WORKING PAPERS

IN

LINGUISTICS

The notes and articles in this series are progress reports on work being carried on by students and faculty in the Department. Because these papers are not finished products, readers are asked not to cite from them without noting their preliminary nature. The authors welcome any comments and suggestions that readers might offer.

Volume 40(9)

2009
(December)

DEPARTMENT OF LINGUISTICS
UNIVERSITY OF HAWAIʻI AT MĀNOA
HONOLULU 96822

DEPARTMENT OF LINGUISTICS FACULTY

2009

Victoria B. Anderson
Byron W. Bender (Emeritus)
Benjamin Bergen
Derek Bickerton (Emeritus)
Robert A. Blust
Robert L. Cheng (Adjunct)
Kenneth W. Cook (Adjunct)
Kamil Deen
Patricia J. Donegan (Co-Graduate Chair)
Katie K. Drager
Emanuel J. Drechsel (Adjunct)
Michael L. Forman (Emeritus)
George W. Grace (Emeritus)
John H. Haig (Adjunct)
Roderick A. Jacobs (Emeritus)
Paul Lassettre
P. Gregory Lee
Patricia A. Lee
Howard P. McKaughan (Emeritus)
William O'Grady (Chair)
Yuko Otsuka
Ann Marie Peters (Emeritus, Co-Graduate Chair)
Kenneth L. Rehg
Lawrence A. Reid (Emeritus)
Amy J. Schafer
Albert J. Schütz, (Emeritus, Editor)
Ho Min Sohn (Adjunct)
Nicholas Thieberger
Laurence C. Thompson (Emeritus)

ISSUES IN THE QUANTITATIVE APPROACH TO SPEECH RHYTHM COMPARISONS[1]

DIANA STOJANOVIC

This paper explores issues related to the quantitative approach to characterizing linguistic rhythm. In particular, it highlights challenges faced by the method in which rhythmic similarity is evaluated by use of rhythm metrics. Predictions made in this paper for the success of such metrics in classifying languages agree with the results in the literature. Explanations are proposed for the cases where discrepancies occur between the results in the literature and the predictions made based on the rhythm-class hypothesis. Criticisms of the current approach lead to a proposed model of perception of speech rhythm and several methods by which perceived rhythmic differences could be quantified more successfully.

1. RHYTHM TYPOLOGY AND THE RHYTHM-CLASS HYPOTHESIS (RCH). In one of the initial formulations of the RCH, Abercrombie (1967) claimed that two posited rhythmic classes differ in the type of isochrony the languages exhibit: isochrony of syllables for syllable-timed languages and isochrony of inter-stress intervals for stress-timed languages. This grouping was based on the perception of salient rhythmic differences between languages such as English or Dutch, called *stressed-timed*, and languages such as Spanish or French, called *syllable-timed*. Two basic proposals were made (Abercrombie 1967): (1) that in the stressed-timed languages stressed syllables occur regularly, while the syllable-timed languages syllables occur regularly, and (2) that languages of the world belong to one or the other class. A third class, *mora-timed*, was later added to accommodate languages such as Japanese that were believed to differ from both of the existing types in that moras occur regularly. Languages included in this group usually have phonological length distinctions.

Because attempts to find evidence of isochrony in the characteristic unit in the speech signal failed (Dauer 1983, Roach 1982), it was proposed that isochrony is a purely perceptual phenomenon (Lehiste 1977) that could not be measured from the acoustic signal. An alternative view of rhythm was put forward by Dauer (1983), who noticed phonological similarities among the original members of the stress-timed group and the syllable-timed group respectively: languages in the stress-timed group have vowel reduction in unstressed syllables and phonotactics that allows complex syllable structure; syllable-timed languages lack vowel reduction and have simple (C)V(C) syllable structure.[2]

Dauer's 1983 model of rhythm posits that rhythm emerges from phonological properties such as syllable structure and vowel reduction, as well as duration of stressed syllables, phonemic vowel length distinction, the effect of intonation on stress, the effect of tone on stress, consonantal phonetic inventory, and function of stress.[3]

The more of the properties a language has, the more stress-timed it is proposed to be. Languages thus are said to lie on a continuum between prototypically syllable-timed at one end (Japanese) and prototypically stress-timed (English) on the other end of continuum. This means that various properties combine, possibly with various levels of importance, towards one resultant perception variable: rhythm. Because the listed properties do not always co-occur, in this view two languages can have different

[1] I would like to thank Profs. Victoria Anderson, Patricia Donegan, Ann Peters, and Albert J. Schütz for their help with this paper. All remaining errors are mine.

[2] Mora-timed languages were not discussed in Dauer 1983; based on the properties of Japanese, Yoruba, and Telugu, the mora-timed group has a syllable structure even simpler than that of the syllable-timed group, namely, mostly (C)V. The (C)VC_1 exists but allows only few selective consonants in C_1 position.

[3] Donegan (personal communication) suggests that the following are important: tendency to diphthongize, vowel harmony, geminate consonants, vowel-length distinctions, many vs. few vowel quality distinctions, contour vs. level tone.

properties but be equally syllable- or stress-timed. In this model, called "continuous uni-dimensional model of rhythm" (Ramus 2002), a strict rhythm-class hypothesis is not true. Instead, languages form a rhythm continuum.

Both infants (Nazzi et al. 1998, Nazzi and Ramus 2003) and adults (Ramus et al. 2003) discriminate among languages based on rhythmic properties, which supports the view that languages can be grouped into different types. Because of that, the idea of rhythmic classes has persisted despite the failure to find measurable evidence of isochrony in the speech signal. Renewed interest in finding evidence for rhythmic differences has occurred with a shift in focus: the distinction of rhythmic classes is not based on isochrony, or lack thereof, among successive units, but is based on a somewhat more relaxed criterion: degree of durational variability among such units. Another difference introduced with the new approach involved a change of unit whose variability is used to characterize rhythm. Instead of syllables and feet (or intervals between two stresses), non-phonological units such as vocalic and consonantal intervals[4] were used. The change of unit was motivated by the results of the studies on infant perception of rhythm. Namely, Nazzi et al. (1998) found that infants, like adults, perceive rhythmic[5] differences between languages. It was shown that infants are able to distinguish speech samples of English from those of French, for instance, but not English from Dutch. Ramus et al. (1999:270) assume that "the infant primarily perceives speech as a succession of vowels of variable durations and intensities, alternating with periods of unanalyzed noise (i.e. consonants)" and suggest that perceiving rhythmic differences must not be based on phonological units such as syllables and feet.

The focus of the new approach was on the formulation of a two-dimensional space in which good exemplars of stress-timed and syllable-timed languages would be separated. Such spaces were defined most of the time by measures that in some way mirror distinguishing phonological properties. Various measures were introduced in the hope of capturing the crucial differences between posited rhythm classes.

The early results were encouraging in that prototypically stress-timed and syllable-timed languages were mapped into opposite corners of the space (Ramus et al. 1999, Grabe and Low 2002). In subsequent studies, in which larger numbers of speakers per language and new languages were tested with various speech materials and speech styles, it was found that (1) empirical results show more support for Dauer's continuum hypothesis than for the strict rhythm class hypothesis (Grabe 2002), and (2) various factors compromise successful cross-linguistic classification. Several serious problems of the quantitative approach based on rhythm measures include the following: (1) within-language inter-speaker differences may be larger than between-class differences (Benton et al. 2007); (2) speech rate (Dellwo and Wagner 2003) and speech style (Benton et al. 2007) may affect metric values more than the posited rhythm class; (3) different metrics produce contradictory classifications: for instance, ΔV and ΔC classify Polish differently (Ramus et al. 1999); (4) different studies obtain contradictory results based on the same rhythm metric (Dellwo 2006 and White and Mattys 2007); and (5) rhythm metrics depend on the segmentation rules (Stojanovic 2008).

Recent proposals include another view of stress- vs. syllable- timing. Nolan and Asu (2009), for instance, propose that stress-timing and syllable-timing are independent dimensions exemplified by all languages, and so one language can express a certain level of stress-timing and a certain other level of syllable-timing.

There is no consensus in the current literature on whether rhythm can be measured from the acoustic signal, or whether it is different from timing (Arvaniti 2009). In fact, some (Pamies Bertrán 1999) propose that speech is not rhythmic at all. Others view rhythm as coupling between nested prosodic units (Cummins 2002).

The goal of this paper is to discuss challenges for the quantitative approach to rhythm classification based on comparison of durational variability through rhythm metrics. Some inconsistencies in the results obtained in the literature are explained, and modifications are proposed towards better capturing certain

[4] In the literature, these are called vocalic and intervocalic intervals.

[5] It was posited that the differences are solely rhythm-based because the samples were filtered to eliminate segmental information

aspects of perceived rhythm in speech. At the end, a model of rhythm is proposed based on the idea that durational differences are perceived by the listeners in conjunction with other properties of the signal.

2. DURATIONAL VARIABILITY. Following the view that rhythm reflects durational patterns in speech, I consider factors that affect duration of vowels, consonants, as well as vocalic and intervocalic intervals. In addition, I discuss the effect of speech rate on change of relative durations of segments within a phrase and highlight the difference between absolute duration measured from the signal and perceived duration experienced by the listeners.

2.1 FACTORS THAT AFFECT DURATION. An excellent review of the literature on various factors that affect segmental durations in English is given in Klatt 1976. In addition to listing durational factors and providing references of experimental studies that support them, Klatt also relates these factors to perception studies of duration and to listeners' ability to use durational differences to help make linguistic decisions. Here, I adapt the factors listed by Klatt to make comparable rules for vowels and consonants. Thus, some of the rules (rule 4 for vowels and rule 5 for consonants) need to be experimentally verified.

