

How stable are acoustic metrics of contrastive speech rhythm?

Lukas Wiget, Laurence White,^{a)} Barbara Schuppler, Izabelle Grenon, Olesya Rauch, and Sven L. Mattys

Department of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol BS8 1TU, United Kingdom

(Received 10 December 2008; revised 22 December 2009; accepted 22 December 2009)

Acoustic metrics of contrastive speech rhythm, based on vocalic and intervocalic interval durations, are intended to capture stable typological differences between languages. They should consequently be robust to variation between speakers, sentence materials, and measurers. This paper assesses the impact of these sources of variation on the metrics %V (proportion of utterance comprised of vocalic intervals), VarcoV (rate-normalized standard deviation of vocalic interval duration), and nPVI-V (a measure of the durational variability between successive pairs of vocalic intervals). Five measurers analyzed the same corpus of speech: five sentences read by six speakers of Standard Southern British English. Differences between sentences were responsible for the greatest variation in rhythm scores. Inter-speaker differences were also a source of significant variability. However, there was relatively little variation due to segmentation differences between measurers following an agreed protocol. An automated phone alignment process was also used: Rhythm scores thus derived showed good agreement with the human measurers. A number of recommendations for researchers wishing to exploit contrastive rhythm metrics are offered in conclusion.

© 2010 Acoustical Society of America. [DOI: 10.1121/1.3293004]

PACS number(s): 43.70.Fq, 43.70.Bk, 43.70.Kv, 43.70.Mn [MSS]

Pages: 1559–1569

I. INTRODUCTION

A. Metrics of contrastive speech rhythm

The study of speech rhythm, once a search for isochronous units of speech (e.g., [Lehiste, 1977](#)), has more recently focused on cross-linguistic variation in durational contrast between stressed and unstressed syllables. [Dauer \(1983\)](#) observed that certain phonetic and phonotactic regularities of syllable construction tend to co-occur between languages, at least when one considers the Romance and Germanic languages of Western Europe. Thus, for example, Spanish, French, and Italian have relatively limited clustering of consonants in onsets and codas, with open consonant-vowel syllables being the predominant pattern, whereas English, Dutch, and German allow more complex consonant clusters, with stress tending to occur on heavy syllables, such as those with complex codas. Furthermore, although all of these Western European languages show lengthening of vowels in stressed syllables, Romance languages—Spanish being the most clear-cut example—have attenuated stress-related vowel lengthening compared to the Germanic languages.

Various acoustic metrics have been devised based on these observations (see [Table I](#) for definitions of the metrics discussed here), all relying on measurement of the duration of vocalic and consonantal intervals within utterances (e.g., [Ramus et al., 1999](#); [Low et al., 2000](#)). Being acoustically based, these intervals take no account of phonological structure, so that coda consonants and immediately following onset consonants are included in the same interval, as are ad-

acent heterosyllabic vowels (see [Ramus et al., 1999](#), for a discussion of the rationale for this procedure).

These rhythm metrics are intended to capture the stable differences between and within languages in degree of temporal stress contrast, as elucidated by [Dauer \(1983\)](#). Given this assumption, such metrics are best at gauging “contrastive” rhythm, i.e., the balance of strong and weak elements in speech, rather than “dynamic” speech timing, i.e., the temporal arrangement of groups of sounds according to a higher-level structure.

[White and Mattys \(2007a, 2007b\)](#) compared the efficacy of various contrastive rhythm metrics in discriminating between languages held to differ in degree of temporal stress contrast. They also looked for evidence of such differences between varieties of English. In those studies, the most effective metrics were found to be %V (the proportion of total utterance duration made up of vocalic rather than consonantal intervals) and VarcoV (the coefficient of variation of vocalic interval duration, i.e., the standard deviation divided by mean vocalic interval duration, to normalize for articulation rate). Another rate-normalized metric of vocalic interval duration, nPVI-V ([Low et al., 2000](#)—see [Table I](#)), was also useful and, unsurprisingly, highly correlated with VarcoV, though it manifested somewhat less discriminatory power.

To illustrate the interpretation of contrastive rhythm metrics, [Fig. 1](#) summarizes results for VarcoV and %V from these and related studies ([White and Mattys, 2007a, 2007b](#); [White et al., 2009](#)). It is assumed that the high VarcoV scores for English, particularly standard Southern British English (SSBE), and for standard Dutch are a reflection of the strong marking of stress by vowel lengthening in these languages. The lower VarcoV scores for French, Italian, and—in particular—Spanish are assumed to reflect smaller durational

^{a)}Author to whom correspondence should be addressed. Electronic mail: laurence.white@bristol.ac.uk

TABLE I. The rhythm metrics considered in the present study.

Metric	Description	Main references
ΔV	Standard deviation of vocalic interval duration.	Ramus <i>et al.</i> (1999)
ΔC	Standard deviation of consonantal interval duration.	Ramus <i>et al.</i> (1999)
%V	Percent of total utterance duration composed of vocalic intervals.	Ramus <i>et al.</i> (1999)
VarcoV	Coefficient of variation of vocalic interval duration (i.e., standard deviation of vocalic interval duration divided by the mean), multiplied by 100.	Dellwo (2006); White and Mattys (2007a, 2007b)
VarcoC	Coefficient of variation of consonantal interval duration (i.e., standard deviation of consonantal interval duration divided by the mean), multiplied by 100.	Dellwo (2006); White and Mattys (2007a, 2007b)
nPVI-V	Normalized pairwise variability index for vocalic intervals. Mean of the differences between successive vocalic intervals divided by their sum, multiplied by 100.	Low <i>et al.</i> (2000); Grabe and Low (2002)
rPVI-C	Pairwise variability index for consonantal intervals. Mean of the differences between successive consonantal intervals.	Low <i>et al.</i> (2000); Grabe and Low (2002)
nPVI-VC	Normalized pairwise variability index for summed vocalic and consonantal intervals. Mean of the differences between successive vocalic+consonantal intervals divided by their sum, multiplied by 100.	Liss <i>et al.</i> (2009)

differences between stressed and unstressed vowels. English and Dutch score low on %V, which is interpreted as indicating the preponderance of consonant clusters compared with the Romance languages. Naturally, %V must also be influenced by patterns of vowel duration, and the shortness of unstressed vowels in Dutch and English is assumed also to contribute to their low %V scores. The influence of vowel duration on both metrics is one factor underpinning the negative correlation between %V and VarcoV, together with the fact that—at least in the languages studied thus far—the trends for vowels and consonants tend to co-occur, as described by Dauer (1983). Thus, the overall picture from such studies is of a gradient variation in temporal stress contrast from low (Castilian Spanish) to high (standard Southern British English).

