Psychological Research (2014) 78:113–123
DOI 10.1007/s00426-013-0480-1

ORIGINAL ARTICLE

Metrical perception of trisyllabic speech rhythms

Fernando Benadon

Received: 23 December 2010 / Accepted: 18 January 2013 / Published online: 16 February 2013

Abstract The perception of duration-based syllabic rhythm was examined within a metrical framework. Participants assessed the duration patterns of four-syllable phrases set within the stress structure XxxX (an Abercrombian trisyllabic foot). Using on-screen sliders, participants created percussive sequences that imitated speech rhythms and analogous non-speech monotone rhythms. There was a tendency to equalize the interval durations for speech stimuli but not for non-speech. Despite the perceptual regularization of syllable durations, different speech phrases were conceived in various rhythmic configurations, pointing to a diversity of perceived meters in speech. In addition, imitations of speech stimuli showed more variability than those of non-speech. Rhythmically skilled listeners exhibited lower variability and were more consistent with vowel-centric estimates when assessing speech stimuli. These findings enable new connections between meter- and duration-based models of speech rhythm perception.

Introduction

This study seeks to lay the groundwork for connecting two areas of speech rhythm that have shown limited overlap in the research literature: meter and syllable duration. Most metrical accounts of speech rhythm provide frameworks for the organization of stress patterns irrespective of duration. In metrical phonology, the syllable level forms a beat-to-beat foundation that is shaped by points of hierarchical stress prominence (Goldsmith, 1990; Halle & Vergnaud, 1990; Hayes, 1995; Liberman, 1975; Liberman & Prince, 1977; Selkirk, 1984). Timing is primarily ordinal, in the sense that syllables must follow one another sequentially without sacrificing the alternation of strong and weak stress. Syllable duration, if addressed, is secondary to the stress hierarchy. For instance, and despite empirical refutations by Cooper & Eady (1986), a pause of unspecified but reasonable magnitude may be represented by the insertion of a "silent grid position" in order to avoid adjacent stressed syllables (Liberman, 1975; Selkirk, 1984). The same essentially relativist approach to syllable duration holds for meter-based comparisons of speech and music (Halle & Lerdahl, 1993; Hayes, 2009; Rodríguez-Vázquez, 2010; Winn & Idsardi, 2008) and for studies of metrical segmentation (Cutler, 1990; Echols, Crowhurst & Childers, 1997; Morgan, 1996; Nazzi & Ramus, 2003). Meter has also been at the forefront of poetry analysis and generative metrics (e.g., Attridge, 1982; Cureton, 1992; Fabb & Halle, 2008; Halle & Keyser, 1971; Hollander, 1989; Kiparsky, 1977; Prince, 1989; Weismiller, 1989). There again, the focus is on patterns of stress rather than duration, as is the case with empirical accounts of meter in speech production by young children (Gerken, 1991), the structure of children's rhyming (Kelly & Rubin, 1988), and the role of rhythm during speech processing (Rothermich, Schmidt-Kassow & Kotz, 2012). When meter and stress are intertwined, attention is paid to interstress (rather than intersyllable) intervals (Couper-Kuhlen, 1993; Cummins & Port, 1998; Port, 2003).

Conversely, empirical assessments of syllable duration have largely bypassed meter, focusing instead on a wide range of rhythmic properties such as speaking rate (Miller, Green & Reeves, 1986; Miller & Volaitis, 1989; Kessinger & Blumstein, 1998), isochrony (Dauer, 1983; Dilley, 1997; Hoequist, 1983; Low, Grabe & Nolan, 2000; McAuley &

F. Benadon (✉)
Department of Performing Arts, American University,
4400 Massachusetts Ave, NW, Washington, DC 20016, USA
e-mail: fernando@american.edu

Dilley, 2004; Ramus & Mehler, 1999; Roach, 1982; Wenk & Wioland, 1982; Williams & Hiller, 1994), intelligibility (Bent, Bradlow & Smith, 2008; Quené & van Delft, 2010), morphosyntactic effects (Greenberg, Carvey, Hitchcock & Chang, 2003; Hamill, 1976; Palmer & Kelly, 1992; Scott, 1982; Turk & Shattuck-Hufnagel, 2000), dialect analysis (Clopper, Pisoni & de Jong, 2005; Jacewicz, Fox & Salmons, 2007), and music-language analogies (Hannon, 2009; Huron & Ollen, 2003; Patel & Daniele, 2003; Patel, Iversen & Rosenberg, 2006; Wenk, 1987). In a study by Fant, Kruckenberg and Nord (1991), four musicians used music notation to transcribe the syllable rhythms of recorded Swedish poems. The authors deemed music notation a viable tool for the study of speech rhythms but left meter unexplored.