2.1.1 VOWELS. All else being equal, vowel V1 is longer than vowel V2 if: (1) V1 is inherently longer than V2 (for instance, /ɒ/ in /dɒl/ 'doll' is longer than /i/ in /dil/ 'deal'), (2) V1 is phonemically long and V2 is phonemically short (for instance, in Hawaiian, /a:/ in *Mānoa* /ma:noa/ is longer than /a/ in *manu* /manu/), (3) V1 is a diphthong and V2 is a monophthong (in English, /ai/ in *my* is longer than /i/ in *me*), (4) V1 is a single vowel and V2 is a part of a hiatus (/i/ in /hi nouz/ 'he knows' is longer than /i/ in /hi ouz/ 'he owes'), (5) V1 is in a word with fewer following syllables (in English, /ʌ/ in *fun* /fʌn/ is longer than /ʌ/ in *funny* /fʌni/, and the V1 of *funny* is longer than that of *funnily* /fʌnili/), (6) V1 is in a phrase-final syllable and V2 is not (/ʌ/ in *sounds like fun* /saunds laik ˈfʌn/ is longer than /ʌ/ in *sounds like a fun movie* /saunds laik ə ˈfʌn muvi/), (7) V1 is in a stressed and V2 in an unstressed syllable (first /i/ in /mini/ 'meany' is longer than /i/ in /fʌni/ 'funny', (8) V1 is in a word with sentence prominence and V2 is not (/ʊ/ in *It's my BOOK, not album* is longer than /ʊ/ in *It's in MY book, not hers*), (9) there is a language-specific rule that makes V1 longer (in English, /ɛ/ in /sɛd/ 'said' is longer than /ɛ/ in /sɛt/ 'set'), and (10) V1 is produced at a slower tempo (speech rate) than V2.

2.1.2 CONSONANTS. All else being equal, consonant C1 is longer than consonant C2 if: (1) C1 is inherently longer than C2 (in English, /m/ is longer than /n/ (Umeda 1977), (2) C1 is phonemically long and C2 is phonemically short (for instance, in Italian, /kk/ in *ecco* /ekko/ 'here it is' is longer than /k / in *eco* /eko/ 'echo'), (3) C1 is a complex consonant, and C2 is a simple consonant (in English, / tʃ / in /tʃɪp/ 'chip' is longer than / ʃ / in /ʃɪp/ 'ship'), (4) C1 is a single consonant and C2 is a part of a cluster (/n/ in /ben/ 'Ben' is longer than /n/ in /bent/ 'bent'), (5) C1 is in a word with fewer syllables (/f/ in *fun* /fʌn/ is longer than /f/ in *funny* /fʌni/), (6) C1 is in a phrase-final syllable and C2 is not (/n/ in *looks like fun* /lʊks laik ˈfʌn/ is longer than /n/ in *looks like a fun movie* / lʊks laik ə ˈfʌn muvi/), (7) C1 is in a stressed and C2 in unstressed syllable (first /m/ in /mimi/ 'Mimi' is longer than the second /m/ in /mimi/ 'Mimi', (8) C1 is in a word with sentence prominence and C2 is not (/b/ in *It's my BOOK, not album* is longer than /b/ in *It's in MY book, not hers*), (9) there is a language-specific rule that makes C1 longer (in English, /t/ in /sɛt/ 'set' is longer than /d/ in /sɛd/ 'said'), and (10) C1 is produced at a slower tempo (speech rate) than C2.

The factors causing the effects listed in (1–10) are respectively: (1) intrinsic duration, (2) phonemic length, (3) complex segment quality, (4) resource-sharing, (5) word length, (6) prosodic phrasing, (7) lexical stress, (8) prosodic prominence effect, (9) language-specific rule, and (10) speech rate. Some of these 10 factors are universal (1, 6, 10), while others are language-specific and either do not affect duration (5, 7, 9) or are not applicable (2, 3, 4, 8) in other languages. We can also group the factors based on their nature into *structural*, *prosodic*, and *pragmatic*. Structural factors include intrinsic duration, phonemic length, complex quality, and language-specific rules (1, 2, 3, 4, 5, 9), prosodic factors include word stress, sentence focus, phrasal edge-lengthening (6, 7, 8), and pragmatic factors include speech rate (10).

Next, I consider intervals that consist of more than one vocalic or consonantal phone. In addition to factors (1–10), which affect individual phones, durational variability of an interval will be higher if its

size, measured in number of phones, is larger. Thus /stɹ/ in /stɹɒŋ/ 'strong' is longer than /ɹ/ in /ɹɒŋ/ 'wrong' and /i i/ in /hi ˈits/ 'he eats' is longer than /i/ in /hi ˈsɪts/ 'he sits'. Typological effects based on syllable structure are briefly discussed.

2.1.3 VOCALIC INTERVALS. Vocalic intervals consist of more than one vowel[6] only when a (C)V syllable is followed by a V(C) syllable, as in *he is* /hi ˈɪz/ or *naïve* /na ˈiv/, i.e., where hiatus occurs. Only languages that allow both (C)V (no coda) and V(C) (no onset) syllable types will have hiatuses. In Levelt and van de Vijver 2004, twelve syllable-type inventories are proposed, out of which seven types may have hiatuses. However, some languages have methods to avoid hiatuses through elision (deleting one of the vowels) or syneloepha (merging of consecutive vowels), or consonant insertion.

Ultimately, the average number of vowels in a vocalic interval depends on the distribution of hiatuses in a given sample. No-coda languages that allow a simple *V* syllable type, such as Cayuvava, Mazateko (listed in Levelt and van de Vijver 2004), and Hawaiian, are more likely to have higher variability of vocalic intervals because the maxium size of the vocalic interval and frequency of hiatuses are higher than in other languages. Languages that have both (C)V and V(C) syllable types but also have one or more closed-syllable types, such as Spanish, Finnish, or English, are likely to have hiatuses with much smaller frequency and size. Vocalic interval size is not a significant cause of durational variability in these languages. Finally, in languages in which the obligatory onset principle applies, durational variability will depend only on the factors affecting single vowels.

2.1.4 INTERVOCALIC INTERVALS. The number of consonants in an intervocalic interval is a function of syllable structure, i.e., the occurrence of consonant clusters in onsets or codas, and combinations that occur across word-boundaries (*he steals* /hi stilz/ as well as *his team* /hɪz tim/). The first factor is determined by the syllable types in a given language, while the second also depends on the way syllables combine to form words and phrases.

In Levelt and van de Vijver 2004, only two out of twelve types will not allow consonant clusters, that is, language like Hua and Cayuvava, which allow only open syllables with no complex onsets. Hawaiian also falls in this group. Other types will show the effect of consonant cluster size on the durational variability of intervocalic intervals. The rule of thumb is that the average size of the consonant cluster will be higher in a language with a higher number of closed-syllable types and the prediction is that, for example, English and Dutch will have higher durational variability of intervocalic intervals than Spanish.

2.2 SPEECH RATE. While it is intuitively clear that a word produced at a higher speech rate is shorter, and that all vowels and consonants have smaller durations, it is not obvious whether each phone will shorten equally, i.e., proportionally as speech rate increases.

Studies on short and long vowels (e.g., Hirata 2004) repeatedly find different amounts of shortening for long and short vowels: the ratio of long to short reduces as the speech rate increases. Other studies show that vowels of different qualities (Flege 1988) and different phonemic lengths (Hirata 2004) change their relative duration, i.e., some shorten or lengthen more than others. If a ratio of two segments S1 and S2 reduces as the speech rate increases, we say that segment S1 is more elastic with respect to speech rate. Due to different elasticities, the relative proportion of each segment in a phrase will be different for different speech rates. This fact will be revisited in section 3, where the effect of speech rate on the value of different measures is discussed.

2.3 FREQUENCY FACTORS. How much variability is present in a speech sample depends on how many factors combine and with what frequencies. For example, all else being equal, a language with a phonemic vowel-length distinction will likely have more variation in vowel duration than a language without it, and a language that allows consonant clusters will have higher variation of intervocalic intervals than a language that does not. Also, a language with higher average cluster size (due to the higher number of consonants allowed in a cluster) will have higher consonant interval variability than a language with smaller average cluster size, and a language with high frequency of clusters will have higher consonant interval variability than a language where clusters are rare. The frequency factor applies

[6] More precisely, vocalic interval consists of one or more syllable nuclei. For instance, syllabic /r/ in Serbian can be a part of a vocalic interval.

in conjunction with structural (high frequency of clusters or long vowels) and prosodic (high frequency of stresses or phrase-level lengthened units[7]) factors.

The frequency factor, unlike other factors, has a stronger effect on the durational variability of a short speech sample, where both frequency of prosodic events and structural parameters vary more from sample to sample within the same language. This leads to higher variability of any measure that is a function of interval or syllable durations, and will impose a constraint on the length of speech samples in the studies of rhythm that use such measures.

2.4 ABSOLUTE VS. PERCEIVED DURATION. In this section, various factors that contribute to durational variability of segments and intervals were reviewed. Factors 1–10 affect the absolute duration of a segment, as measured from the acoustic signal. Not all durational differences, however, are equally perceptible. Those that are smaller than the "just noticeable difference" (JND) are inaudible to listeners.

However, some durational differences *larger* than the JND also seem to be factored out by listeners. For instance, when a speaker produces two vowels with equal intended durations, a high vowel will have a smaller absolute duration than a low vowel, but the two will be perceived as equal by the listeners. The difference in absolute duration in this example is an epiphenomenon of the articulatory movements and does not reflect a speaker's intended durational pattern.