B. Comparison of contrastive rhythm scores between studies

In recent years, contrastive rhythm metrics have been increasingly applied to address a range of issues: E.g., Carter

(2005) looked at the influence of Spanish on the rhythm of Hispanic American English; Patel *et al.* (2006) tested the notion that musical rhythm reflects the rhythm of a culture’s native language; and Liss *et al.* (2009) attempted to discriminate between different types of dysarthric speech. One significant problem in interpreting and comparing such results is that absolute rhythm scores can vary widely for the same languages between published studies. For example, in the studies of Ramus (2002), Grabe and Low (2002), and White and Mattys (2007a), nPVI-V ranged from 30 to 42 for Castilian Spanish and from 63 to 73 for standard Southern British English; for rPVI-C, Grabe and Low’s (2002) score for Spanish was comparable (58) to Ramus’ (2002) score for English (57), despite the strong assumption of much greater consonantal interval variation in English (see Table I for definitions of these metrics).

If contrastive rhythm metrics, as stated above, are intended to capture stable differences between and within languages, such variation in scores represents a significant empirical and theoretical problem. This paper is intended to ascertain whether the variation is merely procedural and can be controlled through improved experimental design, or whether it represents a challenge either to the assumption that languages have stable contrastive rhythmic properties and/or to the assumption that the commonly used rhythm metrics reflect these contrastive rhythmic properties.

There are several obvious potential sources of variation in rhythm scores between studies: The speakers measured, the linguistic materials spoken, and the protocols used for measuring interval duration, together with individual differences between measurers in the application of these protocols. Here we assess the relative variation due to each potential source.

With regard to speakers, there will clearly be variation within any linguistic community in the realization of stress contrasts, as with any segmental or suprasegmental feature, but these differences should be relatively small in comparison with those between languages held to have differing degrees of temporal stress contrast. Variation in articulation

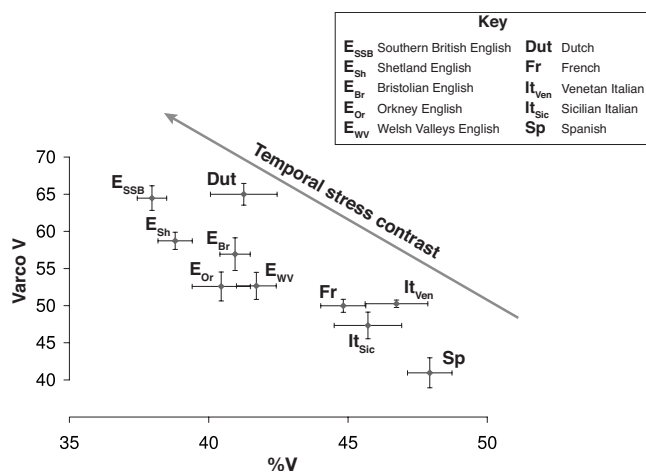


FIG. 1. VarcoV by %V plot showing the scores for Dutch (White and Mattys, 2007a), French (White and Mattys, 2007a), Spanish (White and Mattys, 2007a), two varieties of Italian (White *et al.*, 2009), and several varieties of English (White and Mattys, 2007b).

rate is one important factor that might cause significant deviations even within a fairly homogenous accent group, and it is known that certain non-normalized metrics of variation in vowel and consonant interval duration— ΔV , ΔC , rPVI-C—manifest inverse relationships with articulation rate (i.e., higher scores at lower rates—White and Mattys, 2007a). The metrics with most discriminative power—VarcoV, %V, and nPVI-V as discussed above—are robust to variation in rate, however, and so this potential source of between-speaker variation should not be an issue.

Studies also differ in the speech materials used. Given that rhythm metrics are, at least in part, a reflection of phonological structure, variation in rhythm scores is to be expected according to the linguistic materials recorded. In the extreme case, it is possible, for example, to construct English sentences without consonant clusters, so that they mimic the predominant consonant-vowel-consonant-vowel alternation of a Romance language such as Spanish. In this artificial situation, scores for metrics of consonantal interval variation are indeed more similar for English and Spanish than with more natural unconstrained sentences, as are %V scores (Prieto *et al.*, 2009). Of course, while such contrived sentences serve to make a particular point about rhythm metrics, the ideal materials for contrastive rhythm studies should be a representative reflection of the phonological structures of the languages under consideration. The usual—implicit—solution to this problem has been to use a pseudo-random set of naturalistic sentences, although studies such as that of White and Mattys (2007a) have eliminated approximants from the sentence materials in the interest of consistent application of criteria for the identification of segment boundaries.

With regard to measurers, the extraction of interval durations from speech relies mostly on manual inspection of waveforms and spectrograms, and is therefore subject to the vagaries of individual measurers, who may use different criteria or apply common criteria idiosyncratically. The use of experienced phonetically trained measurers and agreement on a set of measurement criteria between studies (see, for example, Turk *et al.*, 2006) are clear first steps toward the maintenance of consistency. Avoidance of awkward segmental junctures—e.g., approximant-vowel or vowel-approximant boundaries—may also help. This study aims to determine the degree of variability in rhythm metric scores between experienced measurers working to a common segmentation protocol.

One potentially useful method of eliminating inter-measurer variability in speech timing research is the use of an automated phone alignment process based on speech recognition algorithms. In studies using read speech, the recognition software is provided with an orthographic or broad phonetic transcription of the sentences, and identifies the boundaries assumed to be present in the signal given this transcription. The practical benefits of such an approach are that a very large amount of speech can be analyzed at little time cost and with consistency of application of segmentation criteria, in contrast with the laborious and error-prone manual approach. Here we examine whether current methods of automated phone alignment—based on statistical proper-

ties of the signal rather than phonetically defined criteria—are sufficiently reliable to provide consistency of rhythm scores compared to trained human measurers applying agreed phonetic-based protocols.

C. Purpose of experiment

In this paper, we investigate the robustness of various rhythm metrics to variation in speakers, materials, and measurers, based on an analysis of six speakers of SSBE reading five scripted sentences each. The SSBE sentences were segmented into vocalic and intervocalic intervals by five measurers, all phoneticians trained in the identification of speech segment boundaries. We also used an automatic phone alignment process on the SSBE sentences to derive scores for the same rhythm metrics.

