Psychomusicology: Music, Mind, and Brain
2017, Vol. 27, No. 2, 95–112

0275-3987/17/$12.00 http://dx.doi.org/10.1037/pmu0000175

A Musical Model of Speech Rhythm

Steven Brown
McMaster University

Peter Q. Pfordresher
University at Buffalo, State University of New York

Ivan Chow
McMaster University

Research on speech rhythm has been notoriously oblivious to describing actual rhythms in speech. We present here a model of speech rhythm at the sentence level inspired by musical conceptions of meter. We posit that speech is underlain by a basic metricality. However, instead of arguing that speech is isochronous, we propose that utterances can have internal changes of meter, making them "heterometric." In addition, we see 2 rhythmic devices for obviating the need for meter changes within utterances and thus maintaining the stability of the rhythm. Both of them involve subdivisions of component beats into subbeats: 1) subdivisions into 2's and 3's, resulting in duplets and triplets, respectively; and 2) subdivisions according to complex ratios, resulting in polyrhythms. We tested the model acoustically by having a group of 14 participants read unfamiliar sentences aloud and examining the extent to which their timing conformed with the predictions of a priori rhythmic transcriptions of the sentences. The observed patterns of variability in speech timing for these sentences, when measured at the bar level of the transcription, were generally consistent with the musical model.

Keywords: speech, rhythm, meter, timing, music

> ". . . iambic [is] the verse-form closest to speech. There is evidence of this: we speak iambics in conversation with each other very often. . . ."
>
> Aristotle in *Poetics*

Much work on speech rhythm has been driven far more by a desire to classify languages into categories than by the need to elucidate the actual rhythms of spoken utterances. Common approaches to speech rhythm focus, for example, on the variability of syllabic durations within utterances (Grabe & Low, 2002) or the proportion of an utterance's duration that is occupied by vowels (Ramus, Nespor, & Mehler, 1999). But these features do not specify actual rhythms—that is, the temporal patterns of syllable onsets within an utterance—and instead reduce whole languages to descriptive statistics. Knowing that English is 40% vocalic (Ramus et al., 1999) indicates little about the timing of syllable onsets within any given English utterance, even though this information may be useful in differentiating English taxonomically from languages having different types of syllable structure.

Outside of linguistics, though, representations of sentence rhythms are commonplace, and it is unclear why such representations have not had a larger impact on linguistic theories. Poetic verse, song, Shakespearean dialogue, and rap are all based on musical notions of the periodicity of syllable onsets. Consider the rhythmic transcription of the text of the children's song *Twinkle Twinkle* shown in Figure 1a. The rhythm is organized as a two-beat cycle alternating between strong and weak beats. The relative onset-time and relative duration of every syllable in the sentence is specified, hence making this a true representation of a rhythm. Next, the stressed syllables of the disyllabic words fall on the strong beats of the two-beat cycle (i.e., the downbeats), whereas the unstressed syllables fall on the weak beats. Finally, we see that even silence is specified in this transcription in the form of the rest that sits in between "star" and "How," in this case indicating a sentence break.

Regardless of the fact that *Twinkle Twinkle* is a poetic form of speech, its transcription effectively captures the basic elements of what a model of speech rhythm should describe: (a) it specifies a unit of rhythm, in this case the two-beat metrical units that make up each measure of the transcription; (b) it specifies the relative onset-time and relative duration of every syllable in the sentence; and (c) it represents not only the duration but the weight (i.e.,

This article was published Online First May 11, 2017.

Steven Brown, Department of Psychology, Neuroscience & Behaviour, McMaster University; Peter Q. Pfordresher, Department of Psychology, University at Buffalo, State University of New York; Ivan Chow, Department of Psychology, Neuroscience & Behaviour, McMaster University

This work was funded by a grant from the Natural Sciences and Engineering Research Council (NSERC) of Canada to SB and by National Science Foundation Grant BCS-1256964 to P. Q. P. We thank Kyle Weishaar for assistance in data collection, interpretation, and analysis. We thank Stephen Handel for helpful discussion of the concepts and methods covered in this paper.

An early stage of this analysis was presented in poster form at the Speech Prosody conference in 2010 and published as a conference proceeding as: Brown, S., & Weishaar, K. (2010). Speech is "heterometric": The changing rhythms of speech. *Speech Prosody 2010* 100074: 1–4.

Correspondence concerning this article should be addressed to Steven Brown, Department of Psychology, Neuroscience & Behaviour, McMaster University, 1280 Main Street West, Hamilton, ON, Canada, L8S 4K1. E-mail: stebro@mcmaster.ca

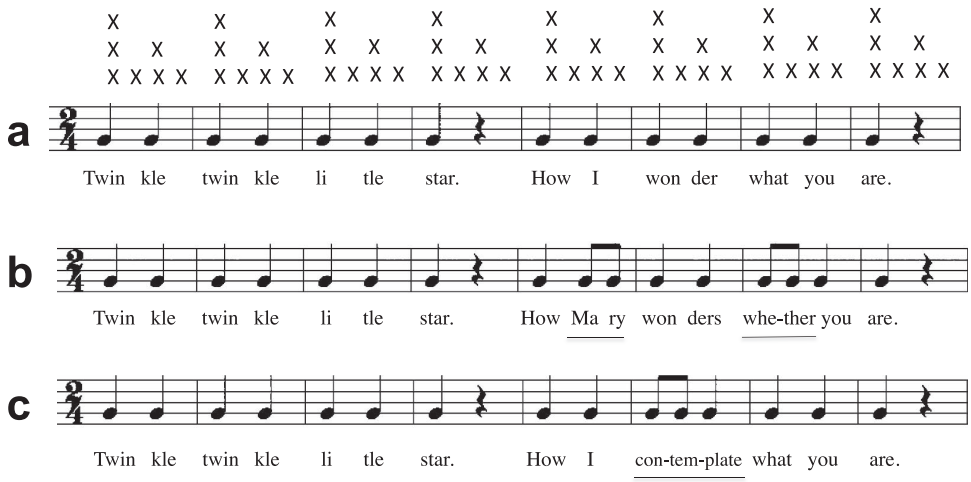

Figure 1. Musical transcription and metrical grid for the sentence tagged "Twinkle". (a) The original version of the text. (b) A version in which two monosyllabic words are converted into trochees (underlined), accompanied by a reduction of the individual quarter notes into duplets of eighth notes. (c) A version in which the dactyl "contemplates" (underlined) replaces the trochee "wonders", accompanied by a reduction of the first quarter note into a duplet of eighth notes.

stress) of each syllable in the sentence, such that prominent syllables fall on strong beats. Each of these three elements has been analyzed in isolation in various models of speech rhythm, but they have rarely been synthesized into a unified model. These three elements have been analyzed, respectively, in isochrony models, rhythm metrics, and metrical phonology. We briefly review these three traditions in phonology before mentioning the only integrated account that we know of, namely Joshua Steele's 1775, treatise *An Essay Toward Establishing the Melody and Measure of Speech to be Expressed and Perpetuated by Peculiar Symbols.* In our study, we report a test of a critical prediction of a musical model of speech, namely, that the production of time intervals between stressed syllables (here called "prominence groups") is based on a music-like representation of metrical structure. In particular, the "meter" of speech can serve to stabilize the timing of prominence groups when the timing of individual syllables varies. At the same time, speech (like music) can feature changes in meter that lead to commensurate changes in the timing of prominence groups.

Isochrony Models

The first issue for speech rhythm relates to specifying a unit of rhythm. Lloyd James (1940, quoted in Pike, 1945) contrasted languages having a rhythm similar to a machine gun with those having a rhythm similar to Morse code. Pike (1945) classified such languages as syllable-timed and stress-timed, respectively, a categorization that is often referred to as the "rhythm class hypothesis" (Abercrombie, 1967; Grabe & Low, 2002). A syllable-timed language is one in which there is equal duration between syllable onsets (in the limiting case, 1/4 time in music), whereas a stress-timed language is one in which there is equal duration between stressed syllables (in the limiting case, 2/4 time in music). A third category of language, namely, mora-timed, was later proposed to account for languages such as Japanese and Tamil (Port, Dalby, & O'Dell, 1987). Suffice it to say that tests of the rhythm class hypothesis have required that a unit of isochrony be found at some level of an utterance and that a failure to find such a unit is evidence against the existence of metrical organization in speech. In reality, many studies have failed to find such isochrony, and this has challenged the whole notion of periodicity and rhythm in speech (Bertran, 1999; Dauer, 1983; Lehiste, 1977; Ramus et al., 1999), or has instead suggested that this phenomenon might be restricted to perception alone, rather than production mechanisms (Nolan & Jeon, 2014; Patel, 2008).

One problem with the rhythm class hypothesis and with the studies that seek to test it is that they require that speech rhythms be isochronous, whereas they give little consideration to *metrical structure*, in other words a regularity of beats and the possibility of subbeats nested within them. Indeed, while a syllable-timed rhythm can be thought of as a one-beat meter (i.e., 1/4 time in music), a stress-timed rhythm can take on a multiplicity of forms, just as is seen with the variety of meter types found in music. The simplest structure is a 2-beat meter, with an alternation between strong and weak beats. However, beats do not necessarily map onto syllables. The phrase "big for a duck" that has been used in speech cycling experiments (Cummins & Port, 1998) can be modeled as a 2-beat cycle (i.e., BIG for a DUCK), but as one in which the two syllables of "for a" occupy one beat rather than two, due to a halving of their duration values. There are far more complex means of creating stress-timed rhythms in speech than that, and so the observation of stress timing per se—even when it can be reliably observed—does not offer a specification of the metrical structure of an utterance.

Implicit in the contrast between syllable timing and stress timing is whether a language has subbeats or not (as mentioned with regard to "big for a duck" above), an issue associated with the durational variability of syllables, as discussed in models of rhythm metrics (see below). This is related to the notion of a

metrical hierarchy in music (Lerdahl & Jackendoff, 1983). Languages classified as stress-timed have a greater variability of syllabic durations than languages classified as syllable-timed (Grabe & Low, 2002), due to mechanisms related to vowel reduction and consonant clustering (Dauer, 1983), among others. Another way of saying this is that languages classified as stressed-timed seem to have a greater number of syllabic durations than languages classified as syllable-timed. Looking back to *Twinkle Twinkle* (see Figure 1), the phrase "how I wonder what you are" is a clear example of syllable timing since there is only a single duration-value for all the syllables; in other words, the phrase is isodurational. But a small change of the phrase to "how *Mary* wonders *whether* you are" (Figure 1b) divides the beats for "I" and "what" into trochees whose syllables contain half the duration of the original words (just as is seen in "big for a duck"). Hence, the modified version contains two syllabic durations, compared to the original isodurational text. Languages whose rhythms readily lend themselves to creating a hierarchical arrangement of beats and subbeats in this manner are far more likely to be classified as stress-timed than languages that restrict this. Quantifying this variability of syllabic durations using the descriptive statistics of rhythm metrics (described in the next section) can be useful in classifying languages, but it tells us nothing at all about the actual rhythm of any given utterance within a language or the processes of subbeat formation that diversify the syllabic durations within utterances. In other words, rhythm metrics do not elucidate the utterance-level processes that apportion relative duration-values to the syllables within a sentence. As O'Dell and Nieminen (1999, p. 1075) noted: "Mathematical formulas estimated from empirical data do not explain anything by themselves, they are just a means of categorizing languages."

The tendency of speech to have not only a metrical structure but also subdivisions of beats is supported by oscillator coupling models, another development within the tradition of isochrony-based research in speech rhythm (Cummins & Port, 1998; O'Dell & Nieminen, 1999; Port, 2003; Tilsen, 2009). Each unit in the phonological hierarchy (e.g., mora, syllable, foot, and stress group) is considered to have its own time scale and thus its own rhythmic oscillator. Tilsen's (2009) multitimescale dynamical model proposed that these multiple time scales are integrated and synchronized to form the rhythmic pattern of speech. Evidence for these models has come from work on repetitive speech entrained to a metronome (Cummins & Port, 1998; Tilsen, 2009), which examines the rhythmic patterns that show the greatest stability, using simple phrases like "big for a duck." Such studies have shown that the stressed syllables of the uttered phrases occur at predictable phases of the metronome cycle, and that such phasing conforms with a "harmonic timing effect" whereby the points of greatest stability occur as integer ratios of the metronome frequency (i.e., 1:2, 1:3). The major implication of such experiments is that "[speech] rhythm is hierarchical, and that elements low in the hierarchy will nest an integral number of times within higher elements" (Cummins & Port, 1998, p. 147), an idea formally similar to the notion of subbeats in music's metrical hierarchy. However, it needs to be pointed out that the use of a metronome in these studies begs the question of whether spontaneous speech in fact contains these rhythms, which is why the present study uses a self-paced paradigm to examine speech rhythm.

