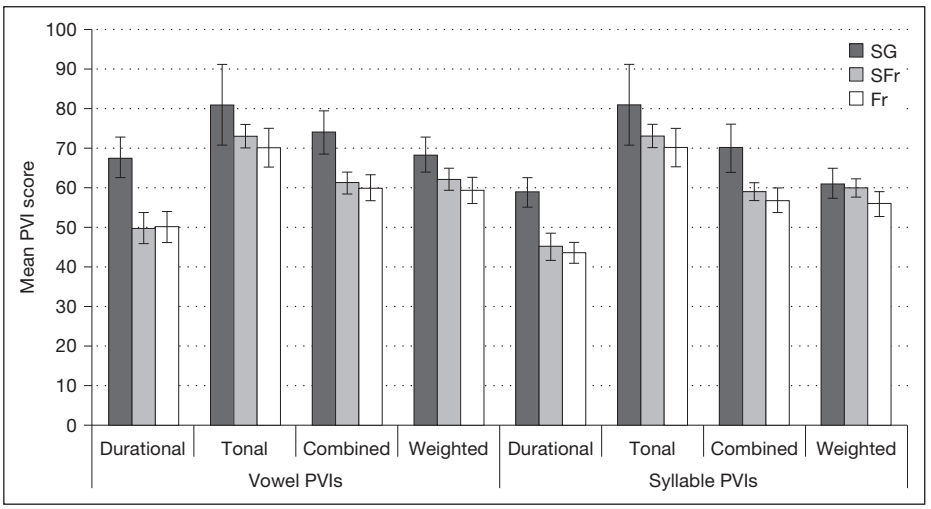


Fig. 1. Mean PVI scores (10 subjects per language); error bars show ± 1 standard deviation. Tonal PVI data are identical for vowel and syllable PVIs.

because syllable variability also includes consonants, which are less durationally variable than vowels in the speech elicited here. For weighted PVIs, durational variability contributed in about the same measure as tonal variability for SFr and Fr, and much more than tonal variability for SG, so this difference between vowel and syllable variability would impact the weighted PVIs too.

This experiment measured vowels and syllables to compare, in the same data set, these acoustic/phonetic and phonological approaches to measuring durational and tonal variability. If we measure produced rhythm (in SG, SFr and Fr) acoustically, as though listening through infant ears, successive vowel intervals show relatively high acoustic variability, also depending on language. Therefore, rhythm unconnected to phonological knowledge of a language (babies hearing speech) may have a relatively clear alternation of perceptually 'weak' and 'strong' elements. If we measure rhythm phonologically, as though listening under the influence of knowledge acquired by adulthood, successive intervals (syllables) show significantly lower acoustic variability than vowels. Therefore, rhythm connected to phonological knowledge of a language (adults hearing speech) may have a perceptually less clear 'weak'-'strong' pattern. This reduced clarity of rhythm results from the fact that meaningful speech is an acoustically and linguistically complex signal with several interacting factors determining its properties at any point. Different approaches to measuring rhythm, e.g. Grabe and Low [2002] (phonetic) and Nolan and Asu [2009] (phonological), might have measured phenomena that are to some extent perceptually distinct. Vowel (and consonant) variability is inextricably linked to syllable variability, but syllable variability is arguably a more appropriate measure if we are concerned with adults' perception of rhythm in language, because they have knowledge of their native-language phonology including syllables (and larger prosodic groups) [Nolan and Asu, 2009].

4.3. Types of PVI (Weighted, Durational, Tonal)

The main effect of *PVI type* and the two-way and three-way interactions with it were significant: *PVI type* [$F(1.08, 29.28) = 265.93$, $p < 0.0001$]; *PVI type* × *language* [$F(2.17, 29.28) = 7.88$, $p < 0.01$]; *interval* × *PVI type* × *language* [$F(2.45, 33.10) = 6.12$, $p < 0.01$]. Planned comparisons between durational PVIs and each of the other PVI types explored the main effect of, and interactions with, *PVI type*.