As discussed above, two metrics have been found to be particularly useful (White and Mattys, 2007a, 2007b): VarcoV, the coefficient of variation of vowel interval duration, and %V, the proportion of utterance duration that is vocalic. The focus of the present study is therefore on these two metrics, together with the widely used nPVI-V (normalized pairwise variability index of vowel interval duration), which we found previously to be highly correlated with VarcoV, while being somewhat less discriminative. As a matter of record, however, the primary analyses were carried out for all rhythm metrics listed in Table I, and are reported in the Appendix. We include the results for nPVI-VC, a metric not utilized by White and Mattys (2007a, 2007b), but subsequently found useful in the classification of dysarthric speech by Liss *et al.* (2009).

II. METHOD

A. Participants and measurers

Six speakers, three females and three males, of Standard Southern British English were recorded. None reported any speech or hearing problems. The measurers were all trained phoneticians.

B. Materials

The five English sentences were the same as those used by White and Mattys (2007a) and are listed here in Appendix A. The sentences were constructed to avoid the approximants /l/, /r/, /j/, and /w/, for reasons discussed in Sec. I B.

C. Recordings

The sentences were recorded at the University of Bristol as part of a longer session comprising (1) short story reading, (2) map description, and (3) sentence reading (the five sentences in the present study preceded by five others). Participants were given time to rehearse the sentences silently before reading them aloud and were instructed to speak in their normal conversational voice at a rate that felt natural and comfortable. Sentence readings that contained errors or disfluencies were repeated. Recordings, direct to disk at a sampling rate of 32 kHz, were made in a sound-attenuated room using a high-quality microphone.

D. Measurements

The boundaries between vocalic and consonantal intervals of the Standard Southern British English sentences were labeled by five phonetically trained measurers (M1–M5). Prior to labeling, the measurers agreed to follow the criteria adopted by [White and Mattys \(2007a\)](#). (Note that the results for measurer M1 are those originally reported in [White and Mattys, 2007a](#).) As these criteria were the sole common guide for the measurers in this study, they are reproduced verbatim.

“This procedure was carried out with reference to standard criteria (e.g., [Peterson and Lehiste, 1960](#)); where labels were associated with the start or end of pitch periods, they were placed at the point of zero crossing on the waveform.

The primary determiner of the placement of a vowel-consonant boundary was the end of the pitch period preceding a break in formant structure associated with a significant drop in waveform amplitude. Additional criteria which facilitated the location of the boundary in certain contexts included:

- Where the vowel offset was glottalized, a change in the shape of successive pitch periods, for example, lengthening or doubling.
- Before fricatives, the onset of visible frication.
- Before nasals, the appearance of nasal formant structure and a waveform amplitude minimum.

The consonant-vowel boundary was the beginning of the pitch period at the onset of vocalic formant structure, where this was associated with the appearance of pitch periods consistent with the body of the vowel (e.g., unfricated and of comparable amplitude). Aspiration following stop release was therefore included within the consonantal interval.” ([White and Mattys, 2007a](#), pp. 506–507).

Following the practice in [White and Mattys, 2007a](#), pauses and disfluencies within utterances were excluded and vowel-vowel or consonant-consonant intervals on either side of the pause were combined. The measurers did not discuss specific labeling criteria, nor did they examine each other’s measurements at any stage prior to extraction and analysis of the resulting durational data. These precautions were intended to ensure that the labeling differences between measurers can be regarded as representative of the differences that occur when independent studies use comparable labeling criteria, but different measurers.

The recorded utterances were also subjected to an automatic phone alignment process. Based on an orthographic transcription of each sentence, a forced alignment was carried out with the open-source speech recognition tool SPRAAK ([Demuyne et al., 2008](#)). Forty-one tri-state monophonic Gaussian mixture models were trained on the British English SpeechDat FDB telephone speech database ([Draxler et al., 1998](#)), separately for male and female speakers. As a 32 ms Hamming window with a step size of 10 ms was used, the duration of the aligned segments is in multiples of 10 ms with a minimum duration of 30 ms. The lexicon used in the forced alignment process was derived from the orthographic transcription using the CELEX lexical database

([Baayen et al., 1995](#)). The resulting phone alignments were then transformed into alternations of vocalic and intervocalic intervals, in order to bring them into the same format as the segmentations by the human measurers.

Scores for the metrics %V, ΔV , ΔC , VarcoV, VarcoC, nPVI-V, rPVI-C, and nPVI-VC (see [Table I](#)) were then computed, separately for the five read sentences produced by each of the six speakers. For the sake of consistency, utterance-initial consonants, where they occurred, were excluded from the measurements.

E. Statistical analysis

We report on the variability in contrastive rhythm scores according to speaker, sentence materials, and measurers. For the analysis of the variability due to speakers, each speaker contributed 30 scores to each rhythm metric (five sentences and six measurers). For the analysis of the variability due to sentence materials, each sentence contributed 36 scores to each rhythm metric (six speakers and six measurers). For the analysis of the variability due to measurers, each measurer produced 30 scores per rhythm metric (six speakers and five sentences). The intraclass correlation coefficients (ICCs) were computed with the R environment’s IRR package ([Gamer et al., 2007](#)). We used a two-way model, which treats both raters and items as randomly sampled from a population ([Shrout and Fleiss, 1979](#)). Intraclass correlation coefficients of 0.40–0.59 are commonly regarded as indicating moderate inter-rater reliability, values of 0.60–0.79 as substantial, and values of 0.80 or larger as outstanding ([Garson, 2009](#)). In addition, in order to provide some sense of whether differences between speakers, sentences, and measurers are potentially problematic for interpretation, we compared the size of those differences to the largest difference reported between languages in [White and Mattys \(2007a\)](#) as a benchmark, namely, the difference between scores for Castilian Spanish and Standard Southern British English (see [Fig. 1](#) for VarcoV and %V values).

III. RESULTS AND DISCUSSION

As discussed above, we focus on the rhythm metrics found by previous studies ([White and Mattys, 2007a, 2007b](#)) to be the most discriminative between and within languages: VarcoV, the coefficient of variation of vocalic interval duration (standard deviation divided by the mean); nPVI, the normalized pairwise variability index of vocalic interval duration (see [Table I](#) for fuller definition); and %V, the proportion of total utterance duration made up of vocalic rather than consonantal intervals.

A. Effect of speaker on %V, VarcoV, and nPVI-V scores

[Figure 2](#) shows the mean %V, VarcoV, and nPVI-V scores of the six Standard Southern British English speakers averaged across sentences and measurers. The means for the other rhythm metrics can be found in [Appendix B](#).