Rhythm Metrics

An important criterion for a theory of speech rhythm is that it should specify the relative durations of all the syllables that comprise an utterance. Very little work in phonology has analyzed syllabic durations. One field that has done so is rhythm metrics, which has devoted itself to providing a quantitative test of the rhythm class hypothesis, with the same emphasis on taxonomic classification of languages. However, instead of analyzing the local rhythmic properties of utterances, rhythm metrics has focused on descriptive statistics of utterances as a whole (Grabe & Low, 2002; Ramus et al., 1999). The principal one has been nPVI (normalized pairwise variability index), which is a measure of the pairwise durational variability of vocalic intervals, but which corrects for the mean duration of each intervocalic interval.

There has been much discussion in the literature about the merits of these rhythmic parameters for classifying languages (Arvaniti, 2009, 2012; White & Mattys, 2007). From our standpoint, the key criticism is that these durational measurements do not provide information about the relative duration of syllables in an utterance within a regular metrical framework. Although these statistics may indeed reflect the rhythmic properties of a language, they are not able to specify the actual rhythm of any given utterance within it. It is worth pointing out that a musical transcription of a sentence, such as that for *Twinkle, Twinkle* in Figure 1 or *Humpty Dumpty* (presented in Figure 5 in the Results section), is able to provide information about durational variability, along similar lines to nPVI (Patel & Daniele, 2003; Patel, Iversen, & Rosenberg, 2006). For example, *Twinkle, Twinkle* is made up exclusively of a single duration-value (i.e., quarter notes in the transcription) and hence shows no durational variability. By contrast, *Humpty Dumpty* is made up of two duration-values (half notes and quarter notes; see Figure 5 for a transcription). The transcription therefore provides information about the variability of syllabic durations within the sentence while at the same time specifying the actual duration-value of each syllable.

Metrical Phonology

Metrical phonology presents a theory of the hierarchical organization of syllable weights within words and higher-level units (Hayes, 1983; Kiparsky, 1977; Liberman & Prince, 1977), as represented through both metrical trees and metrical grids (Goldsmith, 1990). The basic unit of rhythm in this model is the "foot." The two standard disyllabic feet are the trochee (initial stress) and the iamb (final stress). Words and utterances are built up of feet, exactly as is seen in models of poetic meter (Caplan, 2007; Fabb & Halle, 2008). Metrical phonology has offered a rich set of cross-linguistic principles for predicting how stress patterns emerge across the syllables of words and utterances (Hammond, 1995; Nespor & Vogel, 1986). However, its main weakness from our standpoint is that it says nothing about the *relative duration* of syllables at any level of metrical structure, which is a key consideration for the conception of a rhythm. The theory implicitly assumes that all timing units (basically syllables) have equal duration. However, as we alluded to above in our discussion of subbeats, this cannot be the case. Consider again the phrase "how I wonder what you are" from *Twinkle Twinkle*. These syllables would typically be spoken isochronously such that each syllable had the same duration. But now consider a change to "how I

contemplate what you are" (Figure 1c). No native speaker of English would utter the three syllables of "contemplate" with three equal beats, which would sound robotic. They would instead speak the first two syllables as subbeats with roughly half the duration of the third syllable (i.e., a duplet of eighth notes in the musical notation). A theory of speech rhythm requires a model of not just the relative strength but also the relative duration of the syllables in an utterance.

The only integrated account of speech rhythm that we know of is found in the first major treatise on English intonation (Kassler, 2005), preceding Pike's and Abercrombie's proposals of isochrony by nearly two centuries. It is Joshua Steele's *An Essay Toward Establishing the Melody and Measure of Speech to be Expressed and Perpetuated by Peculiar Symbols*, published in 1775 (see Rush [1827/2005] for an acknowledgment of its influence). Steele laid out a detailed musical model of both the melody and rhythm of speech, although we will only concern ourselves with the rhythmic concepts here. He recognized a basic metricality to spoken English, with a preference for 2-beat and 3-beat meters, much as is seen in contemporary oscillator-coupling models (Port, 2003). In addition, he recognized that speech rhythm was based on variations in both the weight and duration of syllables, hence establishing contrasts between strong and weak beats and between long and short beats, respectively. Modern-day metrical phonology provides a detailed theory of syllable weight, but no contemporary approach to speech rhythm in linguistics provides a model of syllabic duration.

The Present Study

The primary objective of the present study is to build upon the prescient but long-forgotten work of Joshua Steele and attempt to reinvigorate the discussion of speech rhythm toward a consideration of the temporal patterning of syllable onsets and durations. We present here a musical analysis of speech rhythm that examines not only the relative prominence of syllables within an utterance (typical of metrical phonology) but the relative duration of syllables as well. Our analytical method is to create an intuitive rhythmic representation of a sentence using musical transcription and to test its rhythmic predictions quantitatively against the acoustic productions of a group of native speakers unfamiliar with the sentence. Within this framework, musical notation serves as a model for speech rhythm.

There is a distinction in music between rhythm and meter (Dowling & Harwood, 1986) that may be applicable to speech rhythms. Whereas *rhythm* refers to a surface pattern of onset times—which in speech may be formed by timespans between syllable onsets—*meter* refers to an abstract temporal framework that helps to structure the production and perception of a rhythm. Meter is based on an inferred pattern of alternating strong and weak accents. Critically, whereas rhythms are typically variable, meter is typically more consistent and stable. We propose that the failure to identify "rhythms" in speech may reflect the failure to apply this music-related distinction. In the present study, we focus on meter, which we consider to be the most critical development of the present model compared to previous work. We place an emphasis not on individual syllables, but on the "bar" level of metrical structure shown in the transcriptions, where we refer to these bars as "prominence groups" (PG's). Subsequent studies will focus on the constituent rhythms (i.e., the variable syllabic level of the transcription). Our notion of a prominence group is similar to the concept of an "inter-stress interval" found in previous research on speech rhythm (Cummins & Port, 1998; Dauer, 1983; Fant, Kruckenberg, & Nord, 1991; Kim & Cole, 2005; Tilsen, 2009).

Although meter is assumed to remain stable across a musical work, occasional changes to meter do occur in music, although much less often than changes to rhythm. Meter change is thus another feature of a musical model that may be well suited to the complexity of speech timing. As such, the present study included sentences predicted to reflect a stable meter—with or without rhythmic variability—as well as sentences with a single internal change in meter, something that we refer to as *heterometric* sentences. We analyzed the timing of the PG's in order to determine whether their variability reflected the kind of stability (or lack thereof) predicted by notated transcriptions of metrical structure.

An important assumption in models of musical timing is that the duration of a measure is stable even when there is variability in the durations of notes. This is how meter functions as a kind of mental frame for the expression and perception of rhythm (cf. Palmer & Krumhansl, 1990). Consider examples that were discussed previously. Most measures in *Twinkle Twinkle* (Figure 1a) comprise two quarter notes, and so it would not be surprising if all measures were produced with the same timing, leading to low variability. However, based on the assumptions of musical meter, the aforementioned variation, "How I contemplate what you are" (Figure 1c), would lead to variability in the rhythmic patterning of syllables (as seen in the notation), and yet the meter would remain consistent. An analysis of timing at the level of meter should be just as consistent for this sentence as for the original version of *Twinkle Twinkle.* As such, our model predicts that the stability of metrical timing should be unaffected by variability in the number of syllables (notes) that are contained in different measures, a parameter that we refer to as "syllable density."

In the present study, we analyzed a music-like metrical framework of speech rhythm against the alternative hypothesis that the timing of PG's should vary as a function of the number of syllables in each measure, in other words the syllable density. Consider, for instance, the possibility that speech rhythms are simply perceptual constructions that are not rooted in actual production (Patel, 2008). If so, then the duration of syllables on average will approximate equality because their variability should just reflect noise in the motor signal or differences in speech articulation that may be only incidentally related to metrical stress. In this case, the duration of PG's would simply reflect how many syllables there are in the measure (i.e., the syllable density), and the utterance-level variability would reflect differences in syllable density across successive measures. Some previous research suggests that speech timing is influenced both by the constraining influence of the metrical foot and by the number of phonemes/syllables within a foot (Fant et al., 1991; Kim & Cole, 2005). However, because such studies have no sentence-level analyses (i.e., feet are dissociated from their sentence context), one cannot draw conclusions about the influence of metrical feet on timing stability across a sentence as a whole, which is a principal goal of the current study's approach to speech rhythm.

Our analyses are based on two sentence-level measures. First, we analyzed variability across PG's in a sentence using the coefficient of variation (CV), which is a standardized measure of

variability. According to the predictions of our model, CV should be influenced by changes in the metrical frame, and be higher for heterometric than isometric sentences, but should not be influenced by the number of syllables per sentence otherwise. Next, we analyzed PG timing in a way that focused on whether the frequency-ratios formed by different meters in a heterometric sentence are borne out in production. We analyzed data both by grouping sentences based on their metrical and rhythmic structure and by examining individual sentences descriptively as well as through regression analysis.

Method

Participants

Fourteen native speakers of Canadian English (12 females, M = 21.9 years, *SD* = 1.2 years) participated. They were recruited from an introductory psychology testing pool, and received course credit for their participation. Upon arrival to the lab, participants filled out questionnaires about their linguistic and musical backgrounds, including any second languages spoken and their level of musical training. Eleven of the 14 participants had some experience with a second language. Nine of the 14 participants had some form of musical training. All participants reported normal hearing.

Stimuli

A sample of nine sentences was generated; all are shown in Table 1. Three of them consisted of *isodurational* sentences for which the notated transcriptions yielded a single duration-value throughout the sentence. We have opted to use the word "isodurational" instead of "isochronous" in describing these sentences since all of them have stress patterns, either in 2/4 or 3/4 time. We wanted to avoid any confusion with definitions of isochrony that require that all elements have identical stress (1/4 meter), such as in the case of a metronome beat. Next, four of the sentences consisted of *isometric* sentences that had a constant meter (either 2/4 or 3/4), but that contained more than one duration-value per sentence, as well as variable numbers of syllables across the measures. Among these four sentences, two of them were isometric counterparts to *heterometric* sentences that contained meter changes within the sentence (either 2/4 to 3/4 or 3/4 to 2/4). Among the isometric/heterometric pairs, one varied focus between two different words in the sentence (TWO yellow shirts vs. two YELLOW shirts) and the other pair contrasted a compound noun (greenhouse) with the associated adjectival phrase (green house). For these four sentences, the emphasized element was written in capital letters when presented to participants (i.e., GREENhouse vs. green HOUSE). Participants were presented with the sentences in standard written format. No rhythmic cues of any kind were used. With the exception of two nursery rhymes (*Twinkle Twinkle* and *Humpty Dumpty*), all sentences were novel and were generated for the experiment, with transcriptions created by the first author.

Procedure

After filling out questionnaires in a testing room, participants were presented with a sheet containing the nine stimulus sentences. They were allowed to practice speaking them aloud a few times for familiarization purposes. The experimenters did not provide cues on how to read the sentences or any of the words within them. They only provided general feedback if participants were speaking too quietly or in a creaky voice, both of which would have affected the acoustic signal we recorded. After this practice phase, the participant moved into a sound booth. Recordings were made using an Apex 181 USB condenser table-mounted microphone. Stimulus sentences were presented to participants using Presentation software (Version 0.70, Neurobehavioral Systems, Berkeley, CA). Participants' responses were recorded using Adobe Audition (Adobe Systems, San Jose, CA) at a 44.1 kHz sampling rate.