I would like to emphasize that differences that are either imperceptible to or factored out by the listeners do not contribute to rhythm of speech. Thus, quantifying durational variability based on absolute durations will reflect perceived variability only if the effects of factors such as intrinsic duration variability are controlled for. This requirement has often been ignored in empirical studies. Later in this paper, I will propose two methods to facilitate measurements of perceived durations.

3. ASSESSMENT OF MEASURES USED AS CORRELATES OF RHYTHM-CLASS. In this section, I define the measures that are used in recent literature as rhythm classifiers. These measures are different functions of durations of appropriate units: phonological units such as syllable nuclei, syllables, or feet; or phonetic units such as vocalic and intervocalic, voiced and voiceless, or sonorant and non-sonorant intervals. In this paper, such measures are called *rhythm metrics* (RM), as they are often referred to in the literature.[8] Furthermore, I examine what kind of variability they are likely to capture, and compare the predictions to the results found in the literature.

3.1 DEFINITIONS, PREDICTED POWER, AND RELATION TO THE RESULTS IN THE LITERATURE. Let us assume that a given fragment of speech consists of vocalic and intervocalic phones grouped into uninterrupted vocalic and intervocalic intervals denoted as

$$C_1V_1C_2V_2C_3V_3...C_nV_nC_{n+1}.$$

Let d_{V_i} denote the duration of the *i*-th vocalic interval V_i $(1 \le i \le n)$, and d_{C_j} – the duration of the *j*-th intervocalic interval C_j $(1 \le j \le n+1)$. If the first syllable has zero-onset, then $d_{C_1} = 0$. If the last syllable is open, then $d_{C_{n+1}} = 0$. Let N_V denote the number of vocalic intervals (N_V is always equal to n), and N_C the number of intervocalic intervals. Note that, by the nature of division, intervals alternate and their numbers do not differ by more than one, i.e., $|N_V - N_C| \le 1$. Let also $\overline{d}_V$ denote the mean value of the vocalic intervals, and $\overline{d}_C$ the mean value of intervocalic intervals. This notation will be used throughout the paper.

3.1.1 PERCENTAGE OF VOWELS OR VOCALIC INTERVALS (%V). This metric is defined as the percent of duration of the speech sample—not counting pauses—that belongs to vowels. It is expressed as a ratio of duration of vowels to total duration

[7] The frequency of phrasal lengthening will be higher if there are more phrases in the sample of a given size, i.e., if phrases are shorter.

[8] In this paper, no a priori claims are made on how successfully such measures correlate with, or represent, the rhythm of speech.

$$\%V = 100 \cdot \frac{\sum_{i=1}^{N_V} d_{V_i}}{\sum_{i=1}^{N_V} d_{V_i} + \sum_{j=1}^{N_C} d_{C_j}},$$

or, after a simple transformation, as a function of the ratio of average durations of intervocalic and vocalic intervals

$$\%V = 100 \cdot \frac{1}{1 + \frac{N_C}{N_V}\frac{\bar{d}_C}{\bar{d}_V}} \approx 100 \cdot \frac{1}{1 + \frac{\bar{d}_C}{\bar{d}_V}}.$$

The formula suggests that this measure may be easily computed online by speakers, especially if it is evaluated regularly on short speech samples.

Intuitively, this rhythm metric distinguishes languages that are *more consonantal* (low *%V*) from those that are *more vocalic* (high *%V*). Note that the languages traditionally labeled as syllable-timed, French, Italian, or Spanish, are often perceived as more vocalic. Thus, the hypothesis found in the literature (Ramus et al. 1999) that *%V* is higher in syllable-timed languages is expected. However, this needs to be examined in turn by considering various factors.

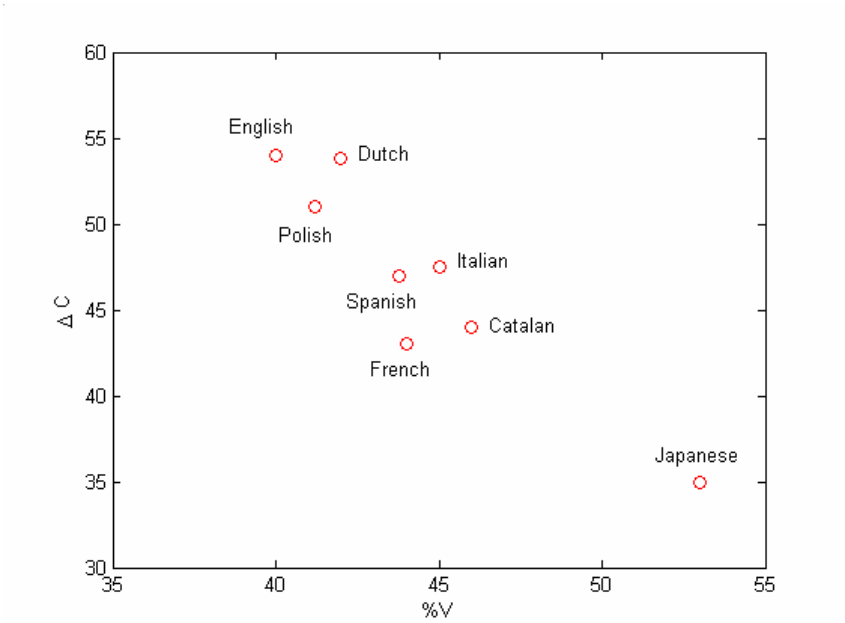
By its nature, a percentage measure depends on the distribution of all elements. Accordingly, *%V* depends on the number and average duration of both vowels and consonants in the speech sample. The value of *%V* will be larger if the frequency of diphthongs, long vowels, and other inherently long vowels is higher, and if the average number of consonants in the intervocalic intervals is smaller, i.e., if the syllable structure is simple. Given that prosodic prominence affects vowel durations more than consonant durations (Fant et al. 1981) the value of *%V* is larger if final lengthening applies to the language under examination, and if it occurs often—i.e., if the phrases are short.

The effect of prominence, however, is not clear. Languages in which prominence affects duration usually have lengthened vowels in prominent syllables, but shortened vowels in unstressed syllables, affecting the value of *%V* in different ways. Thus, the combined effect is not clear. This analysis is supported by the results of Ramus et al. (1999), which show that *%V* is larger in Catalan than in French and Spanish. For the convenience of the readers, results of Ramus et al. for *%V* and Δ*C* are repeated here in figure 1. While the syllabic structures of all three languages are similar, as quantified by the average number of consonants in a cluster, Catalan, unlike Spanish, exhibits vowel reduction in unstressed syllables. However, the value of *%V* obtained by Ramus and colleagues is higher for Catalan than for Spanish. This suggests that the sample sentences for Catalan had a larger average number of consonants in an intervocalic interval, or the lengthening of prominent syllables in Catalan outweighs the shortening in unstressed syllables.[9]

Other results in the same study confirm our predictions: languages with more complex syllabic structure, i.e., larger average number of consonants per intervocalic interval, such as English, Dutch, and Polish, had smaller *%V* than Italian, Spanish, Catalan, and French. The group of Romance languages, in turn, had smaller *%V* than Japanese, which has an even simpler syllable structure (low frequency of clusters) as well as phonemically long vowels.

[9] It is certainly worth examining how reduction in Catalan relates to duration.

FIGURE 1. Results for (*%V*, Δ*C*) from Ramus et al. 1999



3.1.2 STANDARD DEVIATION OF VOCALIC AND INTERVOCALIC INTERVALS (Δ*V*, Δ*C*). Standard deviation is commonly used as a measure of variability, expressed as the average distance from the mean value. It is used for vocalic and intervocalic intervals. The formula for intervocalic intervals is given below:

$$\Delta C = \sqrt{\frac{\sum_{j=1}^{N_C}(d_{C_j} - \overline{d}_C)^2}{N_C - 1}} = \overline{d}_C \cdot \sqrt{\frac{\sum_{j=1}^{N_C}(\frac{d_{C_j}}{\overline{d}_C} - 1)^2}{N_C - 1}}.$$

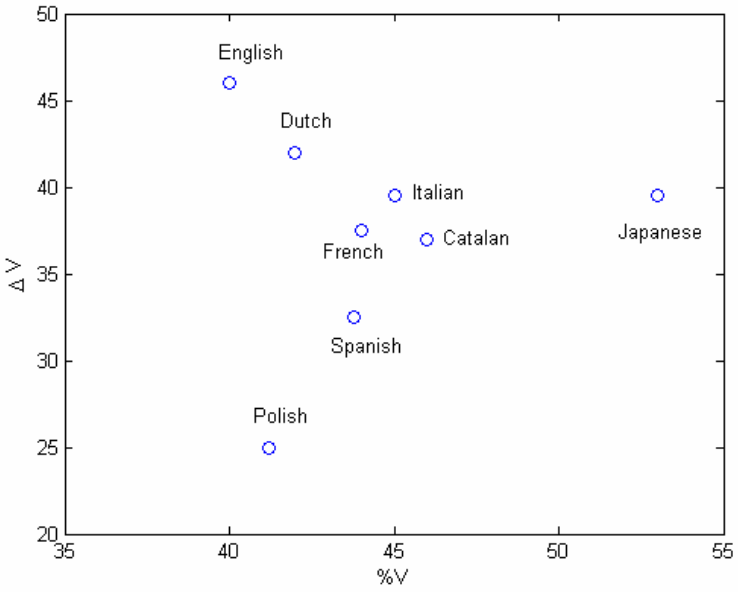
The standard deviation of vocalic and intervocalic intervals measures the overall durational variability of such intervals. Online computation of this metric is less plausible than that of *%V*, because it requires that the overall mean be subtracted from each sample. This would imply that durations of all the elements are memorized, at least for the duration of the sample over which it is calculated. Thus, this metric can be used as a research measure for classifying languages based on the acoustic signal, but it is improbable that it can explain human perception.