A one-way repeated measures analysis of variance (ANOVA) for %V showed a main effect of Speaker, $F(5, 145) = 15.53$, $p < 0.001$. Values for ICCs and 95% con-

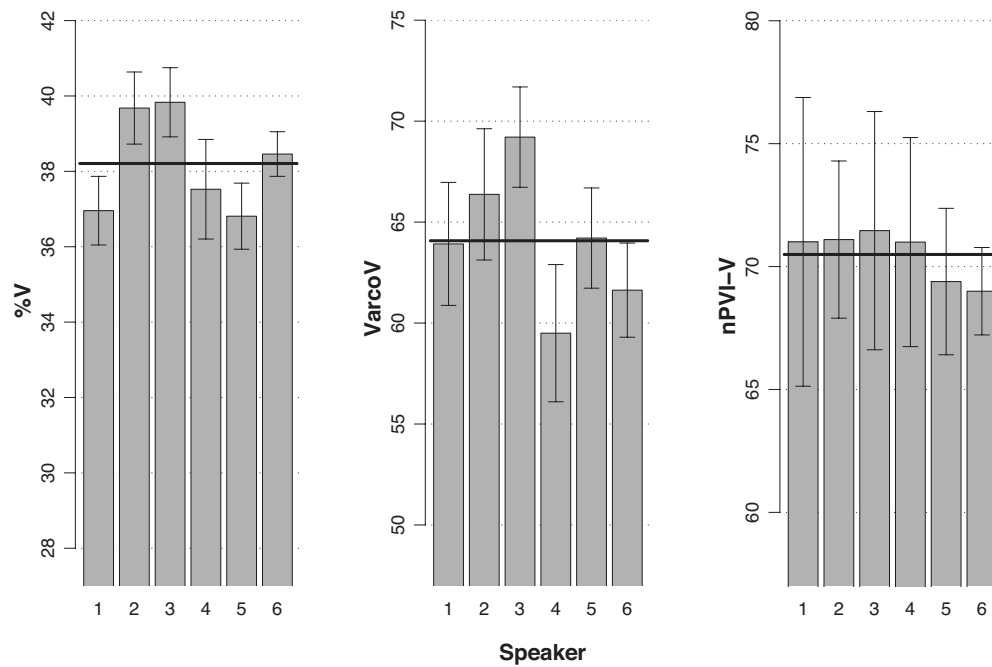


FIG. 2. Means and standard errors of the rhythm metrics %V, VarcoV, and nPVI-V for the six SSBE speakers. The solid horizontal bars indicate the pooled means.

confidence intervals for all metrics and conditions are summarized in Table II. As can be seen, consistency among speakers for %V was moderate. The largest difference in mean %V was 3.0 or 30% of the Spanish-English difference and the mean difference was 1.6 or 16% of Spanish-English. For VarcoV, there was also a main effect of speaker, $F(5, 145) = 13.47$, $p < 0.001$. Consistency among speakers was moderate (Table II). The largest difference in mean VarcoV was 9.7 or 42% of Spanish-English, and the mean difference was 4.2 or 18% of Spanish-English. For nPVI-V, we found no main effect of speaker, $F(5, 145) < 1$. Consistency among speakers was again moderate (Table II), but here the actual differences between the speakers' mean scores were very low: The largest difference in mean nPVI-V was 2.5 or only 7% of Spanish-English, and the mean difference was 1.2 or 3% of Spanish-English.

TABLE II. Intraclass correlation coefficients, among speakers, sentences, and measurers, for the rhythm metrics %V, VarcoV, and nPVI-V (with 95% confidence intervals). Values of 0.40–0.59 are commonly regarded as indicating moderate inter-rater reliability, values of 0.60–0.79 as substantial, and values of 0.80 or larger as outstanding (Garson, 2009).

	%V	VarcoV	nPVI-V
1. Speakers	0.53 [0.38,0.70]	0.56 [0.41,0.72]	0.52 [0.36,0.68]
2. Sentences	0.38 [0.23,0.55]	0.32 [0.17,0.49]	0.01 [-0.08,0.15]
3a. Measurers (including automatic alignment)	0.64 [0.49,0.77]	0.64 [0.49,0.77]	0.52 [0.36,0.68]
3b. Measurers (excluding automatic alignment)	0.67 [0.52,0.80]	0.71 [0.57,0.82]	0.62 [0.47,0.76]

All other rhythm metrics also showed a main effect of speaker (Appendix B). The ICCs of ΔV and nPVI-VC were in the range of those for %V, VarcoV, and nPVI-V; the ICCs for the consonantal interval metrics were lower.

B. Effect of sentence on %V, VarcoV, and nPVI-V scores

Figure 3 shows the mean %V, VarcoV, and nPVI-V scores of the five English sentences averaged across speakers and measurers. The means for the other rhythm metrics can be found in Appendix C.

For %V, a one-way repeated measures ANOVA showed a main effect of sentence, $F(4, 140) = 27.10$, $p < 0.001$. Consistency between sentences, as measured by the ICC, was low (Table II). The largest difference in mean %V between sentences was 4.9 or 49% of Spanish-English and the mean difference was 2.0 or 20% of Spanish-English. For VarcoV, there was also a main effect of sentence, $F(4, 140) = 35.38$, $p < 0.001$. Consistency between sentences was low (Table II). The largest difference in mean VarcoV was 13.0 or 57% of Spanish-English, and the mean difference was 6.5 or 28% of Spanish-English. For nPVI-V, we also found a main effect of sentence, $F(4, 140) = 35.49$, $p < 0.001$. Consistency between sentences was very low (Table II). The largest difference in mean nPVI-V was 15.5 or 42% of Spanish-English, and the mean difference was 9.0 or 25% of Spanish-English.

All other rhythm metrics also showed a main effect of Sentence (Appendix C). The ICCs likewise all indicated a low consistency between sentences. As with the analysis of speaker differences above, the ICCs of the consonantal interval metrics were generally lower than those of the vocalic metrics, with the exception of a very low value for nPVI-V.