The experiment began with a warm-up phase. This consisted of the following tasks: simple conversational speech (e.g., what the participant ate for breakfast that morning); reading of the standard "Rainbow" passage; several coughs; several throat clears; and vocal sweeps up and down the vocal range to obtain the participant's highest and lowest pitches, respectively. Next, *Hickory Dickory Dock* was read aloud by the participant so as to familiarize him or her with the presentation software as well as to allow us to adjust the microphone gain for that participant. This sentence was not analyzed.

Table 1
Stimulus Sentences by Sentence-Timing Category

ISODURATIONAL sentences:
1. *Twinkle.* Twinkle twinkle little star. How I wonder what you are. (2/4)
2. *Balcony.* The balcony facing the Jamison Building was painted with beautiful colors. (3/4)
3. *Mary.* Mary purchased purple flowers Monday morning every week. (2/4)

ISOMETRIC sentences:
4. *Humpty.* Humpty dumpty sat on a wall. Humpty dumpty had a great fall. All the king's horses and all the king's men couldn't put Humpty together again. (3/4)
5. *Pamela.* Pamela purchased beautiful flowers Saturday morning all through the year. (2/4 with 3-against-2 polyrhythms)
6. *Yellow.* Miguel bought two YELLOW shirts at the men's store by the bay. (3/4)
7. *Greenhouse.* Nathaniel writes novels and lives in a GREENhouse built by a farmer. (3/4)

HETEROMETRIC sentences:
8. *Two.* Miguel bought TWO yellow shirts at the men's store by the bay. (2/4 changing to 3/4)
9. *House.* Nathaniel writes novels and lives in a green HOUSE built by a farmer. (3/4 changing to 2/4)

Note: The italicized words display the "tags" used as brief titles for each sentence. The sentences are organized into three sentence-timing categories: isodurational, isometric, and heterometric. After each sentence is its predicted meter, where sentences 8 and 9 are predicted to have internal meter changes. Arrows are used to indicate pairings between sentences that either vary in focus-word (sentences 6 and 8) or that create a contrast between a compound noun and the associated adjectival phrase (sentences 7 and 9).

Participants were then presented with the nine stimulus sentences in random sequence—one at a time—on a computer screen and were asked to read them in an emotionally neutral, conversational voice. Each sentence was displayed on the screen for 10 s during a rehearsal period so that the participant could practice saying it out loud. The participant was then given 15 s to record the utterance fluently twice without error. The second rendition was analyzed. In the event of a speech error, the participant was instructed to repeat the sentence in its entirety. The 14 participants provided nine recordings each, resulting in 126 sentence-samples for analysis. Note that there was no metronome beat or any other entrainment cue in the experiment.

Rhythmic Transcriptions

Each of the nine sentences used in this study was designed to highlight a particular rhythmic principle, as shown in a musical transcription. The major objective of the study was to determine if a group of native speakers would produce renditions of these sentences that conformed with the rhythmic predictions of the a priori transcriptions. The transcriptions were generated by the first author prior to any data collection or analysis. Each sentence was designed to convey a different metrical principle, including 2/4 and 3/4 meter. In the transcriptions presented in the figures below (as in Figure 1 discussed in the Introduction), beats are represented by quarter notes; subbeats are represented by eighth notes for simple divisions or by quarter-note triplets for more-complex divisions. A single arbitrary pitch-level on a clef-less staff is used throughout the transcriptions, since we are only concerned with rhythm in these analyses and not pitch.

The rhythmic transcriptions segmented sentences into a series of stress groups, or what we shall refer to as "prominence groups" (PG), akin to measures of music. We use the term "prominence group" rather the "stress group" in order to accommodate languages such as Cantonese that have no word-level stress but that instead have points of prominence at the sentence level (Chow, Belyk, Tran, & Brown, 2015). This is formally analogous to the rhythmic units proposed in isochrony models of speech rhythm, although our groupings need not be isochronous throughout a sentence (see below). What is common among all of these concepts for both speech and music is that these groups represent interstress intervals (Dauer, 1983). The term "foot" from poetry and metrical phonology requires that the material consist of polysyllabic words. Hence, in the verse "Humpty, Dumpty, sat on a wall" (which is transcribed as three prominence groups in 3/4 meter in Figure 5), "Humpty" and "Dumpty" represent trochaic feet, but the monosyllabic words "sat," "on," "a," and "wall" do not have a true status in foot terminology. However, Nolan and Asu (2009) have applied the foot concept to mean essentially the same thing as an interstress interval in their analyses. Next, a PG differs from an "accentual phrase" (Jun & Fougeron, 2002) in that an accentual phrase can start on an unstressed syllable that leads to the primary stress of a phrase. In other words, it can start on a musical upbeat, whereas a PG can only ever start on a musical downbeat.

By definition, each PG starts with a strong beat, that is, a downbeat, implying a stressed syllable. Unstressed elements—including function words (such as articles and prepositions) or the unstressed syllables of polysyllabic words—should never initiate a PG. For example, in the phrase "the mouse ran up the clock" from the nursery rhyme *Hickory Dickory Dock*, the content words "mouse" and "clock" fall on downbeats, whereas the function word "the" never would. Musical transcriptions of speech rhythm—such as is routinely seen in children's songs—very often break up syntactic units such as noun phrases (e.g., "the mouse") and place them into different rhythmic groups. Moreover, rhythmic groupings may even break up individual words, as is seen below in the sentences containing the names "Miguel" and "Nathaniel" having noninitial stress, where the PG's start with the stressed syllables of "-guel" and "-than," respectively. Finally, because this is a bar-level analysis, each PG extends to the downbeat of the next measure of the transcription.

Analysis of Production

The basic measurement that we derived from the speech signal was the duration of each PG for each sentence, where segmentation and time measurement were done using Praat (Boersma & Weenink, 2014). A critical concern for the segmentation of sentences into PG's relates to the point in the starting syllable of a PG at which the segmentation should occur, the so-called perceptual center or P-center (Pompino-Marschall, 1989; Port, 2003). We validated our segmentation technique using the nursery rhymes, based on the assumption they should be timed in a metrical manner. We examined a host of possibilities for segmentation—including the syllable onset, vowel onset, and the intensity peak of the first vowel—and found that using the point of sonority (voicing) onset as the measurement point, whether of a vowel or a sonorant consonant (nasal, liquid, or glide), provided PG measurements that conformed most strongly to a meter. It is important to note that using sonority onsets in no way biases the analysis of any of the other sentences toward metricality.

The first step in our data analysis was to normalize PG's based on the mean for each utterance (for each participant) so that all PG's could be displayed in a way that reflects relative timing. The distribution of normalized PG values across participants is displayed in boxplots above each notated sentence in Figures 2–8. The major prediction for the study is that isometric sentences should have PG's that are equal throughout, that is, each group should have a normalized mean value of 1.0. For a heterometric sentence that changes in meter from 2/4 to 3/4, the 3/4 groups are predicted to have 1.5 times the duration of the 2/4 groups. Likewise, for a sentence that changes meter from 3/4 to 2/4, the duration of the 2/4 groups is predicted to be 0.67 times that of the 3/4 groups.

We conducted three statistical analyses of these scores. The most basic one involved calculating the variability of PG's across different types of sentences. For normalized PG's, the standard deviation is equivalent to the coefficient of variation (CV), which is defined as the ratio of the standard deviation (SD) to the mean (M). It is a standardized measure of variability that is motivated by the psychophysics of timing. In general, timing variability increases for slower tempos (Wing & Kristofferson, 1973). Because we are interested in timing variability that is independent of speaking rate, CV (the standard deviation of normalized PG's) is an appropriate way to control for such spurious timing variability. Thus, high CV values indicate more-variable timing that is independent of speaking rate. CV's were computed separately for each

spoken utterance (i.e., each participant and sentence) based on the sequence of PG's. We derived a single value of CV for each production of each sentence, reflecting the variability of production across PG's within a single utterance. Because normalized scores already standardize PG's based on the mean per utterance (i.e., the mean is always 1), the standard deviation of the normalized scores is equivalent to the CV of the original measured PG's. Based on preliminary analyses, we removed from consideration PG's from the two nursery rhymes that marked a phrase boundary (PG 4 in both cases), since these boundaries were associated with terminal lengthening of that PG. The mean CV across participants for each sentence is shown in Table 2 in the column labeled "CV PG." Appendix presents illustrative examples of how CV's were computed for individual productions.

The second statistical analysis involved comparisons of selected PG's that might be produced with longer or shorter durations based on properties of the notation. Because the isometric and heterometric sentences both have variable numbers of syllables per PG (unlike the isodurational sentences, which always have identical numbers of syllables per PG), this allowed us to contrast the model prediction—that PG timing reflects the number of beats in a measure—against the alternative hypothesis that PG timing reflects the number of of syllables within each measure. In order to do this, we examined the ratio of the "largest" to the "smallest" PG's in a sentence, labeled as "Ratio PG" in Table 2 (see also Appendix for examples from individual trials). For the isometric sentences, this involved comparing the PG containing the largest number of syllables with that containing the smallest number of syllables. For the heterometric sentences, it involved comparing the PG's associated with a ternary meter (3/4 time) to those having a binary meter (2/4 time). For example, we computed the rate for the isometric sentence called "Pamela" (see Figure 6) by taking the mean normalized PG duration across Groups 1, 3, 5 and 7—all of which have three syllables—to the average of Groups 2, 4, and 6, which have two syllables. If, contrary to our hypothesis, PG duration is based on the number of syllables (i.e., syllable timing), as opposed to the number of beats per measure (i.e., metrical structure), then this ratio should approximate 3:2. For isometric sentences having more than two syllable densities, we used the ratio of the densest PG to the sparsest PG. For example, for the sentence called "Yellow" in Figure 7a, we contrasted PG 2 (5 syllables) with PG 1 (3 syllables), and left out PG 3. Isodurational sentences were excluded from the analysis since there is no basis in their notation for distinguishing PG's that differ in either syllable density or meter.

We ran single-sample *t* tests, comparing the mean of the observed ratios across participants to the predicted ratios (as per the metrical-structure model) of 1.0 for the isometric sentences and 1.5 for the heterometric sentences. A measured value of 1.0 for the isometric sentences would suggest that the duration of the PG's was independent of the number of syllables in the group. A measured value of 1.5 for the heterometric sentences would suggest that speakers observed the meter changes in the sentence, independent of the number of syllables across the PG's. Effect sizes (r^2) and significance levels for this test are shown in Table 2.

The third statistical analysis used linear regression to compare how well the variability in metrical structure (isometric vs. heterometric) predicts the CV for each individual utterance, in contrast to variability in syllable density (number of syllables per notated measure). The isodurational sequences were omitted from this analysis because they have no variability according to either predictor variable. The variable called "CV notation" in Table 2 refers to the variability in syllable density. Because CV is a dimensionless (i.e., ratio-based) measure, variability in the number of syllables is directly comparable to the normalized PG's described earlier. The second predictor was a categorical variable reflecting the sentence-timing category. It was dummy-coded as 0 for isometric and 1 for heterometric types.

Results

Analysis of Individual Sentences

In the figures presented in this section, sentences are shown with their predicted transcriptions, along with boxplots representing the distribution for each normalized PG across participants. The mean CV values across participants are summarized in Table 2 for each sentence in the column labeled "CV PG" (Appendix shows examples of how CV is computed for individual utterances). All raw data are available on request from the authors.