As a measure of overall variability, standard deviation will be bigger if the intervals vary more in their durations. Frequent hiatuses, as in Hawaiian, contribute to a larger value of *ΔV*. For languages in which hiatus is rare, most variability comes from the presence of long vowels and prosodic durational emphasis. Note that prominence lengthening and shortening of non-prominent vowels work in the same direction for *ΔV*, that is, make it bigger. Languages likely to have large *ΔV* include Hawaiian and Japanese (due to hiatus and long vowels), as well as English (due mainly to prosodic effects, both lengthening and reduction, and presence of diphthongs), to mention a few. Note that these languages have large *ΔV* for different reasons and might have quite different rhythmic types. Languages in which prominence is not strongly signaled by durational changes, and which do not have phonemically long vowels, will have small values of *ΔV* caused largely by the intrinsic duration variability of vowels.

The results of Ramus et al. (1999), repeated for the reader's convenience in figure 2, support the above prediction that languages with prominence-induced durational effects, such as English and Dutch, have larger values of *ΔV* than Spanish and French, in which such effects are smaller. Similarly, larger *ΔV* for Catalan than for Spanish can be explained by the comparatively larger effect of prominence

lengthening and, to some extent, by the presence of vowel reduction in Catalan.[10] Values for Italian and Japanese are comparable: slightly larger than those for the rest of Romance languages, but smaller than those of English and Dutch. This is expected based on the presence of long vowels in Japanese and presence of diphthongs and lengthened vowels in open syllables in Italian.[11] Thus, the results do not support the separation into traditional stress-, syllable-, and mora-timing (with mora-timing seen as extreme syllable-timing), but they support well our predictions based on different factors of durational variability. Finally, an extremely small value that Ramus et al. find for *ΔV* in Polish could be explained as due to a strict one-vowel-per-interval distribution, a lack of either diphthongs or long vowels, and a lack of vowel reduction in unstressed syllables (suggesting that prominence-induced lengthening in stressed syllables is not significant).

FIGURE 2. Results for (*%V*, Δ*V*) from Ramus et al. 1999



Standard deviation of durations of intervocalic intervals is dominated by the syllable-structure effect because prominence and phrasing effects are much smaller for consonants than for vowels, and because two consonants in a cluster often have a duration close to double that of a single consonant. For a language with simple syllable structure, intrinsic consonant duration will contribute to larger *ΔC*. These predictions are supported by the results obtained in Ramus et al. 1999: Polish, English, and Dutch in Ramus et al. have large *ΔC*; Catalan, Italian, Spanish, and French have intermediate values (clusters exist but are simpler than in English, Polish, or Dutch); while the value of *ΔC* for Japanese is small (since it has an even simpler syllable structure than Italian and French).

3.1.3 COEFFICIENT OF VARIATION OF VOCALIC AND INTERVOCALIC INTERVALS (*VarcoV, VarcoC*). The dependence of standard deviation on speech rate was pointed out by Ramus (2002) and shown empirically in studies such as Barry et al. 2003 and Dellwo and Wagner 2003. Ramus suggested that dividing standard deviation by the mean would provide necessary rate normalization.[12] The resulting metric is well known in statistics as the coefficient of variation. It is used for both vocalic and intervocalic intervals. Here, the formula is given for vocalic intervals:

[10] In the analysis of results for *%V*, I suggested that effects of vowel reduction are small compared to effects of prominence-induced lengthening.

[11] A certain level of prominence-induced duration can also be posited for Italian.

[12] It was first used in the literature by Dellwo (2006).

$$VarcoV = \frac{\Delta V}{\overline{d}_V} = \sqrt{\frac{\sum_{i=1}^{N_V} (\frac{d_{V_i}}{\overline{d}_V} - 1)^2}{N_V - 1}} .$$

Coefficient of variation has an advantage over standard deviation because it allows comparison of variables with different means (http://www.ats.ucla.edu/stat/). Thus, it is useful for comparison of interval durations at different speech rates.

Varco captures same kind of variability as standard deviation, so similar predictions are made, except that *Varco* is claimed to be able to compare samples produced at different speech rates. Results in the literature are somewhat contradictory: Dellwo (2006) found that *VarcoC* has superior classification capability compared to *ΔC*, in that it normalizes speech rate differences of the samples; however White and Mattys (2007) obtain completely overlapping values of *ΔC* for Italian, French, Dutch, and English. A possible cause for the lack of distinction by this metric may be in the deficiencies of the normalization process, which will be discussed in 3.2.

3.1.4 RAW PAIR-WISE VARIABILITY INDEX (*rPVI*). In its original form, the pair-wise variability index was defined as an average absolute difference in duration between the consecutive vowels in the speech fragment. To distinguish it from its normalized counterpart, this index is called the *raw* pair-wise variability index (*rPVI*). It can be used as a measure for both vocalic and intervocalic variability. Here, the formula is presented for intervocalic intervals:

$$rPVI - \mathrm{C} = \frac{1}{N_\mathrm{C} - 1} \cdot \sum_{k=1}^{N_\mathrm{C}-1} | d_{Ck} - d_{Ck+1} | .$$

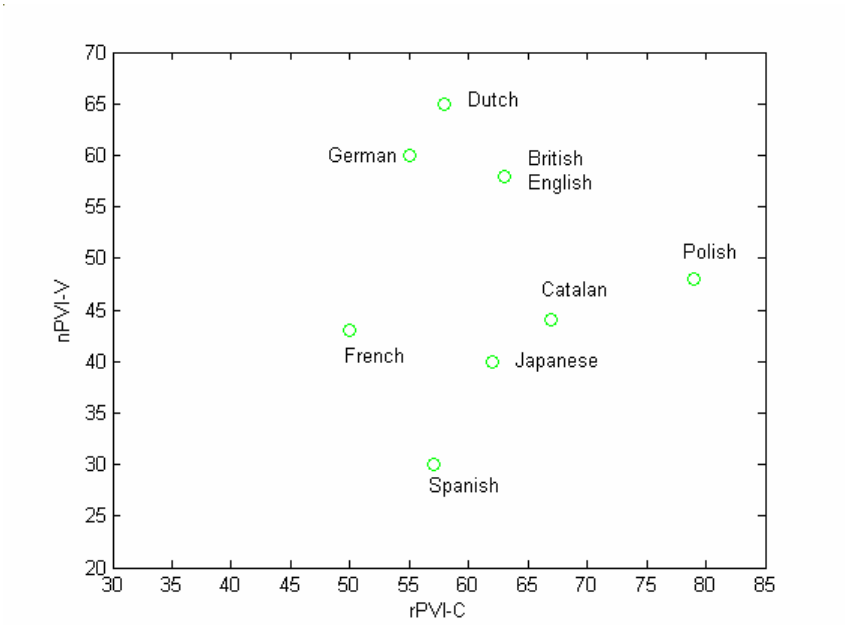
The difference between the standard deviation and the pair-wise variability index is that standard deviation measures overall variability (i.e., dispersion from the average value), while the pair-wise variability measures how much a unit differs on average from the neighboring units of the same type. In that sense, pair-wise variability attempts to quantify the sequential[13] aspect of rhythm—a property that other measures lacked.

Pair-wise variability of intervocalic intervals is affected most by the number of consonants in the interval because, as mentioned in the discussion of the standard deviation, prosodic effects on the duration of consonants are not strong. Thus, this metric is successful in classifying languages based on the complexity of their syllabic structure.

Intrinsic durational variability, as in the case of standard deviation, prevents this metric from expressing prosodic effects on duration, because a non-prominent inherently long neighboring segment may be as long as a prominent inherently short neighboring segment of the same type (e.g., compare duration of unstressed /aɪ/ and stressed /ɪ/ in /maɪ ˈbɪn/ ‘my bin’). If the variability due to syllable structure and intrinsic durations is controlled for, pair-wise variability reflects the prominence and phrasing effects on duration, and as such—because it captures sequential variation—it is predicted to be more successful than the standard deviation.[14] However, these conditions were not often met in previous studies, as this metric was computed on unmodified speech samples that were affected by syllable structure, intrinsic duration, and prosodic factors—all at the same time. Thus, the results show limited success in classifying languages. For instance, the results of Grabe and Low (2002:528), presented in figure 3 for the reader’s convenience, show almost no distinction on this dimension (*rPVI-C*) among languages in three posited rhythm classes. In the results for a larger sample of languages (p. 530), *rPVI-C* does not reflect either the posited rhythm class of a language or the syllabic structure. The metric seems to combine too many factors, including strong dependence on speech rate, and the continuation of its application can be attributed only to the relative success of the accompanying metric—normalized pair-wise variability of vocalic intervals.

[13] Sequential aspect is captured in its simplest form though, through the relation between neighboring units.
[14] Prominent units are usually surrounded by less prominent ones.

FIGURE 3. Results for (*n-PVI-V*, *r-PVI-C*) from Grabe and Low (2002)



3.1.5 NORMALIZED PAIR-WISE VARIABILITY INDEX (*nPVI*). Just like standard deviation, *rPVI* is highly dependent on speech rate, as segments are shorter in faster speech. Thus, the difference between consecutive units is reduced as well. To normalize the raw PVI index, Grabe and Low (2002) divided each term by the local mean, i.e., the term corresponding to the durational difference between intervals *k* and *k*+1 is divided by the average duration of those two intervals, as shown in the formula for *nPVI-V*, the pair-wise variability of vocalic intervals:

$$nPVI-V = \frac{100}{N_V - 1} \cdot \sum_{k=1}^{N_V} \frac{|d_{V_k} - d_{V_{k+1}}|}{\frac{d_{V_k} + d_{V_{k+1}}}{2}} = \frac{200}{N_V - 1} \cdot \sum_{k=1}^{N_V} \frac{|\frac{d_{V_k}}{d_{V_{k+1}}} - 1|}{\frac{d_{V_k}}{d_{V_{k+1}}} + 1}.$$

The alternative formula shows that this metric is a function of the ratio of subsequent elements, and that it results in equal values for equal percent changes between two subsequent elements.