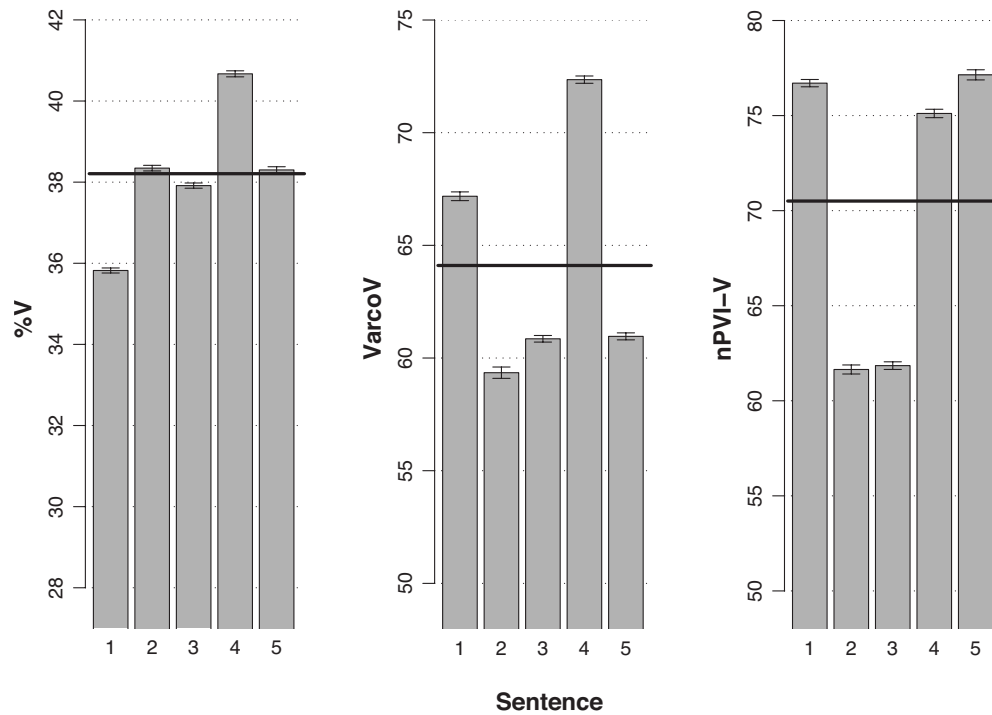


FIG. 3. Means and standard errors of the rhythm metrics %V, VarcoV, and nPVI-V for the five English sentences as read by the SSBE speakers. The solid horizontal bars indicate the pooled means.

C. Effect of measurer on %V, VarcoV, and nPVI-V scores

Figure 4 shows the mean %V, VarcoV, and nPVI-V scores of the five human measurers (M1–M5) and the automatic phone alignment (A) averaged across speakers and sentences. The means for the other rhythm metrics can be found in Appendix D.

For %V, a one-way repeated measures ANOVA showed a main effect of measurer, $F(5, 145)=15.85$, $p<0.001$. Consistency among measurers, as measured by the ICC, was,

however, substantial (Table II). The largest difference in mean %V was 3.7, or 37% of Spanish-English, and the mean difference was 1.4 or 14% of Spanish-English. For VarcoV, we also found a main effect of measurer, $F(5, 145)=10.06$, $p<0.001$. Consistency among measurers was substantial (Table II). The largest difference in mean VarcoV was 6.4 or 28% of Spanish-English, and the mean difference was 3.4 or 15% of Spanish-English. For nPVI-V, there was a main effect of measurer, $F(5, 145)=2.46$, $p<0.05$. Consistency among measurers was moderate (Table II). The largest dif-

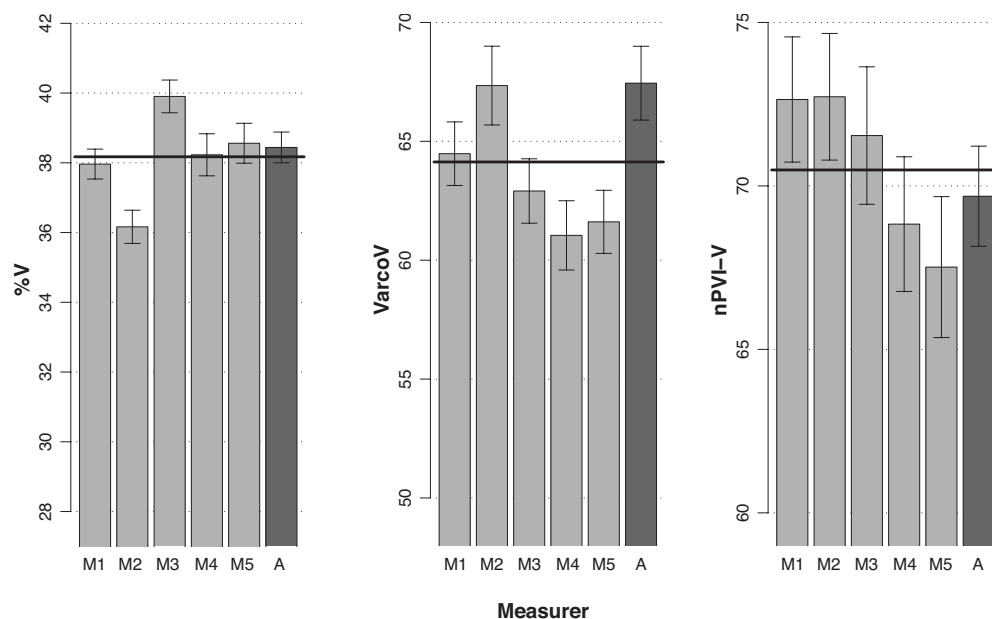


FIG. 4. Means and standard errors of the rhythm metrics %V, VarcoV, and nPVI-V (for SSBE) for the five human measurers (M1–M5) and the automatic phone alignment (A). The solid horizontal bars indicate the pooled means.

ference in mean nPVI-V was 5.2 or 14% of Spanish-English difference, and the mean difference was 2.6 or 7% of Spanish-English.

All other rhythm metrics also showed a main effect of measurer (Appendix D), with the exception of ΔV , where the effect approached significance. The consonantal interval metrics were again generally less consistent across measurers than the vocalic interval metrics.

As for the performance of the automatic phone alignment, Fig. 4 shows that the mean %V and nPVI-V scores were well within the range of the human measurers, and for VarcoV were just outside the human measurers' range. This observation was confirmed by pairwise comparisons between the automatic alignment and mean values for the five human measurers. There was no difference in mean for %V, $t(29) = 0.08$, $p = 0.44$, and nPVI-V, $t(29) = -0.51$, $p = 0.61$. For VarcoV, the mean score of the automatic alignment was higher than that of the human measurers, $t(29) = 3.02$, $p = 0.005$. When we compared the consistency among human measurers with that among all measurers (i.e., including the automatic alignment) we found that the ICCs among the human measurers alone are somewhat higher (Table II); however, the overlapping confidence intervals indicate that these differences were not significant.

D. Discussion

The analyses show that variation between speakers, sentences, and measurers all contribute to variability in rhythm metric scores. Sentences produced the largest variation in scores, with, for example, the mean difference in VarcoV between sentences being 28% of the Spanish-English difference. The differences due to speakers and measurers were generally smaller. In addition, consistency was highest among measurers (generally with an ICC of above 0.60) and lowest between sentences (ICC below 0.40). Between speakers, the ICC was around 0.50. The confidence intervals for sentence ICC and measurer ICC did not overlap for either %V or VarcoV, indicating that, for these two metrics, consistency among sentences was significantly lower than consistency among measurers. With regard to individual rhythm metrics, differences were generally smallest for nPVI-V and largest for VarcoV. Consistency, on the other hand, was slightly lower for nPVI-V than for the other two metrics: This likely to be in part the result of the lack of variation in nPVI-V between speakers; with little variation in the other two variables, there is little scope for sentences to show consistency in their patterns of scores.