What we are left with is a looming question mark regarding how listeners perceive syllabic durations when these are heard within a metrical framework. This is a relevant endeavor because it contextualizes the basic rhythmic unit of speech (the syllable) within an important principle of temporal organization (meter, conceived durationally). Hence, all syllabic sequences are "rhythmic" in that they consist of a series of onsets unfolding in time, whereas—as explained below—a sequence's "metrical" characteristic depends on how the onsets are configured.

Here we focus on four-syllable utterances cast in the pattern XxxX (stressed-unstressed-unstressed-stressed) in order to frame the individual onsets as an Abercrombian foot—a multi-syllabic, stress-initial construct that allows for internal units of variable duration (Abercrombie, 1964). The Abercrombian model therefore coincides with our present definition of meter: the partitioning of a timespan into equal subdivision slots, some of which may not receive an onset. For instance, using o's and _'s to denote sounding onsets and non-sounding subdivisions within the foot, the sequence o_oo(o) contains four subdivision slots of equal duration (a quadruple metric template), of which the second one is silent (the last onset in parentheses closes off the timespan); the interonset sequence can be represented as 2:1:1. Abercrombie (1964) proposed three ratio configurations that could make up the constituent durations of a disyllabic foot: 2:1, 1:2, and 1:1 (or 1.5:1.5 in Abercrombie's notation, in accordance with his foot isochrony convictions). In designing an early speech synthesis model, Witten (1977) expanded Abercrombie's list to include five trisyllabic configurations: 1:1:1, 2:1:1, 1:2:1, 2:2:1, and 2:1:2.[1] The metrically even partitioning just described does not imply that syllabic rhythm is isochronous. The conceptual advantage of meter in general is that it offers temporal reference templates against which rhythmic perception can be modeled.

In theory, it should not be difficult to determine the metrical organization of a given foot if one knows the onset times of its constituent syllables. For instance, a three-syllable string with onsets at 0, 300, and 600 ms consists of two even durations, giving rise to a duple meter. Alternatively, a triple meter rendition of the same string would find the middle syllable either at 200 or 400 ms, yielding a 1:2 or a 2:1 pattern, respectively. But there is no guarantee that the syllable's onset time corresponds to a rhythmic point of perceptual prominence. This is known as the perceptual center (P-center) problem (Morton, Marcus & Frankish, 1976; see Wallin 1901 for a review of late-nineteenth-century preoccupation with essentially the same problem). There is general agreement that the syllable's prevocalic portion largely determines the location of the P-center, with longer consonantal onsets proportionally delaying the P-center (Cooper, Whalen & Fowler, 1986; Howell, 1988). Other less local parameters may be responsible for some additional P-center drift, including the shape of the decay ramp (Scott, 1998), the duration of the entire vowel/syllable [Fox & Lehiste, 1987; Scott, 1993 (Experiment 8)], and the spectral envelope (Harsin, 1997; Howell, 1988; Pompino-Marschall, 1989). The combined empirical evidence suggests that the syllable's perceptual center coincides with a rapid amplitude increase in the center frequency region, generally at or very near the vowel onset (Cummins & Port, 1998; Fowler, 1983; Scott, 1993; Scott, 1998).