Raw pair-wise variability of vocalic intervals, like standard deviation, is high for languages with high frequency of hiatus (supported by the results of Parker Jones (2006) for Hawaiian), for languages with long vowels and diphthongs, and for languages that show durational effects of prominence, including phrasal prominence.

The results of Grabe and Low (2002:528), based on *nPVI-V*, support the distinction of traditional syllable-timed and stress-timed classes. These results can be explained as the joint effect of lengthened stressed vowels and reduced unstressed vowels (resulting in high values of *nPVI-V* for German, Dutch, and British English). While some prominence-induced durational effects exist in French and Spanish, the lack of reduction in unstressed positions, especially in the neighborhood of stressed positions, keeps the value of *nPVI-V* low for these two languages. Additionally, phrasal lengthening likely has less effect on *nPVI-V* than it does on the standard deviation, as the lengthening happens gradually, and changes with respect to the previous syllables are not large.[15]

[15] See Turk and Shatuck-Hufnagel 2007 for a more detailed description of the final lengthening mechanism in American English. Their results show that lengthening is non-linear and that the effect is strongest on the main-stress syllable rime and the phrase-final syllable rime.

We can conclude that the large variations in *nPVI-V* obtained for different languages reflect the joint contributions of various uncorrelated factors that occur in different languages as much as the posited continuum of rhythmic differences. If only prosodic factors were present, a sequential measure like the pair-wise variability index might be successful in capturing differences between languages in which prominence is producing important durational differences and those in which it is not.

3.2 DEPENDENCE OF RHYTHM METRICS ON SPEECH RATE. In the literature, some metrics have been criticized more than others for their dependence on speech rate. In particular, standard deviation and raw pair-wise variability were shown to change significantly when the speech rate changes. This is expected, as they are expressed in time units, milliseconds for instance, and are related to the average duration. Two types of normalizations were proposed in the literature: with respect to average value over the whole sample, and with respect to local average of two subsequent units. Metrics that are normalized by either of these methods are posited in the literature to be speech-rate independent. These metrics are functions of ratios of interval durations, as shown by their formulae in section two.

In order for metrics that are functions of ratios (percentage, *Varco*, and *n-PVI*) to be independent of speech rate, all units are required to change proportionally when the rate changes, say k times. Simple manipulation of the formula shows that lengthening factors cancel out and the metric for the new speech rate results in the same value. The lack of complete invariance of these metrics with respect to speech rate is a result of different elasticities of segments with respect to speech rate, as discussed in the previous section.

Studies in the literature show that metrics change values when speech rate changes, especially the standard deviation measure (Dellwo and Wagner 2003, Barry et al. 2003, among others). The empirical results of a short study I conducted (Stojanovic 2008) show that *all* metrics depend on speech rate, even though the normalized ones vary less. The importance of this dependence, however, is related to the amount of change. If the range of values of a metric for different rates is sufficiently far from the range of values that are obtained for a different rhythm class, then the effect of speech rate on that metric is tolerable. If, however, values of that metric for some rates overlap with the values corresponding to another class, then speech rates for the two languages whose rhythms are compared need to be controlled.

4. PRACTICAL ISSUES RELATED TO COMPUTATIONS OF RHYTHM METRICS.

4.1 SPEECH RATE. As discussed in the previous section, values of all rhythm metrics change when the articulation rate changes, although some metrics vary more than others do. It is important to notice that, due to different elasticity coefficients for different segments, levels of stress, and phonological length, not even so-called normalized metrics are independent of speech rate.

An important question to answer is whether a similar effect characterizes perception—that is, whether speech at higher rates is perceived as having different rhythm, or in terms of duration variability, is less variable. If so, rhythm metrics reflect a real property of speech communication, but if not, then the sensitivity of RMs to speech rate is problematic. More generally, we should ask what the relationship between rhythm and articulation rate is: whether rate is one of the dimensions of rhythm or whether it affects rhythm only as an external factor.

Dellwo and Wagner (2003:473) propose that the average speech rate for a particular language is affected by the language's phonology and phonotactics and that in fact there is a characteristic average speech rate for each language expressed in syllables per second. While the evidence for characteristic average rates may be found, large variation across speakers of the same language undermines the effort to characterize language rhythm based on speech rate. Namely, it is possible that slow speakers of French (which is suggested to have a high characteristic speech rate) overlap in speech rate values with fast speakers of English (which is suggested to have a low characteristic speech rate). Intuition suggests that slow French speech would not be confused perceptually with the fast English speech,[16] but it is of course an empirical question.

[16] Just as a slow waltz is not likely to be confused with a fast tango on the rhythm dimension, although they might be characterized as mid-tempo rhythms.

Two outcomes are possible. If listeners classify samples based on the language and not on the rate, i.e., if fast English is more similar to slow English than to either fast or slow French, then the listeners are likely to rely on rate-invariant durational properties to make decisions, and our goal is to discover such properties. Support for such a model, in which durations are perceived in a relative manner, comes from the research on phonemic vowel length. Kozasa (2005) found that the ratio of long to short vowels in Japanese is different at different speech rates. However, the distinction between long and short vowels is retained. Additional support comes from the fact that listeners deal with significant variability of rates within and across speakers and even within utterances, and thus global rhythmic properties should be evaluated independently of rate.

Another possible outcome is that the perceived rhythm of a language is dominated by the speech rate, i.e., the fast speech samples in different languages are judged as more similar than samples of the same language produced at different rates. If this is true, however, then languages cannot be classified into types based on rhythm: they would be distinguished solely based on how slow or how fast they are spoken.

The only scenario in which it would be possible to maintain the assumption of language rhythm classes is if all the possible rhythms at different speech rates in one language class are more similar to each other, than they are to the rhythms that are characteristic for another rhythm class.

One needs to be careful when investigating how the rhythms of two speech samples relate to each other and whether similarity is judged on rate or the underlying durational pattern. An important starting point would be to clearly define what is understood as *speech rhythm*. Definitions in the literature, such as *patterning of weak and strong, or short and long, beats* imply that pattern is independent of absolute rate and that terms *weak*, *strong*, *short*, and *long* are relative.

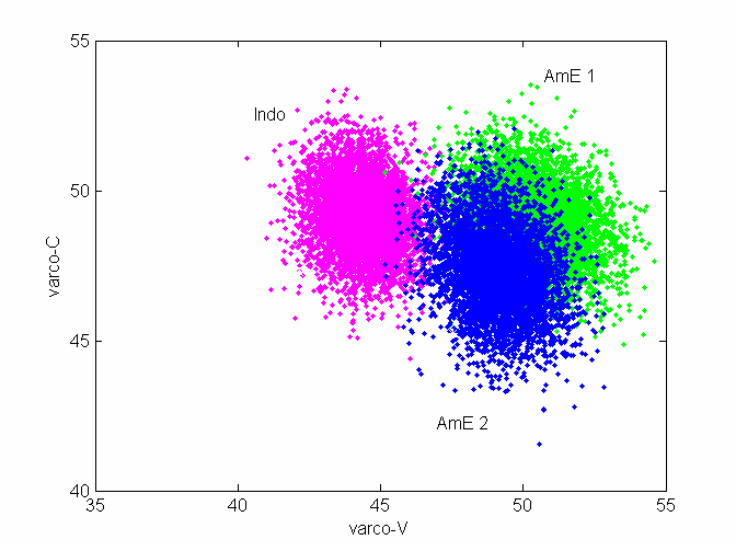
4.2 SAMPLE SIZE. As discussed in 2.3, rhythm metrics are affected by frequency of different phones as well as by the frequency of prominences within a given speech sample. While frequency of a particular phone within a language is sometimes considered constant, on a sample of finite-length phone frequencies vary from sample to sample, particularly for short samples. This implies that values for each metric will vary depending on the chosen speech materials.

One way to achieve stability of phone frequencies, and to a certain extent prominence frequencies, is to use long speech samples, or averages over many short ones. However, such practice raises a question of how rhythm metrics relate to the process of discrimination by human listeners because human listeners do not require long samples of speech in order to make a distinction. This insight, again, shows that even if rhythm metrics are able to distinguish between language rhythm classes, they do not mimic human perception.

4.3 SEGMENTATION. Severe lack of precision in segmentation (Stojanovic 2009) and lack of agreement on the position of the border between two units (Stojanovic 2008) can lead to significant differences in the values of considered metrics. As a result, the segmentation process affects the location of the language in the space defined by rhythm metrics and therefore the grouping of languages in such space.

Results from Stojanovic 2009 for the effect of segmentation precision on values of *Varco-V* and *Varco-C* are given in figure 4. When the uncertainty of segmentation borders is uniformly distributed between 0 ms and 25 ms, values of both metrics vary in a range of 5 points for two American English speakers and one Indonesian speaker. White and Mattys (2007) obtain a 14-point difference for *Varco-V* between French and English, 23 points between Spanish and English, which suggests that uncertainty would not affect classification of French and English (or Spanish and English) into different classes based on *Varco-V*. However, the authors obtain a 3-point difference for *Varco-C* between French and English and only 1 point between Spanish and English; therefore, segmentation uncertainty of 5 points would compromise classification of Spanish, French, and English based on *Varco-C*. Thus, some metrics are affected by the segmentation process more than others.