Isodurational sentences. It is uncontroversial that speech can be metric at times. The limiting case consists of what we are

Table 2
Statistical Timing Measures for Each Sentence

Sentence #	Category	Tag	CV PG	CV notation	Ratio PG	Effect size
1	Isodurational	*Twinkle*	.130	0	N/A	N/A
2	Isodurational	*Balcony*	.108	0	N/A	N/A
3	Isodurational	*Mary*	.168	0	N/A	N/A
4	Isometric	*Humpty*	.146	.221	1.125	.476*
5	Isometric	*Pamela*	.150	.208	.987	.019
6	Isometric	*Yellow*	.109	.250	1.115	.477*
7	Isometric	*Greenhouse*	.161	.160	.929	.142
8	Heterometric	*Two*	.308	.272	1.668	.964*
9	Heterometric	*House*	.246	.391	1.419	.938*

Note. Sentence tags match words highlighted by rectangles in Figures 1–7 and the tags listed in Table 1. CV PG = Coefficients of variation of produced PG's computed for each utterance and then averaged across participants. CV notation = CV based on the number of notes per measure in transcription. Ratio PG = the ratio of the mean PG's associated with dense (or long) measures versus the mean PG's for sparse (or short) measures (see text for details). Effect size = r^2 for t-tests contrasting the mean PG ratio for each sentence to a ratio of 1; * indicates significance of this t-test at $p < .05$.

calling isodurational sentences, in which the meter is stable and in which the notated syllabic durations are all equal. A salient example of this sentence-timing category is a syllable-timed passage of verse, such as *Twinkle Twinkle*. We had participants read this nursery rhyme as a "sanity check" for establishing an operational measurement of metricality in speech. The basic idea behind using this sentence was that, if we were not able observe metricality with this passage (as well as with *Humpty Dumpty* below), it would be unreasonable to detect it in sentences that were not explicitly based on verse-like properties of meter. Figure 2 shows a rhythmic transcription of *Twinkle Twinkle*. The mean CV of the produced PG's for this verse was .130. This value provides a benchmark for the PG-level variability of a sentence that is supposed to be isodurational.

Two novel isodurational sentences were constructed to demonstrate simple duple and simple triple meters, respectively. As with *Twinkle Twinkle*, the syllables in these sentences had only a single duration-value, as represented by the exclusive use of quarter notes in their transcriptions. In addition, these sentences dealt with everyday themes, rather than fanciful ones like *Twinkle Twinkle* and *Humpty Dumpty*. Figure 3 shows the sentence in simple triple meter (3/4 time): *The balcony facing the Jamison building was painted with beautiful colors*, which has the tag name "balcony" in Table 2. The mean CV of produced PG's for this sentence was .108. Hence, even for a completely unfamiliar sentence with no implied verse rhythm, participants were able to read this sentence with a strong sense of meter. A similar though less striking result was obtained with the duple-meter sentence (see Figure 4): *Mary purchased purple flowers Monday morning every week*, whose mean CV value was .168. In examining why additional variability was seen in this sentence compared to the last one, we observed that the fifth PG was unexpectedly short, corresponding with the word "Monday."

Isometric sentences. The second sentence-timing category consisted of sentences with a fixed meter but that had more than one duration-value in the sentence. The isometric sentences allow us to distinguish the predictions of stress-timed and syllable-timed interpretations of sentences in a way that the isodurational sentences do not, since prominence groups now have variable numbers of syllables (see ANOVA analyses below). Figure 5 shows an analysis of the first half of *Humpty Dumpty*, with its combination of 3-syllable and 2-syllable PG's, as well as the associated use of two duration values in the transcription. We were surprised to obtain a high mean CV value of .219 for this verse passage. However, the explanation for this high value was apparent upon examining the duration of the fourth PG. This corresponded with the interval between "wall" and "Humpty," in other words the end of the first sentence and the start of the second one. Clearly, participants were inserting a brief pause after the sentence break. If we eliminate the fourth PG from the analysis, the CV value becomes reduced to .146, more in line with our expectation of metricality for this verse passage.

Figure 6 introduces the first complex rhythmic mechanism into the analysis, namely, polyrhythm. The sentence—*Pamela purchased beautiful flowers Saturday morning all through the year*—creates an alternation between 3-syllable and 2-syllable groupings, all with initial stress. Note that this sentence is matched to the sentence in duple meter described in Figure 4 ("Mary"), except that the disyllables (trochees) are converted to trisyllables (dactyls) in every second bar. The predicted meter does not involve an alternation between triple and duple meters, but instead a constant duple meter in which the 3-syllable units are spoken with the same duration as the 2-syllable units, thereby creating a metrical conflict known as a *polyrhythm*, in this case a 3-against-2 polyrhythm. Had people spoken the sentence in a purely syllable-timed manner, then the 3-syllable groups should have had, on average, 1.5 times the duration of the 2-syllable groups. However, they did not. The average normalized duration value of the four 3-syllable groups was 0.99 and that for the three 2-syllable groups was 1.01. Hence, the 3-syllable groups and 2-syllable groups were spoken, on average, with the same duration, as predicted by a view of speech rhythm based on metrical structure. This sentence, as transcribed in Figure 6, had a mean CV of .150, better than the simple-duple analogue in Figure 4. Hence, this result provides strong evidence that participants spoke this sentence in the polyrhythmic manner shown in the transcription and that the syllables in this sentence were of *two different duration values*, with shorter durations for the syllables in the 3-syllable groupings. Interestingly, the largest source of variability was again seen with the day-word "Saturday,"

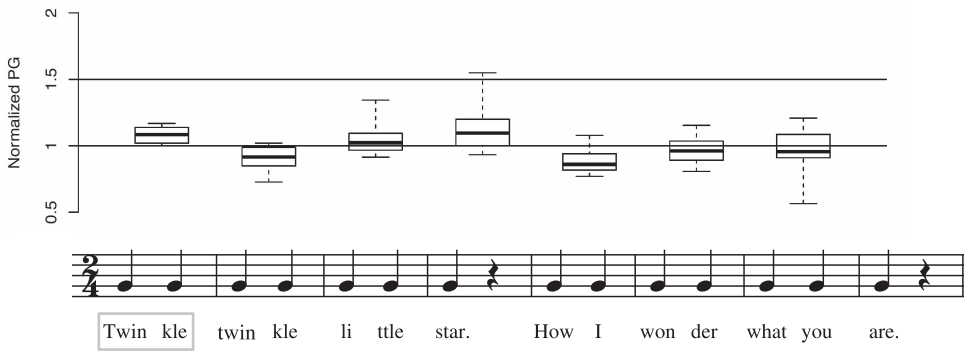

Figure 2. Musical transcription for the sentence tagged "Twinkle" (indicated by the rectangle). Boxplots above the notation display the distribution of normalized PG's across participants. In each boxplot, the rectangle surrounds the interquartile range, the internal line displays the median, and the whiskers span to the most extreme values. Pitches are arbitrary, and thus no clef is displayed.

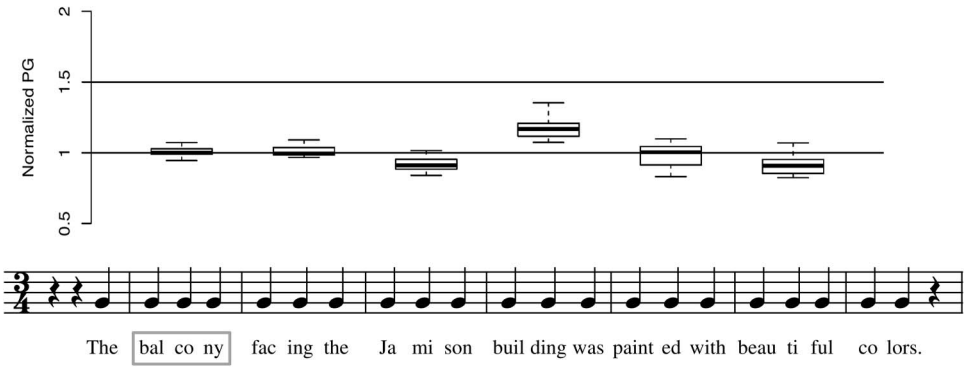

Figure 3. Musical transcription for the sentence tagged "Balcony" (indicated by the rectangle). Boxplots above the notation display the distribution of normalized PG's across participants. In each boxplot, the rectangle surrounds the interquartile range, the internal line displays the median, and the whiskers span to the most extreme values. Pitches are arbitrary, and thus no clef is displayed.

which people spoke in a rushed manner, as with "Monday" in its counterpart sentence in Figure 4 ("Mary").

The last two isometric sentences are each paired with heterometric counterparts below. One of them examines the phenomenon of narrow focus, and the other one compares a compound noun ("greenhouse") with the associated adjectival phrase ("green house"). A common demonstration of prosodic effects in phonology involves taking a single sentence and assigning focus to different words within it (e.g., TWO big dogs vs. two BIG dogs vs. two big DOGS). Words under focus are well known to have pitch accents in the melodic domain (Ladd, 1996), and intonational theories like ToBI that focus on speech melody have provided detailed models of what happens to focused syllables and others in their environment (Beckman & Pierrehumbert, 1986). We emphasize here the rhythmic, rather than the melodic, effects. We present the second sentence first, due to the fact that it fits into our isometric category: *Miguel bought two YELLOW shirts at the men's store by the bay*. This sentence was modeled with a rhythm in simple triple meter (Figure 7a), and the obtained CV value was .109 (see Table 2), one of the lowest values of any sentence in the sample. The fact that this mean CV was lower than the verse passage *Twinkle* is most likely due to the fact that it did not contain a sentence break, which was noted to be a source of variability for the two verse passages. Next, this sentence is the first one discussed thus far that shows durational reductions for function words, as evidenced by the duplets for "at the" and "by the" in the transcription. While the word "yellow" assumes a downbeat position—in keeping with its role as the focus word of the sentence—the notated durations of its syllables are reduced to become eighth notes, something that is not predicted by any current approach to speech rhythm, including metrical phonology. The companion sentence, with a focus on the word TWO, will be discussed in the next section on heterometric sentences.

A related effect to the contrast between two points of focus in a sentence is found in sentences containing compound nouns. In our particular case, we contrasted the compound noun "greenhouse" with the adjectival phrase "green house." As with the focus sentences, we predicted that a downbeat should fall on "green" for "greenhouse" and on "house" for "green house"; the transcriptions reflect this. Figure 8a demonstrates the predicted triple rhythm for the version containing the compound noun: *Nathanial writes novels and lives in a GREENhouse built by a farmer*. The mean CV for this sentence was .161. This CV is comparable to the isoduratonal sentence "Mary," which suggests that "greenhouse" was spoken by

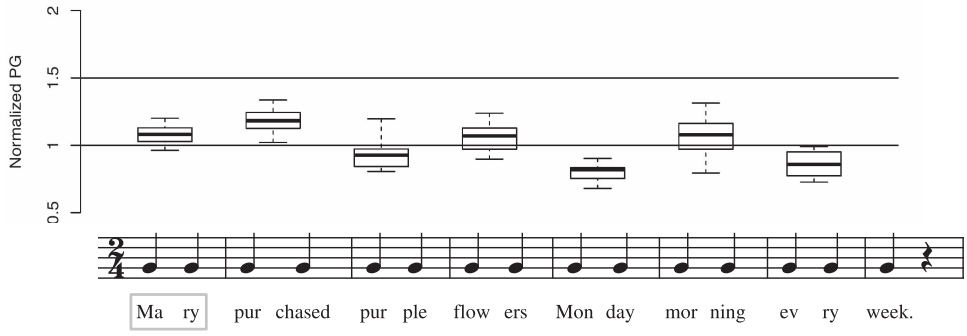

Figure 4. Musical transcription for the sentence tagged "Mary" (indicated by the rectangle). Boxplots above the notation display the distribution of normalized PG's across participants. In each boxplot, the rectangle surrounds the interquartile range, the internal line displays the median, and the whiskers span to the most extreme values. Pitches are arbitrary, and thus no clef is displayed.

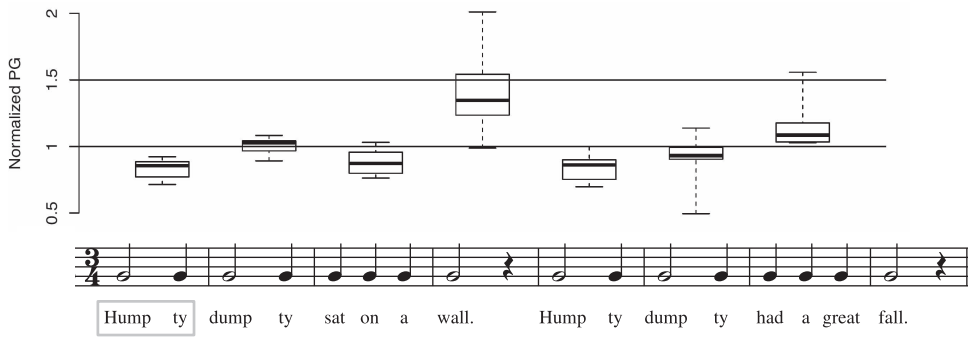

Figure 5. Musical transcription for the sentence tagged "Humpty" (indicated by the rectangle). Boxplots above the notation display the distribution of normalized PG's across participants. In each boxplot, the rectangle surrounds the interquartile range, the internal line displays the median, and the whiskers span to the most extreme values. Pitches are arbitrary, and thus no clef is displayed. It is clear that participants introduce a short pause after the first sentence, as seen in the fourth prominence group.

participants in a fixed triple meter, with 2-against-3 polyrhythms occurring on "greenhouse" and "farmer" (the latter word not being analyzed).