In any case, the reliance on monosyllabic stimuli reported in the P-center literature raises the question of whether the results are extendable to longer utterances. Possibly, vowel-onset centricity gives way to other cognitive factors when the stimuli consist of multi-syllabic strings of the kind used in this study and encountered in actual speech. What might these cognitive factors be? Chief among them is the possibility that syllabic rhythm may be evened out during perception, as has been reported for the perception of interstress intervals. Donovan and Darwin (1979) showed that listeners regularize their taps for speech—but not for noise—when asked to imitate interstress rhythms. The findings were replicated by Scott, Isard and de Boysson-Bardies (1985), who showed that English-speaking listeners regularize not only English sentences but also French (thereby calling into question the claim that English is perceived as stress-timed, whereas allegedly French is not). Participants regularized their interstress tapping to real and garbled speech patterns, but not to noise-burst patterns. The authors concluded that the acoustic complexity of the speech (or speech-like) signal was responsible for the regularization of taps.

Participants in the aforementioned studies were asked to tap the rhythms, which may have introduced sensorimotor

[1] Witten does not explain the omission of two other possible permutations of 1's and 2's—namely, the quadruple 1:1:2 and the quintuple 1:2:2.

constraints. Instead, participants in Darwin & Donovan (1980) adjusted the time intervals between click-like bursts in order to match the interstress intervals in the speech stimuli. The result was a regularization of click intervals. Also, Lehiste (1973) assessed the perception of interstress intervals by having participants underline the longest and shortest feet in four-foot pre-recorded sentences and analogous non-speech sequences. Performance was better with non-speech than with speech, which was interpreted by Lehiste (1977) as a perceptual tendency to regularize foot duration: "if you cannot tell them apart, they must be alike" (p. 257).

Knowing whether listeners regularize syllable duration in metrical contexts can shed light on future models of rhythm perception. The present experiment tested whether metrically framed syllable intervals are perceived as more regular than analogous intervals in tone sequences, a scenario supported by studies of interstress perception in speech. A related but separable question explored here pertains to variability in the perception of syllabic rhythm as compared to non-speech rhythm. On one hand, one could expect that the speech signal's greater complexity (acoustic and syntactic) would result in higher variability. On the other hand, this same complexity might lead to a rhythmic simplification by the listener, resulting in lower variability. The study concludes with an assessment of how the current findings inform the link between duration and meter.

Method

Participants

Thirty-seven members of the American University community took part in the experiment. Musical training ranged from 0 ($n = 1$) to 20 years (median = 10). All but one of the participants were native English speakers; the one exception learned English at a very young age and was highly fluent. Participants gave informed consent in accordance with American University's Institutional Review Board; they also received either a small payment or extra credit in a psychology course in return for their efforts.

Materials

Thirty phrases were recorded by an adult male speaker of General American English. The recordings were made with a Shure SM58 microphone placed near the speaker in a quiet room. Phrases were spoken with a natural intonation and speaking rate. Each phrase was four syllables long and had the stress pattern XxxX (e.g., Watch the cartoon). Apart from this shared feature, the grammatical and phonetic makeup of the phrases was varied. The mean interstress interval (ISI) was 562 ms, as measured from the vowel onset of the first syllable to the vowel onset of the fourth syllable. The complete list appears in the Appendix. For every speech phrase, there was an analogous "piano" version that used a pitched (440 Hz) tone modified from a MIDI piano timbre. Each of the four tones in the piano sequence had the same amplitude; note duration was 80 ms, with a 5 ms rise time and a 60 ms natural decay. The locations of the piano onsets were determined by the vowel onset locations of the corresponding speech phrases, which were marked manually by the author. It is important to emphasize that the piano stimuli were not intended to be exact rhythmic representations of the speech stimuli, since—as noted above in relation to P-centers—such a mapping cannot be determined with objective confidence. Rather, the piano stimuli offered a reasonably close rhythmic version of the speech stimuli, allowing us to draw comparisons between speech and non-speech rhythm perception. Each speech phrase was paired to a piano phrase for a total of 60 phrases (30 speech and 30 piano). In addition to the speech and piano stimuli, the matching task used a non-pitched percussive sound adapted from a MIDI agogo timbre. Its total duration was 30 ms, with instantaneous attack and fast decay.

Design

In order to keep the experiment from overwhelming the participants' concentration, the list of 30 speech–piano pairs was split into two random sub-lists of 15 pairs each. Nineteen participants heard list A and 18 participants heard list B. Thus, each participant heard 30 stimuli: 15 speech phrases and their 15 yoked piano versions. The order of trials was randomized for each participant.