4.3.1. Main Effect of PVI Type

Without accounting for *interval* or *language* (i.e. if we consider just the variation attributable to *PVI type*), durational PVIs were significantly lower than all other PVI types: durational versus tonal [$F(1, 27) = 311.46$, $p < 0.0001$]; durational versus combined [$F(1, 27) = 304.73$, $p < 0.0001$]; durational versus weighted [$F(1, 27) = 613.60$, $p < 0.0001$].

The prediction that tonal PVIs would be similar to durational PVIs, because lengthening and f0 movement co-occur in SG, SFr and Fr, was not supported (note that the PVIs were normalised, so the measurement units are irrelevant when comparing them). A possible explanation for higher tonal than durational PVIs is that tonal movements tended to be either large (i.e. pitch accents) or microfluctuations less than 1 st, whereas shorter vowels, though (of course) shorter than longer ones, still had a considerable millisecond value. For instance, adjacent vowels could have a difference of 150–75 ms and 5–1 st, a ratio of 2:1 for length but 5:1 for pitch. Normalised PVIs account for differences within one measurement unit, e.g. speaking rate (duration) or pitch range (f0 excursion), but do not cover up this potentially important finding that durational variability was of lower magnitude than tonal variability.

4.3.2. PVI Type × Language Interaction

For SFr and Fr, the differences between durational and tonal PVIs, and between durational and weighted PVIs are significantly greater than for SG [$F(2, 27) = 3.48$, $p < 0.05$; $F(2, 27) = 111.32$, $p < 0.0001$]; but the difference between durational and combined PVIs misses significance [$F(2, 27) = 3.30$, $p = 0.052$].

The prediction that SG durational and weighted PVIs would be similar, whereas SFr and Fr weighted PVIs would lie between their durational and tonal ones, was supported. This prediction followed from the prediction that in SG duration would have greater weighting than f0, whereas in SFr and Fr the cues would have more equal weighting. The latter prediction was confirmed in the calculation of weighting values, and an explanation for this cross-linguistic difference in the relative weighting of duration and f0 is given in Cumming [2001b]. To summarise, it seems that SG speakers are most sensitive to the obvious durational variability of SG, whereas SFr and Fr speakers may be sensitive to a more even durational and tonal pattern across syllables, when they perceive rhythm in their native language; SFr listeners' slightly greater sensitivity to f0 than Fr listeners' may result from the difference in precise timing and placement of pitch movements in SFr and Fr.

4.3.3. Interval × PVI Type × Language Interaction

If we take into account *language* and *interval* when comparing PVI types (the three-way interaction), only combined PVIs were significantly different from durational

PVIs [$F(2, 27) = 4.10$, $p < 0.05$], and weighted PVIs were not [$F(2, 27) = 1.97$, $p > 0.05$], though the tonal versus durational PVI difference just failed to reach significance [$F(2, 27) = 3.30$, $p = 0.052$]. This means that the difference between durational and weighted PVIs (seen in the main effect of *PVI type*) can be explained as the effects of *interval* (vowel PVIs > syllable PVIs) and *language* (SG PVIs > SFr and Fr PVIs). Therefore, when it was accounted for that vowel PVIs were higher than syllable PVIs, and SG PVIs were higher than SFr and Fr PVIs, the difference between durational and weighted PVIs was not as great as the difference between tonal and durational PVIs or between combined and durational PVIs. Two important points follow from these findings.

First, the weighted PVIs were a compromise between separate measures of variability for two acoustic cues which turned out to be very different. Viewed in isolation, durational variability was much lower than tonal variability, as explained above, though both are relevant to perceived rhythm. Weighted PVIs combined these two sets of different but related information, thus capturing two dimensions of the multidimensionality of rhythm in a complex speech signal. In fact, unlike combined (non-weighted) PVIs, weighted PVIs captured another 'dimension': the language-specific perceptual relevance of each cue. Thus the difference between combined PVIs and weighted PVIs was lower in SFr and Fr, with fairly equally weighted cues, than in SG, with length weighted much higher than pitch for perceived rhythmicality.