Heterometric sentences. The last category consists of sentences with internal changes in meter. Figure 7b shows the companion sentence to the "yellow" sentence described above, now with the focus on "two": *Miguel bought TWO yellow shirts at the men's store by the bay*. The first thing to notice about this sentence is that a change in focus-word leads to a large change in sentence rhythm, including a switch from an exclusively triple meter for the "yellow"-focus sentence to a duple meter for the initial part of the "two"-focus sentence. To the best of our knowledge, no other approach to speech rhythm accounts for this. As expected, the "yel-" of "yellow" no longer occupies a downbeat, while "two" now does. This is a heterometeric model in which a meter-change occurs from duple to triple meters midway through the sentence. Although there was a great deal of variability for this sentence, it is clear that participants tended to speak this sentence with a meter change, as per the transcription. If one averages the durations of the last PG and divides this by the duration of the first three PG's and (1.43/0.86), the ratio is 1.67, in the vicinity of the predicted value of 1.5. Interestingly, if one ignores the third PG—the one at the point of the meter change—then the ratio of the first two groups to the last one becomes 1.54. Thus, it is likely that a meter change has its most prominent effect on the group that directly precedes it. The transcription for this sentence also shows durational reductions, with duplets for "yellow" and "by the." The alternative transcription of having the sentence be isometric in 2/4 time with "men's store by the bay" being represented as four equal eighth notes was not supported by the productions, which would have given the fourth PG a value close to 1, rather than the observed value of 1.43. Finally, as a result of the change in meter, the mean CV for this sentence was substantially higher than any we have discussed thus far, .308.

The final sentence in the series is the companion to the "greenhouse" sentence: *Nathanial writes novels and lives in a green HOUSE built by a farmer*. Figure 8b shows that the sentence is modeled with a meter change from 3/4 to 2/4 on the word "house" and a durational elongation for the word "house." In fact, the average of the first three PG's to the last two produced a ratio of 1.41, not far from the predicted value of 1.50. However, this occurred with a high amount of between-PG variability in the

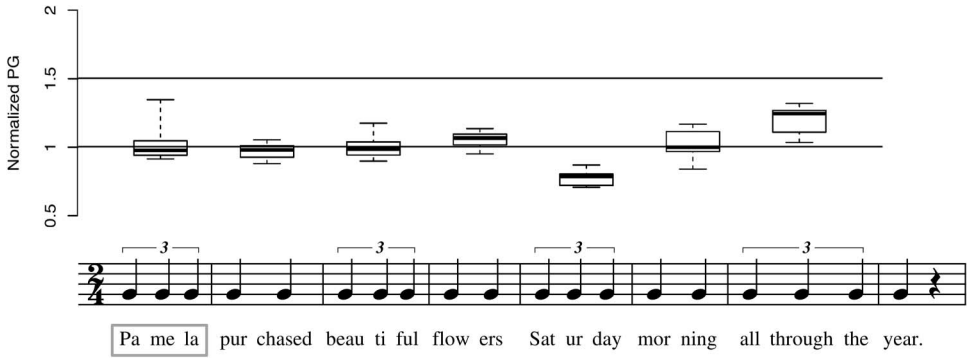

Figure 6. Musical transcription for the sentence tagged "Pamela" (indicated by the rectangle). Boxplots above the notation display the distribution of normalized PG's across participants. In each boxplot, the rectangle surrounds the interquartile range, the internal line displays the median, and the whiskers span to the most extreme values. Pitches are arbitrary, and thus no clef is displayed.

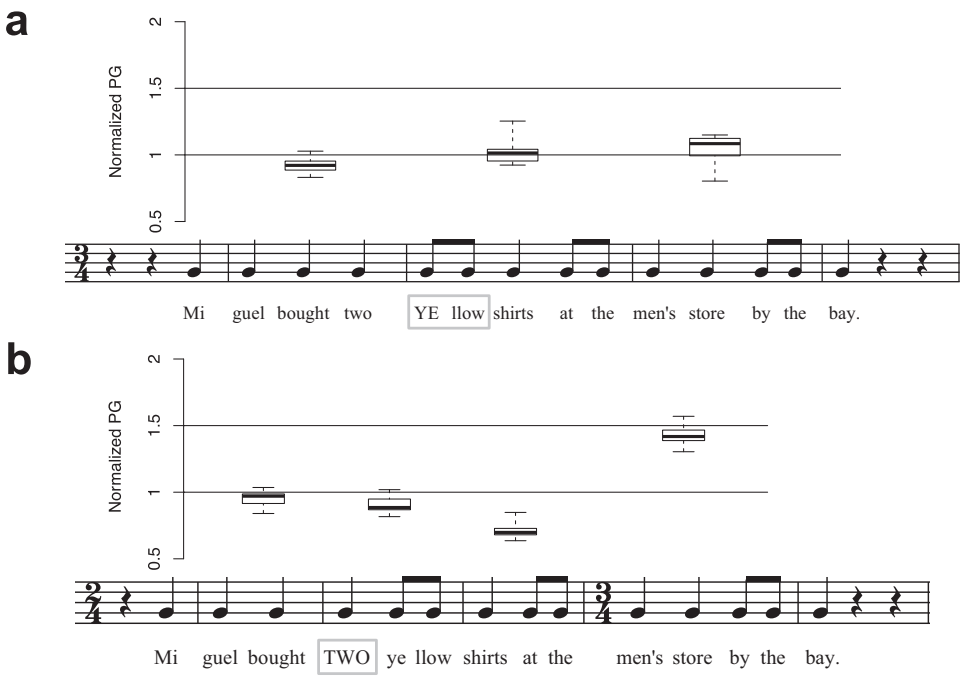

Figure 7. Musical transcription for the sentence tagged "Yellow" (a) and its variant "Two" (b), created by a change in focus. Boxplots above the notation display the distribution of normalized PG's across participants. In each boxplot, the rectangle surrounds the interquartile range, the internal line displays the median, and the whiskers span to the most extreme values. Note that the change in focus results in a change of rhythm compared to the first sentence and that this involves a meter change in the latter half of the sentence. Pitches are arbitrary, and thus no clef is displayed.

durations of the first three groups, which compromises the validity of the findings and of the proposed transcription. In searching for an explanation for this, we listened to the individual recordings and found an obvious source of variability in the results: many of the participants did not provide perceptible emphasis on the intended focus-word "house." It became clear to us after conducting the study that—while an opposition between GREENhouse and green HOUSE is apparent when the two sentences are placed in sequence—"house" is an unnatural word to emphasize when the "green HOUSE" sentence is read in isolation (i.e., when it is not adjacent to its companion sentence). Hence, many participants put *equal* weight on "green" and "house" in this sentence. One line of evidence in support of this is the fact that the ratio of PG3 ("lives in a green") to the mean of PG's 1 + 2 was an unexpected value of 1.28. This is as if the four words of PG 3 were uttered *as four equal quarter notes*, almost as a fusion of the two sentences in Figures 8a and 8b. During the practice session with each participant, we avoided demonstrating sentences or words to participants so that they would not be led to produce our desired rhythms. However, one cost of doing this was that some participants did not create a suitable amount of emphasis on the desired word. If nothing else, the pair of sentences in Figure 8 demonstrates that a change in word pattern (i.e., from compound noun to adjectival phrase) can lead to a clear change in rhythm.

Statistical Analyses of Sentence Types

For these analyses, we grouped sentences according to the three sentence-timing categories described above (see Tables 1 and 2). If, as we predict, metrical variability accounts for speech timing, heterometric sentences should differ from the other two categories. However, if rhythmic variability dominates, then isodurational sentences may differ from both of the other two categories.

We start by analyzing overall PG variability per utterance. Figure 9 shows mean CV (bars) as a function of sentence-timing category. A within-subjects ANOVA was run with a single factor based on three sentence timing categories: isodurational (stable meter and invariant syllable durations), isometric (stable meter but variable timing of syllables within measures), and heterometric. There was a highly significant effect of sentence-timing category on CV's, $F(2, 26) = 118.90$, $p < .001$, $r^2 = .90$. Post hoc tests using a Bonferroni correction showed that heterometric sentences were more variable than either isometric or isodurational sentences, which did not differ from one another. For comparison, the "Notation" line in Figure 9 displays corresponding CV's based on variability in the number of syllables per measure (i.e., syllable density) in the transcriptions (see the "CV Notation" column in Table 2). In contrast to the measured CV values, variability attributable to notated syllable density shows a large increase from the isodurational to the isometric sentences. However, the measured CV's were lower than the CV's predicted from syllable density for both the isometric and heterometric sentences, and were outside the upper limit of the 95% confidence interval in each case. Therefore, metrical structure appears to be a better predictor of PG variability than syllable density, and may to some extent serve

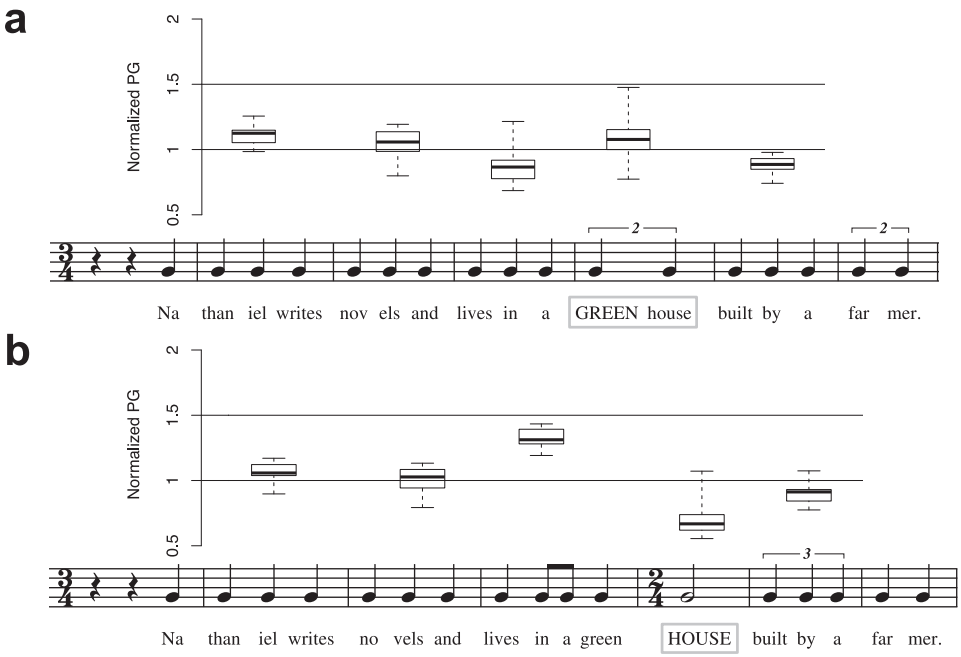

Figure 8. Musical transcription for the compound-noun sentence tagged "Greenhouse" (a) and its adjectival variant "House" (b). Boxplots above the notation display the distribution of normalized PG's across participants. In each boxplot, the rectangle surrounds the interquartile range, the internal line displays the median, and the whiskers span to the most extreme values. Note that the change in wording results in a change of rhythm compared to the first sentence and that this involves a meter change in the latter half of the sentence. Pitches are arbitrary, and thus no clef is displayed.

to stabilize timing. The regression analysis reported below follows up on this possibility.