Interface

An on-screen interface was designed using the graphical programming environment Max/MSP. Clicking on the button labeled *original* played either a speech phrase or a piano phrase (participants did not know which until after clicking the button). The *match* button played a percussive sound four times—the match sequence—with the following onset pattern. The first onset occurred at time 0 ms. The onset location of the fourth percussive sound was the same as that of the ISI of the phrase just heard (speech or piano). This ensured that the total interval of the match sequence was the same as that of the presented sequence. The original and match sequences could not be heard simultaneously. While the locations of the first and fourth percussive onsets were always fixed as just described,

participants could adjust the second and third onsets using two horizontal sliders, each containing a small handle. These middle adjustable onsets will be referred to as *a* and *b*, respectively. Onset *a* could not be placed before the first onset (at time = 0 ms) or after *b*; likewise, onset *b* could not be placed after the fourth onset (at time = ISI) or before *a*. This ensured that the sequence X*ab*X was always preserved.

Procedure

The task consisted of recreating speech and piano rhythms using two on-screen sliders that controlled the timing of a percussive rhythm. Sliding a slider's handle toward the left shifted the location of its corresponding onset (*a* or *b*) to an earlier time, in increments of 1 ms. Sliding a handle toward the right shifted the attack location to a later time, also in 1 ms increments. After setting the two sliders to the desired locations, participants clicked on the *match* button to hear the resulting percussive rhythm. The locations of *a* and *b* were set randomly and automatically (given the above constraints) at the beginning of each trial. Participants were told that at times they would be imitating the rhythm of a spoken phrase, whereas at other times they would be imitating the rhythm of a sequence of piano tones. (Strictly speaking, by "imitation" we are referring to the psychophysical method of adjustment.) Once the percussive match was deemed identical (or close to identical) to the original, participants moved to the next trial by clicking on the *next* button. They were allowed—indeed, encouraged—to tweak their match rhythms as much as they thought necessary in order to achieve a very strong rhythmic match. This often required several back-and-forth comparisons between the original and the match sequence, gradually adjusting the match with one or both sliders. For instance, if the second percussion onset sounded too early as compared to the second syllable (or tone) in the original, the participant nudged slider *a* to the right, then clicked on *match* to re-evaluate the rhythm. The session lasted about 25 min and consisted of 30 trials (15 speech and 15 piano) that were preceded by three additional practice trials, all involving speech. The concept of syllabic rhythm—the conversion of syllables to percussive onsets—was readily and intuitively understood by all participants. Participants wore headphones and were tested individually at a computer workstation in a quiet room.

Data

For each trial, the program recorded the millisecond locations of the second and third percussive onsets (*a* and *b*) as set with the sliders by the participant. Since the fixed fourth onset (the ISI) varied from phrase to phrase, its value was converted to 1.0, and the onset values of *a* and *b* were scaled accordingly between 0 and 1. In other words, the interval from the beginning of the sequence to the onset of the fourth percussive onset was scaled to 1.0, and the two internal onsets were expressed as proportions of this interval. This scaling procedure enabled comparisons across phrases. The data from the two groups were highly similar and were merged during data analysis. The program also counted the number of clicks on the *original* and *match* buttons.

Results

Speech responses underwent *assimilation*—the evening out of syllabic intervals to approximate three equal divisions of the foot. The tendency to even out the durations of the three syllables was specific to speech, as it was not evident in participants' assessments of analogous non-speech stimuli. The degree of assimilation was calculated as the absolute difference between the response onset vector (*a*, *b*) and the idealized even-distribution vector (.333, .667). Thus, a fully assimilated response would locate its two adjustable onsets one-third and two-thirds of the way along the ISI, respectively. Calculating the mean of each participant's absolute difference scores across phrases and comparing the results revealed that mean assimilation was significantly stronger for speech responses than for piano responses, $t(36) = 5.30$, $p < .0001$.