Second, the magnitude of cross-linguistic difference shown by the PVIs was as follows: weighted PVIs < tonal PVIs < combined PVIs < durational PVIs (fig. 1). Therefore, when we quantify rhythm taking into account its acoustic multidimensionality and the language-/variety-specific relevance of its cues, we find that languages usually classified as rhythmically distinct are more similar than universal durational metrics demonstrate. An illustration of what this might mean in terms of cross-linguistic differences in rhythm is as follows: to a group of listeners with a particular native language (X) the rhythm of some other languages (Y) seems different. However, if native speakers of X could hear Y 'through the ears' of native speakers of Y, i.e. have native-like knowledge of Y's phonology including the relative importance of various rhythm cues, these native speakers of X might think that the phenomenon of rhythm shows much similarity between these languages.

4.4. Exchange of Weighting Values Cross-Linguistically

If this idea – that acoustically multidimensional language-specific rhythm metrics provide perceptually informed quantifications of rhythm which show relatively little cross-linguistic divergence – is hard to conceptualise, it may help to consider the opposite effect. If the language-specific weighting values were exchanged between languages (i.e. SG values were used to calculate SFr and Fr PVIs, and vice versa), then we would expect the weighted PVIs to show an even greater cross-linguistic difference than the traditional durational PVIs do. In order to test this, the weighted PVIs were recalculated using the other two languages' weighting values (fig. 2).

Indeed the SG PVIs weighted with SFr and Fr values, and SFr and Fr PVIs weighted with SG values were pushed further apart in numerical distance than the SG and Fr (including SFr) durational PVIs. This implies (using the example from above

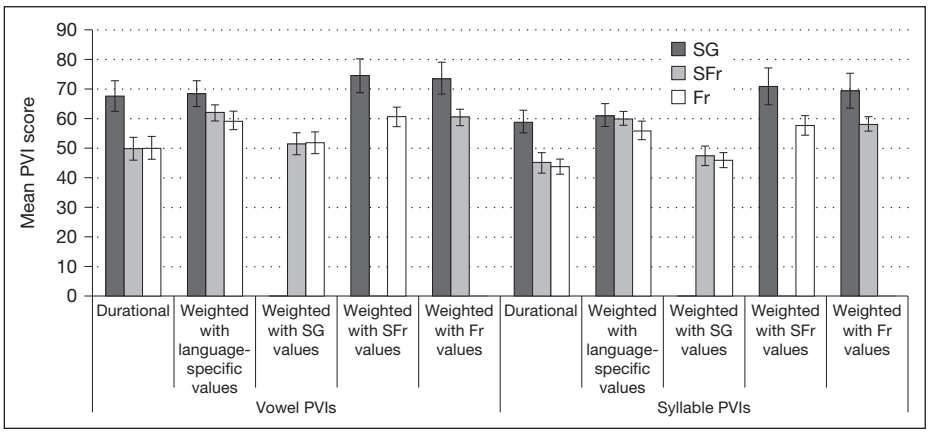


Fig. 2. Mean PVI scores (10 subjects per language) for recalculated PVIs with cross-linguistically exchanged weighting values; error bars show ± 1 standard deviation.

with language X and language Y) that a speaker of X hears language Y as rhythmically more different from X than it physically is, because the speaker of X perceives Y influenced only by the relative importance of various cues to rhythm in X, not by that in Y.

5. Evaluation of Weighted PVIs

The point of this experiment was to suggest a method for making a quantification of produced rhythm perceptually informed, by combining f0 and duration and weighting these according to their language-specific significance in perceived rhythmicality. An evaluation of this suggestion is now needed. According to Nolan and Asu [2009], previous assessments of rhythm metrics have judged performance by how well they correlate with impressionistic rhythm 'types'. Since the present experiment did not assume that a rhythm typology necessarily exists, it did not aim to better categorise these languages into types by generating new numbers. Instead, we need to ask whether weighted PVIs are better than traditional durational ones in making it possible to state what the numbers mean in terms of perception [Barry et al., 2009].

5.1. Improvement on Durational Metrics

In speech, perceived rhythm is induced by a prominence pattern, which depends on how each syllable's unique combination of duration and f0 (and potentially other cues) differs from that of its neighbours, and on the relative perceptual significance of each cue in the language concerned. This acoustic complexity is not fully represented by durational (or indeed tonal) metrics. Weighted PVIs represent a language-specific compromise between durational and tonal variability, so they more adequately reflect

the complex interaction of linguistic factors and acoustic properties that differs cross-linguistically and results in perceived rhythm in language. In sum, perceived rhythm is now, to some extent, represented in numbers: two acoustic cues and cross-linguistic variation in perception are involved. The weighting values for weighted PVIs were derived from adults' perception, who are aware of their native-language phonology. Therefore, weighted PVIs with syllables (i.e. phonologically defined intervals) mean more in terms of representing linguistic rhythm perception than weighted PVIs with vowels (i.e. acoustically defined intervals).