Next, given that the isometric and heterometric sentences both have variable numbers of syllables per PG—unlike the syllable-timed isodurational sentences, which always have identical numbers of syllables per PG—we wanted to test a syllable-timed null interpretation against the models of metrical structure presented in the transcriptions. For this, we examined the ratio of the "largest" to the "smallest" PG's in a sentence. For the isometric sentences, this involved comparing the PG with largest number of syllables to that with the smallest number of syllables. For the heterometric sentences, it involved comparing the PG's in 3/4 time with those in 2/4 time. We ran paired-sample t tests comparing these ratios across sentence timing categories, along with single-sample t tests comparing the mean ratio within each sentence timing category to the predicted ratio of 1.0 for the isometric sentences or 1.5 for the heterometric sentences. Table 2 shows the measured ratios and effect sizes for each of the six sentences ("Ratio PG" and "Effect size" columns), and Figure 10 shows the means graphically. The paired-sample t test on these means was significant and reflected a large effect size, $t(13) = 12.31$, $p < .001$, $r^2 = .92$. Furthermore, the mean for the isometric sentences did not differ significantly from a ratio of 1 (the prediction based on metrical structure), $t(13) = 1.75$, $p > .05$, $r^2 = .19$, whereas the mean for the heterometric sentences did, with a large effect size, $t(13) = 20.39$, $p < .001$, $r^2 = .97$. Heterometric sentences, however, did not differ from a ratio of 1.5, which was the ratio predicted by the change in meters, $t(13) = 1.63$, $p > .10$, $r^2 = .17$. In both cases, the ratio predicted by the model fell within 95% confidence intervals around each sample mean.

It is important to consider how well these sentence-category effects relate to individual sentences. Looking to the isometric sentences, two of them yielded ratios that were not significantly different than 1. Contrary to predictions, though, two other isometric sentences ("Humpty" and "Yellow") had ratios that were significantly greater than 1 (see Table 2). However, the effect sizes for these sentences were considerably smaller than those found for the heterometric sentences (approximately half the size), and their differences from 1 in absolute terms were quite small, on the order of 12%. Overall, the ratios of PG durations in spoken sentences are more strongly attributable to changes in metrical structure than to changes in PG syllable density, although syllable timing does seem to be making a contribution to speech rhythm in some of the isometric sentences.

Finally, we further explored the syllable-timed alternative interpretation of the sentence rhythms using a multiple regression analysis with two predictors. One predictor was based on variability in syllable density, labeled as "CV notation" in Table 2. The other predictor was a dichotomous variable based on the distinction between isometric sentences (including those that are fully isochronous) and heterometric sentences. Both of these predictors were regressed on the variability of PG's simultaneously, and partial regression coefficients were used to determine how well each predictor accounted for this variability independent of the other. The regression equation with both predictors accounted for 53% of the variance across all sentences and participants, $F(2,$

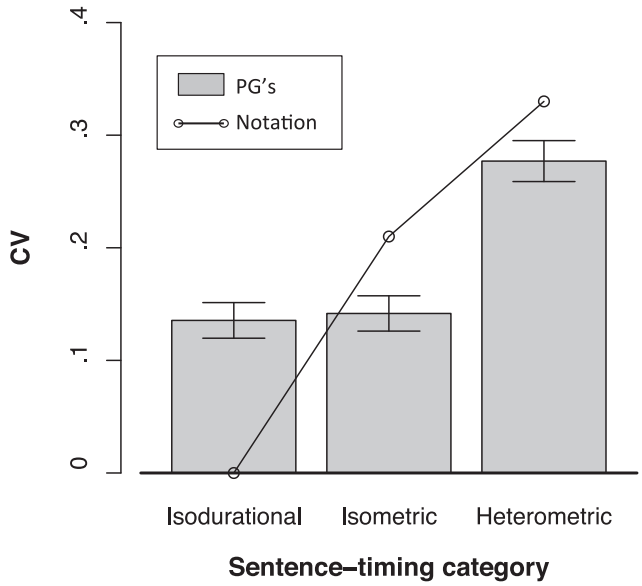

Figure 9. Bar plot showing mean coefficients of variation (CV) for PG's as a function of sentence-timing category. The superimposed line represents CV's based on variability in the number of syllables per notated measure. Error bars display 95% confidence intervals.

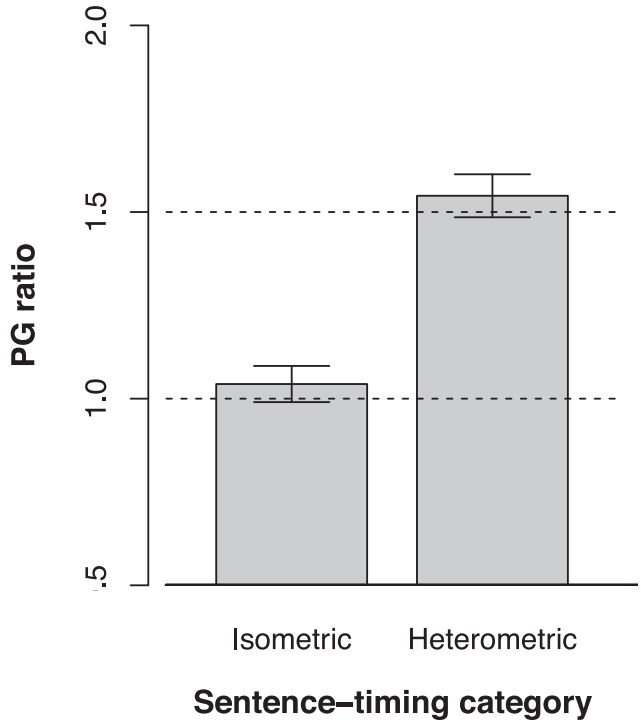

Figure 10. Bar plot showing the ratio of mean PG duration for long PG's to short PG's based on syllable number (isometric sentences) or beats per measure (heterometric sentences). Horizontal lines highlight ratios consistent with the prediction of equally timed PG's for the isometric sentences (1.0), or the long/short ratio based on meter changes for the heterometric sentences (1.5). Error bars display 95% confidence intervals.

123) = 69.23, $p < .001$. More importantly, sentence-timing category accounted for a significant portion of the variance in CV's when controlling for variability in number of syllables per PG, partial $r = .64$, $p < .001$. By contrast, variability in number of syllables per PG did not account for a significant portion of variance when controlling for sentence-timing category, partial $r = -.08$, $p > .10$. These results again suggest that metrical structure did a better job of accounting for the production data than a purely syllable-timed interpretation.

Discussion

We have presented a musical model of speech rhythm, one that shows many similarities to ideas put forth by Joshua Steele in 1775 but that quantifies them experimentally. In particular, we tested how closely the timing of prominence groups in spoken sentences reflects the stability of the notated meter in rhythmic transcriptions of these sentences. Our analyses confirmed these predictions. PG timing was stable when meter remained invariant, regardless of how variable the constituent syllables within PG's were. Conversely, PG timing varied when sentence transcriptions featured a change in meter, and again this variability was independent of how variable the constituent syllables were. A central tenet of the musical model is that speech rhythm can be characterized by a metrical structure. Having provided empirical support for the existence of metrical structure in a corpus of novel sentences, we now elaborate on the implications of the musical model for a theoretical understanding of the components of speech rhythm.

Toward a Musical Model of Speech Rhythm

To the best of our knowledge, there is no contemporary approach to speech rhythm that depicts the temporal pattern of syllable onsets within an utterance, including the relative duration of each syllable. Most approaches are based on either mappings of syllabic stress-points (metrical phonology) or on descriptive statistics of an utterance as a whole (nPVI). Therefore, the musical model fills an important void in the field of speech rhythm. Beyond rhythm alone, a transcription-based approach has the potential to represent *both* pitch and rhythm using the same set of symbols. At the present time, approaches to speech intonation are polarized between rhythmic approaches, such as those mentioned in the Introduction, and pitch-based approaches like ToBI that ignore rhythm (Beckman & Pierrehumbert, 1986; Ladd, 1996). *Musical transcription provides a means of unifying the melodic and rhythmic domains of speech* in a way that has not been attempted since Steele originally proposed this. It is too early to elaborate all of the factors that contribute to meter, subdivision, and meter change in sentences, but we believe that the study of speech rhythm should be dedicated to a search for these principles. The study of speech rhythm is nothing if not a conception of time, more specifically the timing of syllable onsets and a specification of the relative duration of syllables.

Like Steele (1775), we argue that speech rhythm is based on a small number of basic mechanisms as related to the same type of metrical hierarchy found in music, dance, and poetry, namely meters (often 2- and 3-beat meters), subdivisions of component beats into subbeats according to small-integer ratios (typically 2- and 3-beat subdivisions), and metrical conflicts like polyrhythms

(especially conflicts between 2- and 3-beat elements). Scholars of poetry have been aware of such rhythmic devices for centuries (Fabb & Halle, 2008). In addition, we present the novel proposal that meter-change is a central component of speech rhythm, a proposal that helps circumvent many of the problems inherent in the isochrony models of the past. As with models of stress-timing in linguistics, the existence of metrical structure in speech implies that syllables in a sentence need not all have the same duration. However, instead of simply arguing that stresses occur at equal time intervals, the musical model attempts to represent the actual pattern of syllable onsets and thereby provides insight into how and why certain syllables undergo durational compressions or elongations. In addition, there are numerous means of generating stress-timed patterns, as music theory so amply demonstrates. For example, *Twinkle Twinkle* and *Humpty Dumpty* are both stress-timed sentences and yet have different meters and different durational patterns, just as a salsa and a waltz have different metrical patterns. In addition, the same sentence can be uttered with different rhythmic patterns, for example when the focus word is shifted. The results in Figure 7 showed that speakers did indeed change the overall rhythmic properties of the utterances when narrow focus was shifted from "two" to "yellow" in the identical sequence of words. The musical model can represent such changes, including those related to emotional expression, dialects, foreign accents, and even speech pathology (e.g., trigger points for stuttering within a sentence).

We have presented a quantitative method for studying speech rhythm that involves making an intuitive a priori representation of the metrical structure of a sentence, recording a group of speakers reading the sentence aloud, and measuring the extent to which the group's productions conform, on average, with the transcription's temporal predictions. As shown in the Results section, most of our transcriptions were borne out by the productions, suggesting that metricality in speech can be measured reliably and that it can be produced by untrained participants reading completely unfamiliar sentences in the absence of entrainment cues. Not all transcriptions fit the assumptions of our model equally well. The results with certain problematic sentences revealed the fact that different people can read a given sentence in multiple manners. However, the use of musical transcription can accommodate such diversity in production. Transcriptions can be modified based on the observed speech patterns of participants to create multiple rhythmic variants of a given sentence, with a caveat being that meter changes should be minimized. Diversity of this kind across participants was observed by Cummins and Port (1998) in their initial speech cycling study and was represented with musical notation in their paper. In fact, a musical transcription is the only representation of a sentence that can allow a speaker to read an unfamiliar sentence with precision. The metrical grids of metrical phonology (Goldsmith, 1990; Liberman & Prince, 1977) and the diacritical stress-markings of poetic analysis provide far less precise information about relative syllabic durations than is possible with musical notation.

A Unit of Rhythm: Prominence Groups and Musical Meters

An important step toward creating a musical model of speech is to define a unit of rhythm. As with Steele, we propose that the basic unit of speech rhythm is the "prominence group," analogous to a bar or measure in music. The defining feature of a prominence group is that it begins with a strong beat (i.e., a stressed syllable in the case of English), just as a musical measure always begins with a strong beat. Hence, prominence groups always begin with a musical downbeat.

Just as with any description of musical rhythm, each syllable in a sentence transcription is assigned a duration value, an essential feature missing in virtually every other model of speech rhythm. Importantly, these are *relative* duration values, just as in music; an understanding of absolute duration would require a specification of the duration of a note-value at some level of the metrical hierarchy (akin to a metronome marking in music). Transcriptions of our stimulus sentences showed that syllables could differ in their relative duration values. Some syllables could be half the duration of others (i.e., when duplets occurred) and some could be two thirds of others (i.e., when 3-against-2 polyrhythms occurred). Several factors contribute to variability in duration for syllables (Dauer, 1983). For example, consonant clusters generally make syllables longer than simpler syllables (e.g., CCCVCCC vs. CV, where C = consonant and V = vowel). Languages that are classified as stress-timed tend to have more-complex syllable structures than those classified as syllable-timed (Dauer, 1983; O'Dell & Nieminen, 1999), and thus have greater variability of syllable types and durations (Grabe & Low, 2002).