In order to confirm this finding, two additional measures of assimilation were taken on the interval spanning response onsets *a* and *b*: size = $b - a$, and midpoint = $(a + b)/2$; see Fig. 1. An assimilated response would approach .333 for size and .500 for midpoint. Respectively, the errors for size and midpoint are given by $e_{sz} = |\text{size} - .333|$ and $e_{mp} = |\text{midpoint} - .5|$, from which the differences Δ between speech and piano errors were calculated: $\Delta_{sz} = e_{sz(piano)} - e_{sz(speech)}$ and $\Delta_{mp} = e_{mp(piano)} - e_{mp(speech)}$. Assimilation therefore occurs when the difference between speech and piano *e* scores is large, pointing to a relative difference between the two modes of perception. A large positive Δ value denotes greater assimilation for speech, whereas a large negative value denotes greater assimilation for piano; values near zero correspond to little or no assimilation difference. One-sample *t* tests showed significant speech assimilation for size, $t(36) = 7.85$, $p < .0001$, and midpoint, $t(36) = 3.88$, $p < .001$. In other words, the speech *ab* interval was more likely to occupy one third of the foot, and it was more likely to drift toward the center of the foot.

Even though speech tended toward regularization, onset dispersion was greater for speech responses than for piano responses. This was reflected in the click counts: participants clicked the *match* button significantly more often

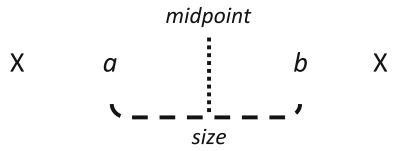

Fig. 1 Two measures of assimilation in the syllabic string X*ab*X. The interval between onsets *a* and *b* is given by *size*, equal to one third (.333) of the interstress interval (ISI) when perfectly assimilated. The half-way point between *a* and *b* is given by *midpoint*, equal to one half (.500) of the ISI when perfectly assimilated

during speech trials as compared to piano trials; $F(1, 36) = 8.74$, MSE = 3.28, $p < .005$. There were otherwise no main effects of button type—*original* vs. *match*—or condition, with median click count being the same (8) for each button across trials. For each stimulus and participant, onset dispersion was measured as the absolute difference between the participant's response onset location and the group's mean onset location. Figure 2 shows the mean dispersion across all phrases (speech and piano) and participants for onsets *a* and *b*. A two-way repeated measures ANOVA showed significant main effects for mode of presentation (speech or piano), $F(1, 36) = 28.21$, MSE = .005, $p < .0001$, as well as onset (*a* or *b*), $F(1, 36) = 24.31$, MSE = .002, $p < .0001$, with a significant interaction, $F(1, 36) = 22.12$, MSE = .001, $p < .0001$. Onset *b* was more dispersed than onset *a* in the speech condition but not in the piano condition; the reason for this disparity between the two onsets remains unclear.

The rhythmic skill of each participant was assessed by measuring his or her success in matching the piano stimuli (no such true values existed for the speech responses). *Skill* was calculated as 1 – *e*, where *e* equals the absolute error value from the correct (piano) stimulus, averaged on *a* and *b* across all piano responses. Each participant also received a *consistency* score of 1 – *d*, where *d* equals the absolute difference between his or her speech response and the group mean, averaged on *a* and *b* across all speech responses. A high consistency score means that a given participant's speech responses tended to lie near the group mean. Note that skill and consistency track performance in separate tasks: skill concerns piano responses (i.e., it measures how closely participants matched the true value of the piano sequence); consistency concerns speech responses (i.e., it measures how close participants were to the group mean of speech responses). Figure 3 shows the correlation between skill and consistency, $r(35) = .49$, $p < .01$. Participants who performed better on the piano trials were closer to the group's speech mean on the speech trials. This finding therefore refines Fig. 2, which showed—without taking skill into account—that speech means were more dispersed than piano means. Less so for rhythmically skilled participants. Skill also correlated highly with the proximity of speech responses to the location of vowel onsets, $r(35) = .66$, $p < .01$.