The weighted PVIs for SG, SFr and Fr show that if we account for the acoustic multidimensionality and language specificity of perceived rhythm when quantifying produced rhythm, the phenomenon might not be as cross-linguistically divergent as is suggested by the data from durational metrics. Cross-linguistic differences in rhythm are apparently quantified by durational metrics, but these metrics might inappropriately capture cross-linguistic differences that are less significant in the rhythm *perceived* by native speakers of a language than the similarities captured by multidimensional language-specific metrics.

That is not to say that durational metrics are not at all informative for phonetic research – some researchers with interests other than cross-linguistic differences in rhythm (e.g. speech timing effects) might be interested in variation in metric scores (e.g. between speakers or speech elicitation methods), and thus would not find the cross-linguistic convergence observed in weighted PVIs particularly enlightening. Moreover, although current durational metrics have clear limitations for the study of rhythm, it is possible that more sophisticated means of investigating purely durational variability could be developed, which might provide insight into how perceptual investigation of rhythm would best proceed.

5.2. Further Room for Improvement of Weighted PVIs

Although weighted PVIs have advanced from traditional durational metrics which told nothing of perceived rhythm, the progress is just one small step. There are some issues with the concept of weighted PVIs that could be addressed in future research. One potential issue is that for syllable PVIs, f0 excursion was only measured across the vowel, but for good theoretical and practical reasons: perceptually relevant f0 movements occur in steady-state vowels [House, 1990], and f0 is lost during many consonants, neither of which affects duration. However, the peak of an f0 rise signalling prominence may occur beyond the prominent vowel itself (or the entire prominence-signalling rise or fall may occur on a syllable adjacent to the one perceived as prominent), as demonstrated by e.g. Fitzpatrick-Cole [1999] for SG and several papers by Ladd and colleagues for other languages [e.g. Ladd et al., 1999; Atterer and Ladd, 2004]. Therefore, further perceptual experiments would be necessary to know how listeners use the alignment of f0 rises and falls as cues to prominence, and this will differ between languages.

Another issue concerning pitch measurement is that the PVI treats all falls and rises equally, despite the fact that some f0 movements are linguistically relevant while others are not, and some cue prominence while others cue boundaries of groupings in speech (though both prominence and boundaries are important in perceived grouping and therefore rhythm). PVIs measuring syllable tonal variability (including weighted

PVIs) ignore these subtle differences in the purpose of pitch contours in speech, which are potentially important in the perception of rhythm, especially since it is argued that the PVIs should be linguistically relevant. It would be interesting and potentially fruitful to measure tonal variability between longer prosodic (e.g. rhythmic) groups, since these might also capture some of the variability relevant to perceived rhythmic structure.

The positive evaluation of weighted PVIs in section 5.1 is based on data concerning only two acoustic cues, from only two languages (including two varieties of one). Other acoustic cues like amplitude and spectral properties are also likely to interact to some extent with duration and f0 in rhythm perception. Thus an even better representation of perceived rhythm would be PVIs that include language-specific weighting of more cues. Wider research with many more prosodically diverse languages is needed to fully assess the usefulness of weighted PVIs.

If such an extensive cross-linguistic investigation involving more acoustic cues were conducted, it is likely that the relative weighting of multiple cues to rhythm would differ between languages, though some would show more similar relative weightings than others. As demonstrated in this experiment, two varieties of the same language had similar but subtly different weighting of duration and f0, which we might expect to be the case for some pairs of languages. It is unlikely, though, that languages would neatly group into 'types' of relative weightings, but rather there would be a complex continuum ranging from those which show one highly dominant cue (like SG did here when just two cues were investigated), to those which show fairly equal weighting of multiple cues (like Fr did here), and various permutations of relative weightings in between. A prediction to test is that weighted PVIs of various languages would converge, i.e. be more similar to each other than their durational PVIs. Although an interpretation was given for the interesting finding of converging weighted PVIs observed in this experiment, it cannot be confirmed with only two languages that this is a general cross-linguistic tendency.