How relevant are these differences to the comparison of results between studies? The largest difference between the English sentences examined here, for example, was more than half the size (57%) of the mean English-Spanish difference, which can be considered the archetypal poles of the contrastive rhythm space, at least among languages studied so far. Thus, a small unrepresentative set of sentences could easily alter the position of the language measured within the rhythm space, as also demonstrated by Prieto *et al.* (2009).

What then are "representative" sentences for the purpose of rhythm analysis? At the segmental level, Prieto *et al.*

TABLE III. The number of syllable transitions without a change in stress (weak-weak or strong-strong), the number of total transitions, and the CRI for the English sentences.

Sentence	Constant transitions	Total transitions	Constant regularity index
1	4	14	0.71
2	6	16	0.63
3	5	14	0.64
4	4	16	0.75
5	2	16	0.88

(2009) clearly showed that the phonological structure of syllables has an impact on rhythm scores, with untypical patterns (such as sentences comprised of a succession of open syllables for English) producing misleading results. At the suprasegmental level, one critical factor is clearly the relative numbers and distribution of strong and weak syllables. Low *et al.* (2000) constructed sentences containing only strong syllables and showed that nPVI-V scores were much lower than those for sentences containing an alternating pattern of strong and weak syllables.

As a further test of the effect of suprasegmental structure on rhythm measures, we introduce the contrast regularity index (CRI), a measure of the regularity of sentence stress patterns. We calculated CRI as $1 - (c/t)$, where c stands for the number of syllable transitions in which the stress remains constant (a weak syllable followed by a weak syllable or a strong syllable followed by another strong syllable), and t for the total number of syllable transitions. The lower the number of syllable-to-syllable transitions where stress is unaltered, the higher the CRI, with a CRI of 1 indicating a strict alternation of strong and weak syllables.

The CRIs for the five English sentences used in this study are listed in Table III. Given that the PVI was devised to reflect the strength of the alternation between successive syllables, it is reassuring that patterns of CRI values quite closely correspond to the nPVI-V values, with the two sentences lowest in CRI (sentences 2 and 3) also having the lowest nPVI-V scores, and the sentences with a high CRI (sentences 1, 4, and 5) having the highest nPVI-V scores.

It is worth noting that, although VarcoV scores were generally highly correlated with nPVI scores, the pattern of scores was very different for sentence 5, which had a low VarcoV score and yet the highest nPVI-V score. This was also the sentence that had the highest CRI value, reflecting its regularly alternating strong-weak pattern, only disrupted by two weak-weak transitions. The ordering of strong and weak syllables is thus critical for nPVI-V, and irrelevant for simple global measures of interval duration variation such as VarcoV, as discussed in Low *et al.*, 2000. This sensitivity to the rhythmic structure of utterances could be seen as a strength of the PVI, but could also be a weakness in studies that lack a sufficiently large and language-typical range of metrical structure.

In summary, it is clear that all three factors—speakers, sentences, and measurers—need to be considered when developing studies and comparing results. We found that the largest source of variation in rhythm scores was the sentences selected for the study: These can either be randomly

sampled or specifically designed to be representative of the language in terms of phonological and suprasegmental structures, as discussed above. Speakers—a moderate source of variation in rhythm scores—can be naturally treated as a random factor and so the usual caveat regarding low speaker numbers applies. There is little corrective action available, however, regarding idiosyncratic differences between measurers apart from the obvious step of ensuring that all measurers conform to an agreed protocol for the identification of segment boundaries. Given that, it is reassuring to note that differences in rhythm scores between measurers are relatively small compared to those arising from speakers and, in particular, sentences, and that consistency between measurers—as indexed by intraclass correlation coefficients—is relatively high.

IV. GENERAL DISCUSSION

In an attempt to assess the reliability of contrastive rhythm metrics used routinely in phonetic research, we examined how such metrics are affected by variation between measurers, sentences, and speakers. We analyzed five English sentences spoken by six speakers of Standard Southern British English, with segmentation into vocalic and intervocalic intervals carried out by five phoneticians, together with an automated alignment using speech recognition software.

The results clearly demonstrate that the nature of the sentence materials recorded is the most important source of variation in rhythm scores. Consistency of scores between sentences was low for all metrics, even those metrics—%V, VarcoV, and nPVI-V—previously found to be most discriminative between language groups and robust to variation in articulation rate (White and Mattys, 2007a, 2007b). Mean differences between SSBE sentences for these three metrics were equivalent to 20%–28% of the differences between Castilian Spanish and Standard Southern British English, which were the largest cross-linguistic differences found in our previous study (White and Mattys, 2007a), with %V the most consistent and nPVI-V the least consistent between sentences. Such differences raise potentially serious problems of interpretation of rhythm scores given the underlying assumption that they should reflect stable rhythmical properties of languages. One solution to this problem is to record a sufficiently large number of randomly sampled sentences to provide a representative selection. Alternatively, materials may be constructed to accurately represent the phonological and metrical structures encountered in the natural speech of the languages under investigation. We have proposed a contrast regularity index as a simple statistical measure of metrical structure; other measures would be required to assess phonological patterns such as the clustering of consonants in onsets and codas. Analyzing recordings of natural spontaneous speech represents a third approach to the problem.

Variation in rhythm scores between speakers was not as marked as between sentences, with mean differences for the three metrics less than 18% of the Spanish-English difference, and particularly low for nPVI-V. Consistency between speakers was nevertheless only moderate and provides a strong indication against taking single-speaker studies to be

in any way representative of population means. Of course, rhythm metrics may be efficacious for comparing scores within a single speaker under different speaking conditions or assessing changes over time, for example, regarding the deterioration in a speaker's prosody due to speech pathology (e.g., Liss *et al.*, 2009), the efficacy of ameliorative speech therapy, or progress to proficiency in second language acquisition. Assuming that the caveats regarding materials are heeded, the use of rhythm metrics in longitudinal studies such as these is one of the potentially most productive areas of their application.

Inter-measurer reliability was relatively good, with mean differences for %V, VarcoV, and nPVI-V between 7% and 15% of the Spanish-English difference. As with speakers, nPVI-V showed the smallest mean difference between measurers, although intraclass correlation coefficients showed that consistency was numerically somewhat lower than for %V and VarcoV.