FIGURE 4. Simulation of the effect of uncertainty on segmentation border placements (presented in Stojanovic 2009)



From the methodological point of view, these results suggest that for a meaningful comparison both across studies and for cross-linguistic data, segmentation rules need to be agreed upon and applied with the required degree of precision.

Proposals to avoid segments (phones) whose durations cannot be reliably determined already exist (Turk et al. 2006). Such requirements, however, impose significant restrictions on possible speech materials. In fact, if only clearly segmentable phones should be included, the speech would need to consist of only vowels and obstruents. We will revisit this idea at the end of the paper.

In the light of articulatory phonology (Browman and Goldstein 1989), which proposes that speech consists of a sequence of overlapping articulatory gestures rather than individual phones, even the question of where a border falls between a voiceless plosive and the following vowel cannot be answered with absolute certainty. Namely, as several gestures are involved in the production of each sound and the transitions are not perfectly timed, there is more than one possible boundary to consider: e.g., closure release, end of burst, start of formant structure, and start of voicing. Conventionally, in a stop-vowel sequence the start of voicing is considered the beginning of the vowel, but based on the saliency of auditory cues, stop release is a more likely contender. Better evidence, though, would be related to where the beats are perceived, and how the duration of the unit is evaluated by the listener. Perception of beats is related to the research on *perceptual centers* (Fowler 1979), which are posited to be *at* or *near* the vowel onset. Descriptions of beat-locations that are more precise are needed, together with a relationship between the distances among consecutive beats and the perceived durational sequence. I plan to address these issues in future research.

A different question we might ask in relation to segmentation is how the practical requirements on precision in measuring durations relate to variation in production and to the JND for the perception of duration. Detrimental values of segmentation noise for some metrics amount to only 20-25 ms at normal speech rate (Stojanovic 2009). While there are some indications that listeners can perceive differences in duration of that order, it would be useful to know whether an artificial modification of durations would be perceived as rhythmically inaccurate, or different from the original. If such differences in interval durations do not change the perceived rhythmic sequence, then metrics used to describe rhythm should be less sensitive to absolute durations.

4.4 RHYTHM-BEARING UNIT AND UNITS USED FOR ANALYZING RHYTHM. A natural choice for a rhythm unit is a *beat*. When talking about the rhythm of language, however, various units have been used with different justifications. Early speech-rhythm theory (Pike 1945, Abercrombie 1967) was based on the hypothesis of isochrony of stresses (or feet), and syllables, under the assumption that these units were

related to two types of breathing and thus two types of rhythmic beats. Many empirical studies were conducted (Lehiste 1977, Roach 1982, Dauer 1983, to mention a few), but no evidence in support of the theory was found and the idea of a relation between syllable and stress production to different kinds of breathing was abandoned.

Vocalic and intervocalic intervals were proposed by Ramus and colleagues (1999) as "neutral" units because of their salient difference to infants, who were claimed to perceive differences between language rhythm classes. The authors summarize the salient differences as: vowels have more energy, are on average longer than consonants, and carry prosodic information. In addition, infants are found to pay more attention to vowels (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, and Mehler 1988) and to be able to count syllables irrespective of syllable-complexity (Bijeljac-Babic, Bertoncini, and Mehler 1993). Finally, they suggest (Ramus et al. 1999:270) that: "the infant primarily perceives speech as a succession of vowels of variable *durations* and *intensities*, alternating with periods of unanalyzed noise (i.e. consonants)."

Ramus and Mehler (1999:512) argued that, if infants are able to discriminate languages based on rhythmic differences, then "the speech signal must contain some prelexical cues that enable language discrimination." Their choice includes succession of vowels, called *vocalic intervals* and periods of noise, called *intervocalic intervals*. Researchers (Grabe and Low (2002), White and Mattys (2007), and many others) followed this approach. Some other proposals, similarly argued for by infants' ability to discriminate among rhythmically different languages, use *voiced* and *voiceless* intervals (Dellwo et al. 2007) or *sonorant* and *non-sonorant* intervals (Galves et al. 2002) as units. Recently, criticisms (e.g., Nolan and Asu 2009) of the fit of such units to rhythms of certain languages prompted a return towards more linguistic units such as syllables and feet.

A comment is in order regarding the choice of rhythm-bearing units and infant perception. There seems to be a misinterpretation of infants' ability with respect to linguistic units. Namely, it seems that the ability to "count the number of syllables (and therefore vowels) in a word, independently of syllable structure or weight" (Ramus et al. 1999:270) argues *for* recognizing syllables, not *against* it. In fact, when syllables consist of nuclei only, being able to count syllables argues against the vocalic interval as a single unit. In addition, if infants perceive durations and intensities, as Ramus et al. (1999) suggest, it is possible that they can extract prominence from the speech signal, and identify stressed (or simply prominent) nuclei. This would be sufficient for perception of feet and for computing foot (or inter-stress interval) durational variability even if consonant stretches in between remain unanalyzed.

The posited perceptual equivalence between possible realizations of phrases such as *She left at noon* /ʃi ˌlɛft ət ˈnun/ and *He brought a broom* /hi ˌbɹɒt ə ˈbɹum/ argue against using absolute interval differences to describe rhythmic sequences, as among corresponding intervals /ʃ/ is longer than /h/, /ft/ is longer than /t/, while /l/ much shorter than /bɹ/. Another line of reasoning can be used to argue against proposals that use voiced and voiceless or sonorant and non-sonorant intervals: it is unlikely that phrases *I owe you a bag* (consisting of all voiced segments) or *I owe you a mule* (consisting of all sonorants) are perceived as single rhythmic units, as the respective proposals would imply. Even when segmental qualities are masked, intensity and pitch cues seem sufficient to recover several beats from such utterances.

Based on the assumption that beats can be recovered using suprasegmental cues, vocalic intervals should be separated into different vowels (nuclei), or at least the prominent vowels should be considered independent intervals from the neighbors. In this way, a prosodic effect that applies on the syllable that is prominent with respect to its neighbors within the same interval can be captured, and such a case can be distinguished from a simple hiatus of three consecutive equally prominent vowels.

Similarly, we can argue against considering consonant intervals as relevant to rhythmic structure. Intrinsic durations aside, the coda of one syllable and the onset of the following syllable, especially across a word boundary, may experience quite different prosodic treatments, based on the length of the words in which they occur and the prominence that their respective syllables receive. If considered together, the composite effect might be quite difficult to analyze.

To sum up, none of the three divisions into contrastive elements seems to correspond to the beats and pauses dichotomy used to describe the rhythm of music. Based on agreement of speakers and listeners on duration judgments (Tuller and Fowler 1980), it seems that rhythmic sequences do exist in the speech signal, but need to be recovered from the acoustic signal using the perceptual filters of listeners' auditory and processing systems. Studies establishing *equivalent durations*, i.e., explaining how the durations of different types of syllables are perceived, would be a step toward understanding the rhythmic nature of speech.

4.5 SHOULD FINAL LENGTHENING COUNT TOWARDS VARIABILITY? Phrasing is one of the prosodic factors that affect our perception of durational and intonational patterns in a given language. An example of a durational effect is phrase-final lengthening, which is observed to various degrees in many languages.

An argument for including syllables at the phrase-edge into computation of variability is that final lengthening is salient to listeners and contributes to the overall perception of speech rhythm. In fact, Arvaniti (2009) posits that prominence and phrasing are the two properties that define the speech rhythm.

On the other hand, one might argue that durational variability caused by final lengthening exists universally and thus does not contribute to the perceived difference between two languages. For instance, Fant et al. (1991) show that French and English, usually considered prototypes of two rhythm classes, differ clearly in phrase-internal prominence realization, but not in the durational effect of final lengthening. Given that most rhythm metrics reflect average durational variability, it is possible that short phrases with phrase-final lengthening but no phrase-internal prominence and long phrases with internal prominence and phrase-final lengthening will result in the same value of the rhythm metric. This would result in missing the distinction between two significantly different speech samples.

In their first study on differences between British English and Singapore English, Low, Grabe, and Nolan (2000) eliminated phrase-final syllables. In this way, they focused on comparison of durational variability that does not include phrase-edge effects. In the subsequent study (Grabe and Low 2002), when comparing 18 different languages, the authors opted against exclusion of phrase-final syllables, arguing that it would be a difficult to determine phrase boundaries in languages they did not speak.

Putting such difficulties aside for a moment, a reasonable solution is to quantify differences for both a version in which the final units are excluded and one in which they are present. In such a way, different components of the durational and intonational structure of two languages can be compared.

5. ISSUES RELATED TO QUANTIFYING RHYTHM.

5.1 SEGMENTAL CONTENT MASKS PERCEIVED DURATIONAL VARIABILITY. Different views of linguistic rhythm so far include the following: (1) rhythm cannot be observed in the acoustic signal but is only a perceptual phenomenon (Lehiste 1977), (2) rhythm is a result of phonological properties of language such as syllabic structure and processes such as vowel reduction (Dauer 1983), and (3) rhythm is present in the acoustic signal as variability of vocalic and intervocalic units (Ramus et al. 1999, Grabe and Low 2002). Based on the discussion in this paper, none of these views seems to correctly describe the relation between perceived speech rhythm and the acoustic signal, although the first assessment is closest to our view. I propose that, while information about the rhythmic sequence—expressed as a variation of prominence—is present in the acoustic signal, it is masked by the variation in segmental durations.[17] This variation is factored out by the native listener, based on the known information about segmental qualities, pitch, and possibly other acoustic properties.[18]

I argue, contra Dauer's proposal (1983), that syllabic structure is only a frequent typological correlate of rhythm, and that perceived rhythmic differences among languages are due to prosodic factors. Starting from the assumption (the thesis of this paper) that perceived rhythmic sequences differ from sequences based on absolute durations seen in the speech signal, and from the fact that beats (perceptual centers) in

[17] For example, a complex onset takes much longer to produce than a simple one, even though this does not affect the perceived rhythmic sequence.