The statistical method suggested in this experiment for linking perceptual findings to production data, by developing an already popular method for quantifying rhythm, was a first attempt and has room for improvement. Currently a perceptual experiment like that reported in Cumming [2011b] needs to be run to obtain language-specific cue weightings. Since a cross-linguistically extensive investigation is proposed, this would be a time-consuming task, though it could be shared by several researchers who could each adopt the same method to make results comparable. Large-scale cross-linguistic rhythm research linking perception and production should not be avoided just because it is challenging. In fact, this research is precisely what is needed, particularly in light of the outcomes of this series of experiments.

Within this future research linking rhythm perception and production, if rhythm metrics remain in use, they need to integrate acoustic measurements in a perceptually motivated way, like the weighted PVIs. The aim of the weighted PVIs was not to 'jump on the bandwagon [of rhythm-metric experiments]' [Kohler, 2009b, p. 6], but to try and steer the bandwagon around to another possible direction, which might lead to a more enlightening destination. However, weighted PVIs are not necessarily the best method for investigating rhythm, and, along with other metrics, their validity in achieving the ultimate goal of better understanding rhythm should continue to be thoroughly questioned. They could be a helpful tool, amongst other experimental techniques, each a means to an end, not an end in itself.

6. Conclusion

The experiment reported in this paper followed on from two previous rhythm perception experiments [Cumming, 2010b, 2011b]. Two major implications of these experiments' findings for current duration-based rhythm research are that it should investigate not just duration but also f0, and consider that rhythm *perception* differs cross-linguistically, to avoid a 'one-size-fits-all' method when comparing cross-linguistic rhythmic differences in production. These two problems with rhythm metrics in their current state were the motivation behind the adaptation of the durational PVI to acoustically multidimensional language-specifically weighted PVIs.

The main conclusion that can be drawn from these different PVIs is that the cross-linguistic pattern of rhythm production observed with weighted PVIs is noticeably less divergent than the cross-linguistic pattern observed with durational PVIs. SG and Fr (including SFr) are rhythmically distinct according to durational PVIs, but rhythmically more similar according to weighted PVIs. Furthermore, SG and Fr (including SFr) are rhythmically even more distinct according to PVIs weighted with the other languages' relative weightings of duration and f0. Therefore, according to acoustically multidimensional language-specific metrics, rhythm may be less cross-linguistically divergent than durational metrics suggest. Moreover, the cross-linguistic differences captured by durational metrics may be less significant in the rhythm *perceived* by native speakers of a language than the similarities captured by multidimensional language-specific metrics.

It can also be concluded that it is now possible to represent in numbers produced rhythm along with perceived rhythm. The weighted PVIs are to a certain extent perceptually informed – they reflect the complex interaction of two acoustic properties, as well as their relative importance in each language, which all contributes to perceived rhythm. The weighted PVIs would be even more perceptually informed if they involved more acoustic cues to rhythm, and the conclusions drawn from these PVIs would be more robust if they were applied to many more prosodically diverse languages.

7. Acknowledgements

I would like to thank Francis Nolan, Bill Barry, Brechtje Post, Spyros Armosti and Hae-Sung Jeon for comments, suggestions and discussion of the work at various stages, two reviewers for their helpful comments on an earlier version of this article, and Stephan Schmid and Daniel Elmiger for providing locations to test participants in Switzerland. This work forms part of research conducted for the author's PhD which was funded by an AHRC doctoral award.

References

Arvaniti, A.: Rhythm, timing, and the timing of rhythm. Phonetica *66:* 46–63 (2009).

Atterer, M.; Ladd, D.R.: On the phonetics and phonology of 'segmental anchoring' of F0: evidence from German. J. Phonet. *32:* 177–197 (2004).

Barry, W.J.; Andreeva, B.; Koreman, J.: Do rhythm measures reflect perceived rhythm? Phonetica *66*: 78–94 (2009).