A practical implication of these findings is that it should be possible to use several measurers within one study without substantial loss of reliability. Our measurers provided segmentations based on their own interpretation of a set of written instructions without consulting each other. By ensuring that measurers discuss problem cases and update their joint segmentation protocol accordingly, we expect that differences between measurers could be reducible to a non-significant level. In addition, of particular practical interest is our finding that the automatic phone alignment produced rhythm metric scores that were comparable to those of the human measurers. Of the three metrics examined here, only for VarcoV was the score of the automatic alignment different from the mean score of the human measurers, although the automatic alignment still performed within the range of the human measurers. Automatic phone alignment could, therefore, be used in future studies of contrastive speech rhythm, assuming an adequate supply of training data for the languages under investigation.

Finally, as this paper is essentially a methodological study, we present some recommendations for researchers intending to utilize contrastive rhythm metrics. The first two recapitulate what we have learnt from our previous work (White and Mattys, 2007a, 2007b); the others summarize the findings of the present study.

- (1) %V, VarcoV, and nPVI-V are robust to variation in articulation rate and are effective at discriminating between language varieties previously held to differ in terms of contrastive rhythm. However, as all rhythm metrics have limitations, it is safest to use %V in combination with either VarcoV or nPVI-V rather than rely on a single metric.
- (2) Results for non-rate-normalized metrics (ΔV , ΔC , and rPVI-C) are difficult to interpret and not reliably discriminative, likewise all metrics of consonantal interval variation (VarcoC, in addition to the non-rate-normalized ΔC and rPVI-C). Furthermore, they show relatively poor consistency between speakers, sentences, and measurers.
- (3) Single-speaker or low-N studies should absolutely be avoided where speakers are intended to be representative

of a particular linguistic group. Contrastive rhythm metrics may be useful in single-speaker longitudinal studies, however.

- (4) Rhythm scores are strongly affected by the particular linguistic materials used. Either a large sample of sentences should be used or materials should be constructed to be representative of the relevant phonological and metrical properties of the language under study.
- (5) Where several measurers are used, they should work according to an agreed protocol for the identification of segment boundaries. Furthermore, discussion and comparison of difficult cases between measurers, which were avoided in the current study, should help minimize variation in rhythm scores.
- (6) Contrastive rhythm scores obtained through automatic alignment show good agreement with those obtained from human measurers, assuming sufficient training data

are available for the language in question. Of course, use of automated methods would allow much larger sampling of speakers and sentences.

- (7) Do not rely too heavily on contrastive rhythm metrics or over-interpret the results of studies that use them. They merely provide an approximate indication of the degree of temporal stress contrast in a language and are susceptible to extraneous variation from multiple sources.

ACKNOWLEDGMENTS

This research was supported by a grant from the Leverhulme Trust (United Kingdom) to S. L. Mattys (Grant No. F/00182/BG) and a Research Training Network grant from the Marie Curie Foundation (Grant No. MRTN-CT-2006-035561). We thank Esther de Leeuw for help with segmentation.

APPENDIX A: THE SENTENCES READ BY THE SSBE SPEAKERS IN THIS STUDY

The following are the sentences read by the SSBE speakers.

- (1) The supermarket chain shut down because of poor management.
- (2) Much more money must be donated to make this department succeed.
- (3) In this famous coffee shop they serve the best doughnuts in town.
- (4) The chairman decided to pave over the shopping center garden.
- (5) The standards committee met this afternoon in an open meeting.

APPENDIX B: SPEAKER COMPARISONS

Means (and standard errors) of the rhythm metrics ΔV , ΔC , VarcoC, rPVI-C, and nPVI-VC according to speaker; the ICC among speakers (with 95% confidence intervals); and the *F*-test for a main effect of speaker.

	Sp 1	Sp 2	Sp 3	Sp 4	Sp 5	Sp 6	ICC	<i>F</i> -test
ΔV	53.8 (1.4)	55.7 (1.3)	55.5 (0.6)	45.0 (1.3)	42.9 (0.6)	46.4 (1.1)	0.45 [0.29,0.62]	$F(5, 145)=48.06,$ $p < 0.001$
ΔC	69.3 (2.5)	55.1 (2.0)	58.8 (2.4)	60.4 (2.1)	54.0 (2.2)	57.7 (1.4)	0.26 [0.12,0.45]	$F(5, 145)=8.94,$ $p < 0.001$
VarcoC	48.2 (1.6)	42.8 (1.2)	47.8 (1.4)	47.6 (1.1)	46.4 (1.2)	48.2 (1.5)	0.25 [0.12,0.44]	$F(5, 145)=3.21,$ $p=0.009$
rPVI-C	82.9 (3.8)	66.5 (2.5)	65.7 (3.0)	72.6 (2.8)	66.4 (2.7)	67.9 (2.5)	0.37 [0.22,0.56]	$F(5, 145)=8.32,$ $p < 0.001$
nPVI-VC	40.9 (1.3)	37.2 (1.3)	41.5 (1.7)	40.2 (1.1)	40.9 (1.4)	34.2 (0.8)	0.61 [0.46,0.75]	$F(5, 145)=12.32,$ $p < 0.001$

APPENDIX C: SENTENCE COMPARISONS

Means (and standard errors) of the rhythm metrics ΔV , ΔC , VarcoC, rPVI-C, and nPVI-VC according to sentence read; the ICC among sentences (with 95% confidence intervals); and the *F*-test for a main effect of sentence.

	Sen 1	Sen 2	Sen 3	Sen 4	Sen 5	ICC	<i>F</i> -test
ΔV	51.9 (1.3)	48.3 (1.4)	49.8 (1.0)	55.3 (1.3)	44.1 (1.0)	0.49 [0.33,0.65]	$F(4, 140)=23.50,$ $p < 0.001$

ΔC	59.3 (0.9)	65.8 (2.3)	59.4 (1.6)	55.8 (2.7)	55.8 (2.1)	0.22 [0.09,0.40]	$F(4, 140)=5.26,$ $p < 0.001$
VarcoC	43.0 (0.7)	49.9 (1.1)	44.2 (1.0)	49.4 (1.6)	47.6 (1.3)	0.05 [-0.05,0.20]	$F(4, 140)=7.07,$ $p < 0.001$
rPVI-C	63.3 (1.5)	74.6 (3.0)	84.1 (1.8)	65.5 (3.4)	64.1 (2.2)	0.25 [0.11,0.43]	$F(4, 140)=16.72,$ $p < 0.001$
nPVI-VC	41.9 (0.9)	33.9 (0.7)	35.2 (0.8)	36.9 (0.9)	47.8 (1.3)	0.16 [0.04,0.33]	$F(4, 140)=44.48,$ $p < 0.001$

APPENDIX D: MEASURER COMPARISONS

Means (and standard errors) of the rhythm metrics ΔV , ΔC , VarcoC, rPVI-C, and nPVI-VC according to measurer (human measurers M1–M5 and automatic phone alignment A); the ICC among measurers (with 95% confidence intervals); and the F -test for a main effect of measurer.