Syntax did not appear to be a primary influence on assimilation. One might expect syllables 3 and 4 to be perceived as closer together—a later *b*—when they are coupled, for instance by cliticizing or by forming part of the same word. Conversely, syllables 3 and 4 could be perceived as more separated—an earlier *b*—when there is a clear boundary before the last syllable, such as between two content words. Such timing effects are well documented in speech production studies (e.g., Turk & Shattuck-Hufnagel 2000, and references therein). Table 1 shows the location of onset *b* for phrases whose placement of this onset differed significantly ($p < .05$) between paired speech and piano responses. In 16 out of 18 cases, *b* was closer to .667 in speech than in piano, as predicted by the assimilation effect. The two exceptions were *Painless return* and *Rally the troops*, which would suggest that *b* occurred later as a result of the phrase's syntactic

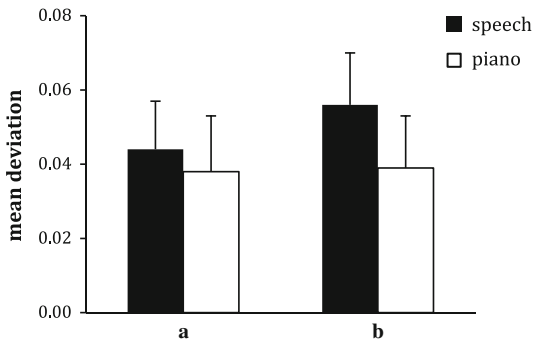

Fig. 2 Mean deviations from the group mean for onsets *a* and *b* were greater for speech than for piano responses. Error bars give standard deviations

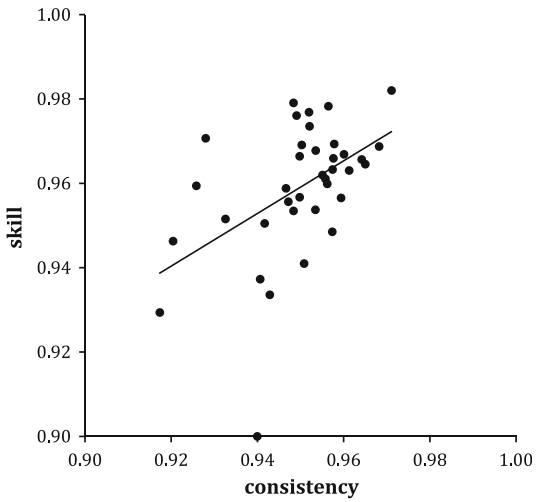

Fig. 3 Participants who were more accurate on the piano trials (high skill) were closer to the group mean on the speech trials (high consistency)

Table 1 Mean location of onset *b* when significantly different for speech vs. piano responses

	Speech	Piano	$p<$
Answer my plea	0.637	0.573	0.01
Delicate glass	0.577	0.509	0.02
Dissonant note	0.549	0.498	0.001
Don't blink an eye	0.741	0.770	0.03
Eat macaroons	0.722	0.773	0.01
Flawless design	0.713	0.789	0.001
Lenny's TV	0.692	0.744	0.01
Memory lane	0.609	0.551	0.02
Ned is so brave	0.655	0.623	0.03
Painless return	0.736	0.667	0.001
Paint it all green	0.556	0.510	0.03
Panic set in	0.711	0.740	0.03
Permanent bliss	0.539	0.463	0.01
Pick the right one	0.574	0.520	0.01
Rally the troops	0.689	0.660	0.05
Silly but true	0.632	0.470	0.001
Temple of doom	0.598	0.496	0.001
Watch the cartoon	0.656	0.616	0.03

structure (*re-turn* and *the troops*). However, several other phrases on the list provide counterexamples. Compared to the paired piano responses, *b* occurred earlier in the speech responses to *Flawless design*, *Eat macaroons*, *Lenny's TV*, and *Don't blink an eye*, even though a later *b* would be expected owing to the word membership of *de-sign*, *-a-roons*, and *T-V*, and the clitic group *an eye*. Moreover, *b* occurred later in speech responses to phrases with stronger final boundaries—examples include *Delicate glass*, *Dissonant note*, *Memory lane*, and *Permanent bliss*. It may be concluded from Table 1 that the boundary strength between syllables 3 and 4 did not determine (primarily) the behavior of onset *b* in the speech responses. However, since the study's selection of phrases did not control for syntax and possible word segmentation effects, further investigation is required to tease out possible confounds.