Barry, W.J.; Andreeva, B.; Russo, M.; Dimitrova, S.; Kostadinova, T.: Do rhythm measures tell us anything about language type? 15th Int. Congr. Phonet. Sci.,, Barcelona 2003, pp. 2693–2696.

Bertinetto, P.M.; Bertini, C.: On modeling the rhythm of natural languages. Speech Prosody 2008, Campinas 2008.

Blevins, J.: The syllable in phonological theory; in Goldsmith, The handbook of phonological theory, pp. 206–244 (Blackwell, Oxford 1995).
Boersma, P.; Weenik, D.: Praat: doing phonetics by computer (version 5.1.04). Computer program, retrieved 10th April 2009, from http://www.praat.org/ (2009).
Classe, A.: The rhythm of English prose (Blackwell, Oxford 1939).
Cumming, R.E.: Speech rhythm: the language-specific integration of pitch and duration; doct. thesis University of Cambridge (2010a).
Cumming, R.E.: The interdependence of tonal and durational cues in the perception of rhythmic groups. Phonetica *67:* 219–242 (2010b).
Cumming, R.E.: The effect of dynamic fundamental frequency on the perception of duration. J. Phonet. *39:* 375–387 (2011a).
Cumming, R.E.: The language-specific interdependence of tonal and durational cues in perceived rhythmicality. Phonetica *68:* 1–25 (2011b).
Dellwo, V.: Rhythm and speech rate: a variation coefficient for DC; in Karnowski, Szigeti, Language and language processing. Proc. 38th Linguistics Colloquium, Piliscsaba 2003, pp. 231–241 (Lang, Frankfurt 2006).
Deterding, D.: The measurement of rhythm: a comparison of Singapore and British English. J. Phonet. *29:* 217–230 (2001).
Di Cristo, A.: Vers une modélisation de l'accentuation du français: première partie. French Lang. Stud. *9:* 143–179 (1999).
Di Cristo, A.: Vers une modélisation de l'accentuation du français: deuxième partie. French Lang. Stud. *10:* 27–44 (2000).
Ferragne, E.: Étude phonétique des dialectes modernes de l'anglais des îles britanniques: vers l'identification automatique du dialect; doct. thesis Université Lyon 2 (2008).
Fitzpatrick-Cole, J.: The Alpine intonation of Bern Swiss German. 14th Int. Congr. Phonet. Sci., San Francisco 1999, pp. 941–944.
Fleischer, J.; Schmid, S.: Illustrations of the IPA: Zurich German. J. Int. Phonet. Ass. *36:* 243–253 (2006).
Fletcher, J.: The prosody of speech: timing and rhythm; in Hardcastle, Laver, Gibbon, Handbook of the phonetic sciences; 2nd ed., pp. 523–602 (Blackwell, Oxford 2010).
Fudge, E.C.: Syllables. J. Ling. *5:* 253–286 (1969).
Galloway, R.E.: Bilinguals' interacting phonologies? A study of speech production in French-Swiss German bilinguals; MPhil thesis University of Cambridge (2007).
Grabe, E.: Variation adds to prosodic typology. Speech Prosody 2002, Aix-en-Provence 2002.
Grabe, E.; Low, E.L.: Durational variability in speech and the rhythm class hypothesis; in Warner, Gussenhoven, Papers in Laboratory Phonology. VII, pp. 515–543 (Mouton de Gruyter, Berlin 2002).
Grabe, E.; Post, B.; Watson, I.M.C.: The acquisition of rhythmic patterns in English and French. 14th International Congress of Phonetic Sciences, San Francisco 1999.
Handbook of the International Phonetic Association (Cambridge University Press, Cambridge 1999).
Häsler, K.; Hove, I.; Siebenhaar, B.: Die Prosodie des Schweizerdeutschen – Erkenntnisse aus der sprachsynthetischen Modellierung von Dialekten. Linguistik Online *24:* 187–224 (2005).
House, D.: Tonal perception in speech (Lund University Press, Lund 1990).
Jun, S.-A.; Fougeron, C.: A phonological model of French intonation; in Botinis, Intonation: analysis, modelling and technology, pp. 209–242 (Kluwer Academic, Dordrecht, London 2000).
Kohler, K.J.: Rhythm in speech and language: a new research paradigm. Phonetica *66:* 29–45 (2009a).
Kohler, K.J.: Whither speech rhythm research? Phonetica *66:* 5–14 (2009b).
Ladd, D. R.; Faulkner, D.; Faulkner, H.; Schepman, A.: Constant 'segmental anchoring' of F0 movements under changes in speech rate. J. acoust. Soc. Am. *106:* 1543–1554 (1999).
Lee, C.S., Todd, N.P.M.: Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora. Cognition *93:* 225–254 (2004).
Low, E.L.: Intonation patterns in Singapore English; MPhil thesis University of Cambridge (1994).
Low, E.L.: Prosodic prominence in Singapore English; doct. thesis University of Cambridge (1998).
Low, E.L.; Grabe, E.; Nolan, F.: Quantitative characterizations of speech rhythm: syllable-timing in Singapore English. Lang. Speech *43:* 377–401 (2000).
Miller, J.S.: Swiss French prosody: intonation, rate and speaking style in the Vaud canton; doct. thesis University of Illinois at Urbana-Champaign (2007).
Niebuhr, O.: F0-based rhythm effects on the perception of local syllable prominence. Phonetica *66:* 95–112 (2009).
Nolan, F.: Intonational equivalence: an experimental evaluation of pitch scales. 15th Int. Congr. Phonet. Sci., Barcelona 2003.
Nolan, F.; Asu, E.L.: The Pairwise Variability Index and coexisting rhythms in language. Phonetica *66:* 64–77 (2009).
Pamies Bertrán, A.: Prosodic typology: on the dichotomy between stress-timed and syllable-timed languages. Lang. Design *2:* 103–130 (1999).
Payne, E.; Post, B.; Astruc, L.; Prieto, P.; del Mar Vanrell, M.: Rhythmic modification in child directed speech. Oxford Univ. Working Papers Ling., Phil. Phonet. *12:* 123–144 (2009).
Peterson, G.E.; Lehiste, I.: Duration of syllable nuclei in English. J. acoust. Soc. Am. *32:* 693–703 (1960).
Post, B.: Tonal and phrasal structures in French intonation; doct. thesis Catholic University of Nijmegen (2000).