	M1	M2	M3	M4	M5	A	ICC	F -test
ΔV	49.2 (1.4)	49.5 (1.4)	50.2 (1.4)	50.1 (1.4)	48.8 (1.5)	51.5 (1.7)	0.79 [0.69,0.88]	$F(5, 145)=2.02,$ $p=0.080$
ΔC	58.6 (1.9)	59.1 (1.9)	57.0 (1.9)	68.1 (3.1)	58.6 (2.0)	53.9 (1.9)	0.42 [0.27,0.60]	$F(5, 145)=8.20,$ $p < 0.001$
VarcoC	46.8 (1.2)	45.4 (1.2)	47.4 (1.2)	50.8 (1.8)	46.4 (1.2)	44.1 (1.3)	0.43 [0.27,0.61]	$F(5, 145)=4.92,$ $p < 0.001$
rPVI-C	69.9 (2.6)	69.1 (2.5)	68.0 (2.8)	83.1 (4.0)	68.5 (2.8)	63.4 (2.7)	0.54 [0.39,0.70]	$F(5, 145)=11.23,$ $p < 0.001$
nPVI-VC	39.7 (1.5)	38.3 (1.5)	39.6 (1.3)	40.8 (1.5)	39.3 (1.4)	36.9 (0.9)	0.68 [0.55,0.81]	$F(5, 145)=3.07,$ $p=0.011$

Baayen, R. H., Piepenbrock, R., and Gulikers, L. (1995). The CELEX lexical database. Release 2 (CD-ROM) (Linguistic Data Consortium, University of Pennsylvania, PA).

Carter, P. M. (2005). "Quantifying rhythmic differences between Spanish, English, and Hispanic English," in *Theoretical and Experimental Approaches to Romance Linguistics: Selected Papers From the 34th Linguistic Symposium on Romance Languages*, Current Issues in Linguistic Theory Vol. 272, edited by R. S. Gess and E. J. Rubin (John Benjamins, Amsterdam), pp. 63–75.

Dauer, R. M. (1983). "Stress-timing and syllable-timing reanalyzed," *J. Phonetics* 11, 51–62.

Dellwo, V. (2006). "Rhythm and speech rate: A variation coefficient for deltaC," in *Language and Language Processing: Proceedings of the 38th Linguistic Colloquium*, Peter Lang, Frankfurt, pp. 231–241.

Demuyne, K., Roelens, J., Van Compernelle, D., and Wambacq, P. (2008). "Sprak: An open source speech recognition and automatic annotation kit," in *Interspeech 2008*, Brisbane (CDROM).

Draxler, C., van den Heuvel, H., and Tropic, H. (1998). "SpeechDat experiences in creating large multilingual speech databases for teleservices," in *Proceedings of the LREC98*, Granada, pp. 361–366.

Gamer, M., Lemon, J., and Fellows, I. (2007). "irr: Various coefficients of interrater reliability and agreement," R package version 0.70.

Garson, G. D. (2009). "Reliability analysis," <http://faculty.chass.ncsu.edu/garson/PA765/reliab.htm> (Last viewed 7/12/2009).

Grabe, E., and Low, E. L. (2002). "Durational variability in speech and the rhythm class hypothesis," in *Laboratory Phonology 7*, edited by C. Gussenhoven and N. Warner (Mouton de Gruyter, Berlin), pp. 515–546.

Lehiste, I. (1977). "Isochrony reconsidered," *J. Phonetics* 5, 253–256.

Liss, J. M., White, L., Mattys, S. L., Lansford, K., Lotto, A. J., Spitzer, S. M., and Caviness, J. N. (2009). "Quantifying speech rhythm abnormalities in the dysarthrias," *J. Speech Lang. Hear. Res.* 52, 1334–1352.

Low, E. L., Grabe, E., and Nolan, F. (2000). "Quantitative characterisations of speech rhythm: 'Syllable-timing' in Singapore English," *Lang Speech* 43, 377–401.

Patel, A. D., Iversen, J. R., and Rosenberg, J. C. (2006). "Comparing the rhythm and melody of speech and music: The case of British English and French," *J. Acoust. Soc. Am.* 119, 3034–3047.

Peterson, G. E., and Lehiste, I. (1960). "Duration of syllable nuclei in English," *J. Acoust. Soc. Am.* 32, 693–703.

Prieto, P., Vanrell, M. M., Astruc, L., Payne, E., and Post, B. (2009). "Top-down planning of rhythm: Evidence from Catalan, English, and Spanish," paper presented at *Phonetics and Phonology in Iberia 2009*, Las Palmas de Gran Canaria, June.

Ramus, F. (2002). "Acoustic correlates of linguistic rhythm: Perspectives," in *Proceedings of the Speech Prosody 2002*, Aix-en-Provence, pp. 115–120.

Ramus, F., Nespors, M., and Mehler, J. (1999). "Correlates of linguistic rhythm in the speech signal," *Cognition* 73, 265–292.

Shrout, P. E., and Fleiss, J. L. (1979). "Intraclass correlations: Uses in assessing rater reliability," *Psychol. Bull.* 86, 420–428.

Turk, A., Nakai, S., and Sugahara, M. (2006). "Acoustic segment durations in prosodic research: A practical guide," in *Methods in Empirical Prosody Research*, edited by S. Sudhoff, D. Lenertova, R. Meyer, S. Pappert, P. Augurzy, I. Mleinek, N. Richter, and J. Schliesser (de Gruyter, Berlin), pp. 1–28.

- White, L., and Mattys, S. L. (2007a). "Calibrating rhythm: First language and second language studies," *J. Phonetics* 35, 501–522.
- White, L., and Mattys, S. L. (2007b). "Rhythmic typology and variation in first and second languages," in *Segmental and Prosodic Issues in Romance Phonology*, edited by P. Prieto, J. Mascaró, and M.-J. Solé (John Benjamins, Amsterdam), pp. 237–257.
- White, L., Payne, E., and Mattys, S. L. (2009). "Rhythmic and prosodic contrast in Venetan and Sicilian Italian," in *Phonetics and Phonology: Interactions and Interrelations*, edited by M. Vigario, S. Frota, and M. J. Freitas (John Benjamins, Amsterdam), pp. 137–158.