[18] If underlying rhythmic sequences were isochronous for a given speech sample, this view would be consistent with Lehiste's (1977) proposal that isochrony is perceptual.

speech do not always coincide with segmental borders, I conclude that rhythmic structure in speech is best observed when segmental effects on durations in the acoustic signal are controlled or eliminated.[19]

Segmental and interval durational effects, which act as noise with respect to the rhythmic sequence, do not only prevent us from finding invariance in rhythmic properties of the speech samples within a language or language class, but they also compromise our ability to find cross-linguistic similarities and differences. Results of the empirical studies in the literature agree with perception results only when the rhythmic similarities or differences are supported by segmental durational effects (as when a language with high prosodic variability has high phonotactic variability).

The frequent practice of analyzing paragraphs that represent translations of the same story into different languages is employed in order to control for type of speech, overall mood, and repetitions that exist in the text. This approach may somewhat reduce the number of factors that contribute to durational variability, but it does not eliminate the segmental and phonotactic durational effects. Possible ways to eliminate these segmental and phonotactic effects on durations are considered next.

5.2 WAYS TO ELIMINATE SEGMENTAL EFFECTS. Low-pass filtering has been used frequently in preparation of stimuli for rhythm perception to eliminate segmental (lexical) information from the signal. This technique eliminates information necessary to perceive segmental qualities if the relevant cues are above the cut-off frequency of the filter, but it preserves the true durations of segments in the signal. Such filtered stimuli are then presented to the listeners, who judge whether two samples are rhythmically the same or different. Additionally, to isolate perceptual effects of duration, pitch variations over the samples can be neutralized as well, as described below.

Criticizing low-pass filtering as a method, Ramus and Mehler (1999:513) suggest that low-pass filtering "does not allow one to know which properties of the signal are eliminated and which are preserved." A more serious critique of this method is that if rhythm is perceived holistically, based on duration, pitch, intensity, and segmental quality, then eliminating certain factors without adjusting for their joint effects will change our perception of rhythm. There are some indications that pitch shape influences perceived duration, at least in some languages (Lehiste 1976, Kozasa 2005, Cumming 2008, Lippus et al. 2009). For example, given two time intervals of equal duration, the one that carries flat pitch will be perceived as shorter than the one that carries a contour pitch. Thus, when pitch is simply flattened, two intervals will be perceived as equal in length. Another example is that, when perceived as equally long, /i/ is shorter in absolute duration than /a/. If low-pass filtering masks the vowel quality, however, /i/ will be perceived as shorter than /a/ because information on vowel quality that normally can be used for factoring out segmental effect will not be present. This casts doubt on filtering as a technique for preparing data for experiments on perceived duration, unless such durational differences between samples are smaller than the JND for duration.

Ramus and Mehler (1999) propose resynthesis as a better technique with which one can change signal properties independently. During resynthesis, a transformation is applied to segments so that their durations are preserved, but the qualities are replaced by different segments, depending on the scenario. In a *saltanaj* version, segments are replaced by the representative for each group (e.g., all fricatives are replaced by /s/). In a *sasasa* version, all consonants are replaced by /s/. In a *flat sasasa*, pitch contours are replaced by flat pitch. Ramus and Mehler argue that with resynthesis we can test the perceptual effects of phonotactics, intonation, and rhythm independently. However, if adjustments to durations are not performed in resynthesized versions, for instance correction for duration based on pitch flattening, the process will suffer from the same problem as filtering, i.e., change of perceived durations.

Testing the underlying rhythmic sequence free of segmental and structural effects is possible via two methods: (1) reiterant speech, and (2) tapped sequences, which are described next.

5.3 DISCOVERING UNDERLYING PROSODIC STRUCTURE. One way to control for the effects of intrinsic durations and syllabic structure is to consider a reiterant variant of the original speech sample. In these variants, each syllable is replaced by the *reiterant syllable*, for instance /da/,[20] but the prosody of the

[19] Speech rate variability within utterances needs to be controlled for as well.

[20] The vowel occurs in its reduced variant if the prosody of the language in question requires it.

reiterated phrase is matched to that of the original phrase. In a reiterated phrase, only prosodic factors affect the duration of segments, and thus quantified durational variability corresponds more closely to *intended* rhythm. In a study (Barry et al. 2009) that compared textual[21] and reiterated versions of speech for three sample languages, it was found that the rhythm metrics *%V* and *nPVI*-syllable reflected rhythmic differences between different meters. However, reiterant speech has not been used to compare differences among languages from different posited rhythm classes. Future studies might benefit from this method in that the prosodic effects on duration of vowels and consonants in the reiterated sequences are not mediated by structural effects.

Another way to access underlying rhythmic structure is to consider tapped imitation of the given sample. In an experimental setting, participants are asked to imitate the original sequence with a series of taps, so that one tap corresponds to each syllable. In this way, a perceived time distance between two syllabic beats is reflected in the interval between two consecutive taps. To my knowledge, no study so far has used tapped sequences to quantify durational variability through rhythm metrics. One reason might be that tapping is a difficult task: not all speakers are able to mimic the rhythmic sequence. Another reason for the lack of such studies is that there is no agreed upon criterion for judging similarity of the original and tapped sequences. Despite difficulties, I believe that such a study would be rewarding in attempts to quantify rhythmic differences.

5.4 RHYTHM AND PITCH. Most studies so far have examined only durational variability in an attempt to classify languages into different rhythm groups. However, rhythm is sometimes defined more holistically using a combination of duration, intensity, pitch, and vowel spectral qualities. There are two basic questions on the relation between pitch and rhythm. The first one is whether pitch affects our perception of duration. The second is whether pitch variability alone can tell us something about differences perceived between two languages described as being rhythmically different.

As mentioned in 5.2, some studies suggest that pitch shape influences perceived duration. Therefore, in addition to intrinsic duration, pitch effects on the duration of subsequent units need to be included in the model of rhythm.

Coming back to the second question, we would like to know whether the variability of pitch levels and contours is significantly different in languages that are perceived as rhythmically different. Some support for this can be found in the literature on automatic language classification (Rouas 2007) and the perception of rhythm by infants (Ramus 2002). Thus, a model in which pitch and duration are combined into a general prosodic model is expected to explain human perception more faithfully.

5.5 MODEL. According to work by Fowler (1979), the most likely successful model of rhythm production and perception should be based on the existence of underlying rhythmic sequences, known to both speakers and listeners, onto which segmental information is mapped to encode meaning.[22] Because of the inertia of articulatory movements, the segmental information distorts the temporal template of the underlying sequence but does not destroy it, as the listener is able to decode it.

Fowler (1983) proposes that these underlying rhythmic units are vowels. There are two possibilities with respect to timing of consonantal gestures. In the first case, consonantal gestures completely overlap with the intended vocalic intervals and the points of "beats" corresponding to the underlying sequences can be found somewhere in the signal. This view is consistent with the theory of perceptual centers (Morton et al. 1976, Fowler 1979). Alternatively, consonantal gestures partially overlap with the vocalic ones, but additional time is allowed to accommodate full articulation of the consonant (or a cluster). In this view, the beats of the intended rhythmic sequence are separated by the intervals corresponding to the additional time allowed for the production of consonants, and thus the absolute duration of the realized speech sequence is different from the duration of the intended sequence.

Another question is worth considering in relation to how listeners classify languages based on prosodic properties. The approaches discussed in this paper rely on computation of durational statistics

[21] In textual versions, segmental qualities are preserved.

[22] Donegan and Stampe (2004:30) call this process *putting syllables onto accents*, contrary to the view in which *beats/accents are determined by segments*, such as in Duaer's (1983) proposal.

(percentages, standard deviations, coefficients of variations) of speech samples. In a different approach, languages or language groups are differentiated based on recognition of characteristic prosodic events. The second model differs from the first in that not all information in the signal is equally important or prominent, and in that rhythm perception is better modeled by pattern recognition than by statistical computation. Such a model would benefit from evidence that differentiation by listeners can be accomplished based on short speech samples. It is also supported anecdotally by naïve listeners' explanations of how they differentiate speech samples.

In sum, to model rhythmic and prosodic classification, the following need to be investigated: (1) computation of perceived durations based on absolute durations, pitch, intensity, and segmental quality, (2) non-linear mapping of the intended rhythmic sequences to produced speech sequences, and (3) comparison of statistical and pattern recognition approaches.

6. CONCLUSION. The goal of this paper was to examine in detail the quantitative approach to rhythm in which rhythmic similarity is evaluated by use of rhythm metrics. After a discussion of the factors that cause durational variability of segments and intervals in a speech sample, predictions were made about what kind of classification each rhythm metric can accomplish. These predictions were shown to be supported by the results in the literature. The discrepancies between the results and the predictions made based on the rhythm class hypothesis were clearly explained by individual variability factors.

In the second part of the paper, problems with the approach, such as dependence on speech rate and syllabic structure, were pointed out and discussed. A particularly problematic issue for the current model of rhythm is that only absolute duration is in fact being measured. This is because the effects of pitch, segmental quality, and intensity on perceived duration are not being taken into account.

Finally, proposals were made on how to remedy deficiencies of the current approach while keeping the idea of difference in perceived variability as a criterion for distinction. To this end, a model was discussed that will be tested in the continuation of the present work.

REFERENCES

ABERCROMBIE, DAVID. 1967. *Elements of general phonetics*. Chicago: Aldine Pub. Co.