Speech cycling experiments in which short phrases, such as "big for a duck," are entrained to a metronome beat show that duple and triple meters are stable metrical structures for such productions (Cummins & Port, 1998; Tilsen, 2009), arguing that the regular beats of meters are strong attractors for syllables onsets, especially in the case of stressed syllables. This was seen to be the case in our test sentences, all of which involved duple and/or triple meters. Such is the case as well for much poetry and sung text throughout the world and across historical time. As we argue below, our proposal of heterometers in speech is quite different from saying that speech is arhythmic or nonmetric. It is instead a means of countering such ideas by arguing that meters can change not only across sentences but within them as well.

Simple Subdivisions of Beats: Duplets and Triplets

A reasonable optimality rule for speech rhythm would be to minimize meter changes within a sentence. To this end, we can imagine two major meter-preserving mechanisms in speech. Both of them involve creating subdivisions of the basic beat into sub-beats and thus generating a *metrical hierarchy* for the phrase: (a) subdividing beats according to 2's and 3's to generate duplets and triplets, respectively, and (b) subdividing beats in a complex fashion to generate polyrhythms (discussed in the next section). In music's metrical hierarchy, subdivisions of beats generally take the form of *small integer ratios*, such as duplets (each one having one half the duration of the basic beat) and triplets (each one having one third the duration of the basic beat), and our results show this to be the case in speech as well. Such duplets and triplets reflect the fact that syllable durations are compressed in speech. For languages like English, there are well-characterized phenomena like vowel reduction that lead to corresponding reductions in syllable duration for unstressed syllables in polysyllabic words. Likewise, certain function words, such as clitics, articles, and

many prepositions, are monosyllabic words that tend to get uttered in a highly reduced manner. Hence, both syllable stress and syntactic role become factors in defining compressions in syllable duration. This was seen in several of the test sentences in the present study, including the phrases "in a" (duplet) and "built by a" (triplet). It is also observed in studies of metronome-entrained speech (Cummins & Port, 1998; Tilsen, 2009), for example "big for a duck," where the function words "for" and "a" undergo durational reduction compared to the content words "big" and "duck." The idea that the beats in a speech meter can be divided into subbeats according to small integer ratios is consistent with the "harmonic timing effect" seen in these studies, in which neural oscillators are proposed to operate at harmonic fractions of beats, especially halves and thirds, thereby attracting perceptual attention to these locations (Port, 2003).

Complex Subdivisions of Beats: Polyrhythms

A related meter-preserving rhythmic device of subdivision is polyrhythm, a device with no precedent in speech cycling experiments but that is present in Steele's (1775) transcriptions. In music, the concept of a polyrhythm implies a conflict between incompatible rhythms. For example, if two people were to simultaneously tap a 3-beat and 2-beat rhythm, respectively, against the same drumbeat, this would create a 3-against-2 polyrhythm, since 3 and 2 are not divisible by a common integer (except 1). Polyrhythm is another manifestation of the phenomenon of subdivision, but one in which the beats are not mutually divisible as simple integer ratios. The results of the present study demonstrate that polyrhythms are a natural part of speech, providing further support for a musical interpretation of speech rhythm. The sentence presented in Figure 6 (*Pamela*) created an alternation between trisyllabic (dactylic) and disyllabic (trochaic) groupings, all having initial stress. As predicted by our transcription, participants read this sentence such that the trisyllabic and disyllabic groups occupied equal time intervals, as would be the case if the sentence were read as a musical polyrhythm with two different syllabic duration values. It is interesting to point out that pianists are sometimes taught to perform polyrhythms between their two hands using short *sentences* as their metrical guides (e.g., "hot cup of tea" approximates a 3:2 polyrhythm). Such a method could only work if the sentences themselves embodied these polyrhythms.

Heterometers: Changes of Meter Within a Sentence

A natural sentence spoken by an individual will not have the rhythmic simplicity of a passage of composed verse. A significant departure of our model from classic models of isochrony is that it posits the occurrence of meter changes within sentences, for example, from a triple meter to a duple meter. Hence, we propose that sentences can be heterometric, and that meter-change is a central feature of speech rhythm, especially in longer or more-complex sentences. This was demonstrated most clearly in the sentence *Miguel bought TWO yellow shirts at the men's store by the bay*, where the first half of the sentence was spoken in a 2/4 m and the second half in a 3/4 m. The location of greatest imprecision in the sentence was the bar containing the meter change, as might be predicted by an oscillator-coupling model.

The notion of meter change might provide one solution to critiques that have been historically levied against models of speech isochrony (e.g., Lehiste, 1977; Nolan & Jeon, 2014). At the same time, the heterometric sentences provided the least reliable results in this study and therefore require further study in order to understand their properties. However, we believe that a model of speech rhythm that makes allowance for meter change is a necessity in order to account for the obvious complexity of spontaneous speech, a topic that we have not broached in the present study.

When meter-changes occur (and sometimes even when they do not), the tempo can change as well. In other words, the durational value of the basic beat can become shorter or longer. Hence, another important feature of speech rhythm is not only changes in the metric groupings across a sentence but also changes in the duration-value of the beats within that meter, in other words *tempo change*. Tempo modulation is an important aspect of expressive timing in musical performance (Friberg, Bresin, & Sundberg, 2006; Repp, 1992, 1994). Hence, we believe that it will also turn out to be a significant factor in expressive intonation for speech. The musical model of speech rhythm, with its explicit attempt to model syllable durations, provides a promising means of representing speech prosody.

Cross-Linguistic Considerations

What are the determinants of these rhythmic mechanisms cross-linguistically? At least two interdependent factors seem to be strong candidates: polysyllabilicity of words and the presence of syllabic stress within words. Languages like English that have polysyllabic words with lexical stress probably lend themselves to having meter changes in sentences. Languages that are more monosyllabic will probably have more-constant meters. But even a language like Cantonese that has a simpler syllable structure than English, and is thus less prone to meter change, still shows subdivisions of beats in a pervasive manner, most especially on function words (Chow, Brown, Poon, & Weishaar, 2010). Hence, subdivision of beats might be a more general rhythmic mechanism than heterometers.

In our opinion, the classic dichotomy between stress-timed and syllable-timed languages is in serious need of an overhaul. Speech rhythm seems to be inherently based on stress timing (Dauer, 1983; Fant et al., 1991), even for languages that lack word-level stress, like Cantonese (Chow et al., 2010), Korean, and Tamil (Nolan & Jeon, 2014). A similar conclusion was reached by Fant et al. (1991) in a comparison of Swedish, English and French production of the same text translated into their respective languages. What seems to vary across languages are the kinds of features we have talked about: the durational variability of constituents that sit between stress points (i.e., subbeats); the presence of meter changes; and the presence of tempo changes. We suspect that there is no language that is based on constant strings of isochronous syllables. Instead, one should find, at one end of the spectrum, rhythmically simpler languages that have few subdivisions of beats, relatively constant meters, and relatively constant tempos. At the other end should be rhythmically complex languages that have greater numbers of subdivisions of beats, more frequent meter changes, and more frequent tempo changes. From our experience with this analysis, Cantonese and English might represent prototypes of these two varieties of speech rhythms, respectively. This jibes perfectly with the well-established notion that languages differ in the durational variability of their syllables (Grabe & Low, 2002; Ramus et al., 1999).

Nolan and Jeon (2014) have argued that speech is, in reality, arhythmic, and that the notion of speech rhythm is nothing more than

a metaphor. We have argued throughout this paper that equating rhythm with isochrony is a mistake, and that the absence of isochrony does not necessitate that speech be arhythmic. In our view, speech is rhythmic, but it is based on a complex set of rhythmic patterns. Much music too is based on complex rhythms. In fact, our notion of a heterometer is taken directly from the literature on musical rhythm. The application of musical notions of rhythm to speech has thus far been dominated by a single unsuccessful concept, namely isochrony. We believe that a more sophisticated understanding of rhythm, one that takes full advantage of the rich tool kit offered by musical analysis, can enlighten the nature of speech rhythm.

Limitations

This work suffers from several significant limitations, some of which are pervasive in the linguistics literature overall. For example, the study was based on read speech, rather than spontaneous speech. Spontaneous speech is far more complicated rhythmically than read sentences, not least because of the presence of pauses, fillers, speech errors, and the frequent use of sentence fragments. In fact, the majority of studies of speech rhythm in production are based on read speech (Cummins & Port, 1998; Lee & Todd, 2004; Tilsen, 2009). Next, one of our heterometric sentences, *House*, showed a high level of between-PG variability. We feel that this was due in part to our need to avoid influencing the participants' productions by demonstrating the sentences and revealing the rhythms that we were seeking. However, upon analysis, it was clear that several of the participants failed to achieve contrastive stress in *House*. The next phase of the work needs to focus on complex sentences and on creating multiple models for single sentences.

Next, two of the isometric sentences, while having a relatively low ratio for component PG's containing more syllables compared to those having fewer syllables (i.e., ratios of 1.125 and 1.115, respectively), were still found by the *t* tests to be significantly different than 1.0, suggesting that syllable timing did make a contribution to these sentences beyond what was predicted by metrical structure alone. While this finding represents a limitation in the context of the current transcriptions, it also suggests avenues for further explorations of the rhythmic properties of such sentences. For sentences that do not conform well with transcriptions, they can be examined post hoc to try to infer where the inaccuracy might emanate from. At least two major sources can be examined. One is that there is a large level of inter-individual variation in the data. Another is that the a priori transcription is inadequate. In such a case, the observed production data can suggest alternative transcriptions for the sentence, which could then be analyzed in a follow-up experiment. There might even be situations in which there is a bimodal distribution in the pattern of production, for example due to differences in the pronunciation of certain words. Consider the rhythmic contrast between "The | president | purchased in | SURance" and "The | president | purchased | INsurance", with their alternative prominence groupings.

Finally, it is important to point out that the present analysis is a bar-level analysis, where the primary durational unit that is analyzed is the PG. A more detailed analysis would focus on the syllable level. For example, while the bar-level analysis of *Pamela* (see Figure 6) showed that the dactyl "Pamela" was spoken with the same duration as the trochee "purchased," a syllable-level analysis could further verify (or not) that the three syllables of Pamela each have 2/3 the duration of each of the two syllables of purchased. However, even for musical works that are in simple meters, notes often vary from one another in duration value due to factors related to expressive timing (Repp, 1992), such as rubato. Speech further complicates matters by adding phonetic (articulatory) diversity onto the timing units, thereby contributing an additional source of timing variability that would have to be taken into account in a syllable-level analysis of speech rhythm.

Conclusions

Our musical model posits a small number of fundamental rhythmic mechanisms that should be applicable across languages. We see a basic similarity of speech rhythm to the hierarchical structure of musical rhythm through an organization of sentences into prominence groups headed by strong beats. Next, we posit that meter-change is central to speech rhythm, and thus that speech is often heterometric rather than isochronous. Tempo changes can also occur during the course of an utterance, altering the duration values of beats. In addition, we see two meter-preserving rhythmic mechanisms involving subdivisions of beats into subbeats: (a) subdivisions according to 2's and 3's to generate duplets and triplets, respectively; and (b) subdivisions according to complex ratios to generate polyrhythms. Although the relative importance of these mechanisms varies across languages, it is likely that all of them are present in some form in all languages.