Ramus, F.; Dupoux, E.; Mehler, J.: The psychological reality of rhythm classes: perceptual studies. 15th Int. Congr. Phonet. Sci., Barcelona 2003, pp. 1–6.
Ramus, F.; Nespor, M.; Mehler, J.: Correlates of linguistic rhythm in the speech signal. Cognition *73:* 263–292 (1999).
Reese, J.: Swiss German: the modern Alemannic vernacular in and around Zurich (Lincom Europa, Munich 2007).
Schmid, S.: Un nouveau fondement phonétique pour la typologie rhythmique des langues? Workshop for the 10ème anniversaire du Laboratoire d'Analyse Informatique de la Parole (LAIP), Lausanne 2001.
Steele, J.: Prosodia rationalis: or, an essay towards establishing the melody and measure of speech (Nichols, London 1775).
Tilsen, S.; Johnson, K.: Low-frequency Fourier analysis of speech rhythm. J. acoust. Soc. Am. *124:* EL34-EL39 (2008).
Tranel, B.: The sounds of French (Cambridge University Press, Cambridge 1987).
Vaissière, J.: Rhythm, accentuation and final lengthening in French; in Sundberg, Nord, Carlson, Music, language, speech and brain. Proc. int. Symp. at the Wenner-Gren Center, Stockholm 1990, pp. 108–120 (1991).
Walker, D.C.: French sound structure (University of Calgary Press, Alberta 2001).
White, L.; Mattys, S.: Calibrating rhythm: first language and second language studies. J. Phonet. *35*: 501–522 (2007).
White, L.; Mattys, S.; Series, L.; Gage, S.: Rhythm metrics predict rhythmic discrimination. 16th Int. Congr. Phonet. Sci., Saarbrücken 2007., pp. 1009–1112.
Wiget, L.; White, L.; Schuppler, B.; Grenon, I.; Rauch, O.; Mattys, S.: How stable are acoustic metrics of contrastive speech rhythm? J. acoust. Soc. Am. *127:* 1559–1569 (2010).