ARVANITI, AMALIA. 2009. Rhythm, timing and the timing of rhythm. *Phonetica* 66:46–63.

BARRY, WILLIAM; BISTRA ANDREEVA; and JACQUES KOREMAN. 2009. Do rhythm measures reflect perceived rhythm? *Phonetica* 66:78–94.

BARRY, WILLIAM; BISTRA ANDREEVA; MICHELA RUSSO; SNEZHINA DIMITROVA; and TANJA KOSTADINOVA. 2003. Do rhythm measures tell us anything about language type? In *Proceedings of the 15th International Congress of Phonetic Sciences,* ed. by D. Recasens, M. J. Solé, and J. Romero. 2693–96. Barcelona.

BENTON, MATTHEW; LIZ DOCKENDORF; WENHUA JIN; YANG LIU; AND JEROLD A. EDMONDSON. 2007. The continuum of speech rhythm: Computational testing of speech rhythm of large corpora from natural Chinese and English speech. In *Proceedings of the 16th International Congress of Phonetic Sciences*, ed. by J. Trouvain and W. J. Barry, 1269–72. Saarbrücken.

BERTONCINI, JOSIANE; RANKA BIJELJAC-BABIC, PETER W. JUSCZYK, LORI J. KENNEDY, and JACQUES MEHLER. 1988. An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General* 117(1):21–33.

BIJELJAC-BABIC, RANKA; JOSIANE BERTONCINI; and JACQUES MEHLER. 1993. How do four-day-old infants categorize multisyllabic utterances? *Developmental Psychology* 29:711–21.

BROWMAN, CATHERINE P., and LOUIS GOLDSTEIN. 1989. Articulatory gestures as phonological units. *Phonology* 6:201–51.

CUMMING, RUTH. 2008. Should rhythm metrics take account of fundamental frequency? *Cambridge Occasional Papers Linguistics* 4:1–16.

CUMMINS, FRED. 2002. Speech rhythm and rhythmic taxonomy. In *Proceedings of Speech Prosody 2002*:121–24. Aix-en-Provence, France, April 11–13.

DAUER, REBECCA M. 1983. Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics* 11:51–62.

DELLWO, VOLKER. 2006. Rhythm and speech rate: A variation coefficient for Δ*C*. In *Language and Language-processing*, ed. by Pawel Karnowski and Imre Syigeti, 231–41. Frankfurt am Main: Peter Lang.

DELLWO, VOLKER; ADRIAN FOURCIN; and EVELYN ABBERTON. 2007. Rhythmical classification of languages based on voice parameters. In *Proceedings of the 16th International Congress of Phonetic Sciences*, ed. by J. Trouvain and W.J. Barry, 1129–32. Saarbrücken.

DELLWO, VOLKER, and PETRA WAGNER. 2003. Relations between language rhythm and speech rate. In *Proceedings of the 15th International Congress of Phonetic Sciences,* ed. by D. Recasens, M. J. Solé, and J. Romero, 471–74. Barcelona.

DONEGAN, PATRICIA, and DAVID STAMPE. 2004. Rhythm and the synthetic drift of Munda. In *Yearbook of South Asian languages and linguistics*, ed. by Rajendra Singh, 3–36. Berlin, New York: Mouton de Gruyter.

FANT, GUNNAR; ANITA KRUCKENBERG; and LENNART NORD. 1991. Durational correlates of stress in Swedish, French and English. *Journal of Phonetics* 19:351–65.

FLEGE, JAMES EMIL. 1988. Effects of speaking rate on tongue position and velocity of movement in vowel production. *Journal of the Acoustical Society of America* 84(3):901–16.

FOWLER, CAROL A. 1979. "Perceptual centers" in speech production and perception. *Perception & Psychophysics* 25(5):375–88.

FOWLER, CAROL A. 1983. Converging sources of evidence on spoken and perceived rhythms of speech: Cyclic production of vowels in monosyllabic stress feet. *Journal of Experimental Psychology: General* 112(3):386–412.

GALVES, ANTONIO; JESUS GARCIA; DENISE DUARTE; and CHARLOTTE GALVES. 2002. Sonority as a basis for rhythmic class discrimination. In *Proceedings of Speech Prosody 2002,* Aix-en-Provence, France, April 11–13.

GRABE, ESTHER, and EE LING LOW. 2002. Durational variability in speech and the Rhythm Class Hypothesis. In *Laboratory Phonology 7*, ed. by Carlos Gussenhoven and Natasha Warner, 515–46. Berlin, New York: Mouton de Gruyter.

GRABE, ESTHER. 2002. Variation adds to prosodic typology. In *Proceedings of Speech Prosody 2002* Aix-en-Provence, France April 11–13.

HIRATA, YUKARI. 2004. Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics* 32(4):565–89.

KLATT, DENNIS. 1976. Linguistic uses of segmental duration in English: Acoustic and perceptual evidence. *Journal of the Acoustical Society of America* 59(5):1208–21.

KOZASA, TOMOKO. 2005. An acoustic and perceptual investigation of long vowels in Japanese and Pohnpeian. University of Hawai'i at Mānoa PhD dissertation.

LEHISTE, ILSE. 1976. Influence of fundamental frequency patterns on the perception of duration. *Journal of Phonetics* 4:113–17.

LEHISTE, ILSE. 1977. Isochrony reconsidered. *Journal of Phonetics* 5:253–63.

LEVELT, CLARA C., AND RUBEN VAN DE VIJVER. 2004. Syllable types in crosslinguistic and developmental grammars. In *Fixing priorities: Constraints in phonological acquisition*, ed. by René Kager, Joe Pater, and Wim Zonneveld, 204–18. Cambridge: Cambridge University Press.

LIPPUS, PÄRTEL; KARL PAJUSALU; and JÜRI ALLIK. 2009. The tonal component of Estonian quantity in native and non-native perception. *Journal of Phonetics* 37:388–96.

LOW, EE LING; ESTHER GRABE; and FRANCIS NOLAN. 2000. Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech* 43(4):377–401.

MORTON, JOHN; STEVE MARCUS, and CLIVE FRANKISH. 1976. Theoretical note: Perceptual centers (P-centers). *Psychological Review* 83(5):405–8.

NAZZI, THIERRY; JOSIANE BERTONCINI; and JACQUES MEHLER. 1998. Language discrimination by newborns: Toward understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance* 24(3):756–66.

NAZZI, THIERRY, and FRANCK RAMUS. 2003. Perception and acquisition of linguistic rhythm by infants. *Speech Communication* 41(1):233–43.

NOLAN, FRANCIS, and EVA LIINA ASU. 2009. The pairwise variability index and coexisting rhythms in language. *Phonetica* 66:46–63.

PAMIES BERTRÁN, ANTONIO. 1999. Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages. *Language Design* 2:103–30.

PARKER JONES, 'ŌIWI. 2006. Durational variability and stress-timing in Hawaiian. In *Proceedings of the 11th Australian International Conference on Speech Science and Technology*, ed. by Paul Warren and C. I. Watson. Auckland, New Zealand (CD-ROM).

PIKE, KENNETH L. 1945. *The intonation of American English*. Ann Arbor: University of Michigan Press.

RAMUS, FRANCK. 2002. Acoustic correlates of linguistic rhythm: Perspectives. In *Proceedings of Speech Prosody 2002,* Aix-en-Provence, France, April 11–13.

RAMUS, FRANCK, and JACQUES MEHLER. 1999. Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America* 105(1):512–21.

RAMUS, FRANCK; MARINA NESPOR; and JACQUES MEHLER. 1999. Correlates of linguistic rhythm in the speech signal. *Cognition* 73:265–92.

RAMUS, FRANCK; EMMANUEL DUPOUX; and JACQUES MEHLER. 2003. The psychological reality of rhythm classes: Perceptual studies. In *Proceedings of the 15th International Congress of Phonetic Sciences*, Barcelona, 337–42.

ROACH, PETER. 1982. On the distinction between "stress-timed" and "syllable-timed" languages. In *Linguistic controversies*, ed. by David Crystal. 73–79. Bungay, Suffolk: The Chaucer Press Ltd.

ROUAS, JEAN-LUC. 2007. Automatic prosodic variations modeling for language and dialect discrimination, *IEEE Transactions of Audio, Speech Language Processing* 15(6):1904–11.

STOJANOVIC, DIANA. 2008. Impact of segmentation rules on the rhythm metrics. Poster presented at the Acoustical Society meeting. Miami. November 17–21.

STOJANOVIC, DIANA. 2009. Modeling segmentation precision and inter-segmenter variability. Poster presented at the Acoustical Society meeting, San Antonio. October 26–30.

TULLER, BETTY, and CAROL A. FOWLER. 1980. Some articulatory correlates of perceptual isochrony. *Perception & Psychophysics* 27(4):277–83.

TURK, ALICE; SATSUKI NAKAI; and MARIKO SUGAHARA. 2006. Acoustic segment durations in prosodic research: A practical guide. In *Methods in empirical prosody research*, ed. by Stefan Sudhoff, Denisa Lenertová, Roland Meyer, Sandra Pappert, Petra Augurzky, Ina Mleinek, Nicole Richter and Johannes Schließer, 1–28. Berlin, New York: De Gruyter.

TURK, ALICE E., and STEPHANIE SHATTUCK-HUFNAGEL. 2007. Multiple targets of phrase-final lengthening in American English words. *Journal of Phonetics* 35:445–72.

UMEDA, NORIKO. 1977. Consonant duration in American English. *Journal of the Acoustical Society of America* 61(3):846–58.

WHITE, LAURENCE, and SVEN L. MATTYS. 2007. Calibrating rhythm: First language and second language studies. *Journal of Phonetics* 35:501–22.

stojanov@hawaii.edu