The cognitive implication of the musical model of speech is not that speech is an example of music but instead that speech and music share an underlying prosodic system (Lerdahl, 2001). At the rhythmic level, this system is characterized by a basic metricality involving 2- and 3-beat meters and subbeats. At the melodic level, this involves features like declination, pitch accents, affective expression, and perhaps overall melodic contour as well. There are numerous examples of metric speech (Cummins, 2013), but many of them are driven in an explicit manner by entrainment signals, such as musical beats (e.g., rap) or mutual entrainment with other individuals (e.g., the chanting of political slogans). However, when it comes to conversational speech, we believe that, to the extent that the rhythms that we posit do operate at all, these rhythms should be occurring in an implicit and unconscious manner, as driven by some type of internal oscillator at the level of the production mechanism (Cummins & Port, 1998; Port, 2003; Tilsen, 2009). Much work is needed to explore the question of whether spontaneous speech has an underlying metricality at the level of production (Turk & Shattuck-Hufnagel, 2013). One thing that will complicate such an analysis is the emotional prosody that accompanies spontaneous speech. Studies of the expressive performance of notated music make a distinction between "the score" (i.e., musical notation) and "performance," where performance is seen as an expressive deviation from notation (Friberg et al., 2006). The big question for the field of speech rhythm is whether the brain contains a "score" for the production of spontaneously produced speech. The musical model of speech rhythm, to the extent that it can provide scores for spoken sentences, offers a null hypothesis against which other generative models can be tested.

References

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh, United Kingdom: Edinburgh University Press.

Aristotle. (1996). *Poetics*. [Written roughly 335 BCE]. In M. Heath (Trans). London, United Kingdom: Penguin Books.

Arvaniti, A. (2009). Rhythm, timing and the timing of rhythm. *Phonetica, 66,* 46–63. http://dx.doi.org/10.1159/000208930

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics, 40,* 351–373. http://dx.doi.org/10.1016/j.wocn.2012.02.003

Beckman, M., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3,* 15–70.

Bertran, A. P. (1999). Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages. *Language Design, 2,* 103–130.

Boersma, P., & Weenink, D. (2014). *Praat* [a computer software used for acoustic speech analysis]. Amsterdam, the Netherlands: Phonetic Sciences, University of Amsterdam.

Caplan, D. (2007). *Poetic form: An introduction*. New York, NY: Pearson.

Chow, I., Belyk, M., Tran, V., & Brown, S. (2015). Syllable synchronization and the P-center in Cantonese. *Journal of Phonetics, 49,* 55–66. http://dx.doi.org/10.1016/j.wocn.2014.10.006

Chow, I., Brown, S., Poon, M., & Weishaar, K. (2010). A musical template for phrasal rhythm in spoken Cantonese. *Speech Prosody, 100078,* 1–4.

Cummins, F. (2013). Joint speech: The missing link between speech and music? *Percepta, 1,* 17–32.

Cummins, F., & Port, R. (1998). Rhythmic constraints of stress timing in English. *Journal of Phonetics, 26,* 145–171. http://dx.doi.org/10.1006/jpho.1998.0070

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

Dowling, W. J., & Harwood, D. L. (1986). *Music cognition*. New York, NY: Academic Press.

Fabb, N., & Halle, M. (2008). *Meter in poetry: A new theory*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511755040

Fant, G., Kruckenberg, A., & Nord, L. (1991). Durational correlates of stress in Swedish, English and French. *Journal of Phonetics, 19,* 351–365.

Friberg, A., Bresin, R., & Sundberg, J. (2006). Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology, 2,* 145–161. http://dx.doi.org/10.2478/v10053-008-0052-x

Goldsmith, J. A. (1990). *Autosegmental & metrical phonology*. Oxford, UK: Basil Blackwell.

Grabe, E., & Low, L. (2002). Durational variability in speech and the rhythm class hypothesis. In N. Warner & C. Gussenhoven (Eds.), *Papers in laboratory phonology 7* (pp. 515–546). Berlin, Germany: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110197105.515

Hammond, M. (1995). Metrical phonology. *Annual Review of Anthropology, 24,* 313–342. http://dx.doi.org/10.1146/annurev.an.24.100195.001525

Hayes, B. (1983). A grid-based theory of English meter. *Linguistic Inquiry, 14,* 357–393.

Jun, S. A., & Fougeron, C. (2002). Realizations of accentual phrase in French intonation. *Probus, 14,* 147–172. http://dx.doi.org/10.1515/prbs.2002.002

Kassler, J. C. (2005). Representing speech through musical notation. *Journal of Musicological Research, 24,* 227–239. http://dx.doi.org/10.1080/01411890500233965

Kim, H., & Cole, J. (2005). The stress foot as a unit of planned timing: Evidence from shortening in the prosodic phrase. In *9th European conference on speech communication and technology, Eurospeech interspeech* (pp. 2365–2368). Lisbon, Portugal.

Kiparsky, P. (1977). The rhythmic structure of English verse. *Linguistic Inquiry, 8,* 189–247.

Ladd, R. (1996). *Intonational phonology*. Cambridge, UK: Cambridge University Press.

Lee, C. S., & Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: Application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition, 93,* 225–254. http://dx.doi.org/10.1016/j.cognition.2003.10.012

Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics, 5,* 253–263.

Lerdahl, F. (2001). The sounds of poetry viewed as music. *Annals of the New York Academy of Sciences, 930,* 337–354. http://dx.doi.org/10.1111/j.1749-6632.2001.tb05743.x

Lerdahl, F., & Jackendoff, R. (1983). *A generative theory of tonal music*. Cambridge, MA: MIT Press.

Liberman, M., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry, 8,* 249–336.

Lloyd James, A. (1940). *Speech signals in telephony*. London, UK: Sir I. Pitman & Sons.

Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordecht, the Netherlands: Foris Publications.

Nolan, F., & Asu, E. L. (2009). The pairwise variability index and coexisting rhythms in language. *Phonetica, 66,* 64–77. http://dx.doi.org/10.1159/000208931

Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society of London Series B: Biological Sciences*, *369,* 20130396. http://dx.doi.org/10.1098/rstb.2013.0396

O'Dell, M. L., & Nieminen, T. (1999). Coupled oscillator model of speech rhythm in English. In J. Ohala, Y. Hasegawa, M. Ohala, D. Granville, & A. Bailey (Eds.), *Proceedings of the XIVth International congress of phonetic sciences* (Vol. 2, pp. 1075–1078). Berkeley, CA: University of California.

Palmer, C., & Krumhansl, C. L. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance, 16,* 728–741. http://dx.doi.org/10.1037/0096-1523.16.4.728

Patel, A. D. (2008). *Music, language and the brain*. Oxford, UK: Oxford University Press.

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition, 87,* B35–B45. http://dx.doi.org/10.1016/S0010-0277(02)00187-7

Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *The Journal of the Acoustical Society of America, 119,* 3034–3047. http://dx.doi.org/10.1121/1.2179657

Pike, K. (1945). *The intonation of American English*. Ann Arbor, MI: University of Michigan Press.

Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-center phenomenon. *Journal of Phonetics, 17,* 175–192.

Port, R. (2003). Meter and speech. *Journal of Phonetics, 31,* 599–611. http://dx.doi.org/10.1016/j.wocn.2003.08.001

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America, 81,* 1574–1585. http://dx.doi.org/10.1121/1.394510

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73,* 265–292. http://dx.doi.org/10.1016/S0010-0277(99)00058-X

Repp, B. H. (1992). Diversity and commonality in music performance: An analysis of timing microstructure in Schumann's "Träumerei". *The Journal of the Acoustical Society of America, 92,* 2546–2568. http://dx.doi.org/10.1121/1.404425

Repp, B. H. (1994). Relational invariance of expressive microstructure across global tempo changes in music performance: An exploratory study. *Psychological Research, 56,* 269–284. http://dx.doi.org/10.1007/BF00419657

Rush, J. (2005). *Philosophy of the human voice*. Whitefish, MT: Kessinger Publishing. (Original work published 1827)

Steele, J. (1775). *An essay towards establishing the melody and measure of speech to be expressed and perpetuated by peculiar symbols*. Reprinted as part of the Gale Eighteenth Century Collections Online print editions. Farmington Hills, MI: Gale Cengage Learning.

Tilsen, S. (2009). Multitimescale dynamical interactions between speech rhythm and gesture. *Cognitive Science, 33,* 839–879. http://dx.doi.org/10.1111/j.1551-6709.2009.01037.x

Turk, A., & Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory Phonology, 4,* 93–118. http://dx.doi.org/10.1515/lp-2013-0005

White, L., & Mattys, S. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics, 35,* 501–522. http://dx.doi.org/10.1016/j.wocn.2007.02.003

Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception and Psychophysics, 14,* 5–12. http://dx.doi.org/10.3758/BF03198607

Appendix

Examples of Computation for Statistical Measures

We present here examples of how measures of prominence group (PG) timing are computed for individual sentences and productions from our data set.

We start with an isometric sentence, in which the number of syllables per PG varies. **PG durations** (from the start of one PG to the start of the next) are shown in milliseconds for two performances, a relatively slow and a relatively fast one (see Table A1).

Several points are worth noting. First, because the **coefficient of variation** (CV)—which is the ratio of the standard deviation to the mean—standardizes timing variability, the CV values for both performances are highly similar despite the fact that the slower performance has more-variable PG's than the faster performance. Second, the PG durations have a moderate negative correlation with the number of syllables per PG, which runs counter to the predictions of a syllable-timing model. This is related to a third observation, namely that variability in performances is lower than variability in the number of syllables across PG's.

For this sentence, the **PG ratio** would come from averaging durations of PG's with 3 syllables (PG numbers 1, 3, 5, and 7) and dividing that average by the average duration of PG's having two syllables (PG numbers 2, 4, and 6). For the slow performance, this ratio is 506/529 = .96, and for the fast performance it is 418/436 = .96. In both cases, the ratio is very close to 1, indicating that PG's were produced almost equivalently despite differences in the number of syllables (i.e., syllable density) per PG.

Table A1
Sentence = "Pamela"

			Slow PG's		Fast PG's	
PG	Text	# syllables	Raw	Norm.	Raw	Norm.
1	Pamela	3	486	.94	394	.93
2	purchased	2	507	.98	425	1.00
3	beautiful	3	485	.94	446	1.05
4	flowers	2	571	1.11	458	1.08
5	Saturday	3	386	.75	303	.71
6	morning	2	510	.98	425	1.00
7	all through the	3	667	1.29	528	1.24
8	year.	1	(N/A)		(N/A)	
	M =	2.6	516	1.00	424	1.00
	SD =	.53	86	.17	68	.16
	CV (SD/M) =	.21	.17		.16	
	r(# syllables, PG) =		−.14		−.14	

Note. Norm. = normalized, CV = coefficient of variation. CV and r are identical for raw and normalized PG's.

Table A2
Sentence = "Two"

			Slow PG's		Fast PG's	
PG	Text	# syllables	Raw	Norm.	Raw	Norm.
1	(Mi)guel bought	3	729	.97	627	.99
2	two yellow	3	664	.88	593	.94
3	shirts at the	3	518	.69	418	.66
4	men's store by the	4	1106	1.47	893	1.41
5	bay.	2	(N/A)		(N/A)	
	M =	3.00	754	1.00	633	1.00
	SD =	.82	251	.33	196	.31
	CV (SD/M) =	.27	.33		.31	
	r(# syllables, PG) =		.61		.55	

Note. Norm. = normalized, CV = coefficient of variation. CV and r are identical for raw and normalized PG's. The first syllable of PG1 is treated as an "upbeat" and not counted in that PG duration.

The most critical prediction of the model, however, has to do with comparisons between isometric and heterometric sentences (see Table A2). We now illustrate PG timing with two productions of a heterometric sentence.

As can be seen, both fast and slow performances lead to approximately double the amount of variability across PG's than was found in the isometric sentence, despite the fact that variability in the number of syllables across PG's was much more closely matched, differing by only 6%. Also, the heterometric sentence led to strong positive correlations between PG durations and number of syllables, but this reflects the fact that the PG with the greatest number of syllables (4) was also associated with a change in meter from 2/4 to 3/4.

For a heterometric sentence such as this one, the PG ratio was based on dividing the duration of the single long (3-beat) metrical frame (PG number 4) by the mean duration of PG's with the shorter (2-beat) metrical frame (PG numbers 1–3). For the slow performance, this ratio is 1106/637 = 1.74, and for the fast performance it is 893/546 = 1.64. In both cases, the PG ratio from production approximates the ratio of the number of *beats* per notated meter associated with PG's (3:2 = 1.5), more so than the ratio based on the number of *syllables* associated with PG's (4:3 = 1.3).

Received February 5, 2016
Revision received March 29, 2017
Accepted March 30, 2017 ■