Discussion

The perceptual equalization of syllabic intervals observed here is reminiscent of analogous findings for speech interstress intervals, as noted above (Darwin & Donovan, 1980; Donovan & Darwin, 1979; Lehiste, 1973; Scott et al., 1985). Also relevant is a study by Repp, Windsor and Desain (2002) in which musicians performed piano melodies at increasingly faster tempos; all permutations of the 1:2:3 ratio pattern were tested. Assimilation occurred at fast tempos between the two long intervals, while the short interval remained proportionally constant.[2]

In addition to undergoing assimilation, speech rhythms elicited higher response variability than non-speech sequences. This is not entirely surprising, since the piano stimuli consisted of uniform acoustic events separated by silence, while the speech stimuli traversed a continuum of timbres and amplitudes with no clearly defined segmental boundaries. The syntactical and acoustic complexities of speech may have precluded the formation of clearly defined rhythmic nuclei, leading different participants to assign onset markers to different sub-regions of the same speech sequence. Furthermore, time interval estimation has been shown to increase as a result of stimulus complexity regardless of modality, with complexity being measured in terms of the difficulty of the task (Fraisse, 1956), the amount of sound "filler" contained in the interval (Repp, 2008; Thomas & Brown, 1974; Wearden, Norton, Martin, & Montford-Bebb, 2007), the number of perceived changes in the stimulus (Block, 1978; Schiffman & Bobko, 1974), and the amount of memory storage (Ornstein, 1969) or neural energy needed to code the information (Pariyadath & Eagleman, 2007). Speech phrases may have been perceived as longer than their piano counterparts, thereby presenting more possibilities for the selection of onset locations. Alternatively, it could have been the case that the phonetic complexity of speech, coupled with shared knowledge of syntactical rules, could have led to more rigid interpretations of the speech rhythm. But this was not the case. The results agree with Allen's (1972) conclusion that the syllable beat is like a "broad slur" rather than a single point in time.

A participant's level of rhythmic skill, as measured by his or her ability to mimic the piano stimuli accurately, influenced the dispersion in speech responses, with more skilled participants showing lower variability. Why would this be the case? The data showed that the level of rhythmic skill was significantly correlated with how close speech responses were to the vowel onset locations in the speech stimuli. This suggests that skilled listeners were more consistent with their speech responses because they may have relied on vowel onsets as reference cues. If, as the P-center literature suggests, vowel onsets are the most reliable markers for a syllable's "moment of occurrence," then the claim applies more appropriately to rhythmically skilled listeners. However, assimilation was not influenced by skill, indicating that skill only modulates certain aspects of syllabic rhythm perception.

[2] Yet, the opposite occurred in Repp, London & Keller (2005), where musicians tapped all permutations of the metrically complex patterns 2:2:3 and 2:3:3. The two nominally different ratios were now contrasted toward a 2:1 (or 1:2) ratio.

The observed assimilation was not strong enough to coalesce the speech responses into an isochronous 1:1:1 template. Quite the contrary. The duration patterns of speech responses exhibited much variety, with different phrases displaying different onset distributions. It helps to visualize the response data with the triangular graph shown in Fig. 4 (adapted from Desain & Honing, 2003). The three coordinates correspond to the three temporal intervals in the phrase, scaled from 0 to 1. The two measures of assimilation discussed above, size and midpoint, correspond, respectively, to vertical and horizontal motion along the graph. This is illustrated by the four reference points corresponding to simple metrical configurations (music notation equivalents are provided for reference). Hence, an intuitive measure of similarity between two rhythms (points) is given by their proximity in the graph.

Figure 5 zooms into the graph's center to show responses for three speech–piano pairs. While the piano responses for *Answer my plea* cluster around the long–short–long pattern of the piano stimulus (the filled square), the speech responses lie higher, around the 1:1:1 landmark—a clear size assimilation of the *ab* interval. Assimilation of *ab*'s midpoint cannot be assessed for this phrase because the piano stimulus already lies roughly halfway between the left and right endpoints of the graph. In *Flawless design*, speech points lie to the left of piano points—the *ab* interval has shifted horizontally toward the midpoint of the graph. Both types of assimilation are visible in *Temple of doom*: the speech points have shifted upward and to the right with respect to the piano points.

Notwithstanding the general tendency toward assimilation, speech responses across phrases covered a lot of rhythmic terrain, suggesting that the rhythmic organization of the foot can be perceived in a number of metric configurations. This is illustrated in Fig. 6, which gives examples of

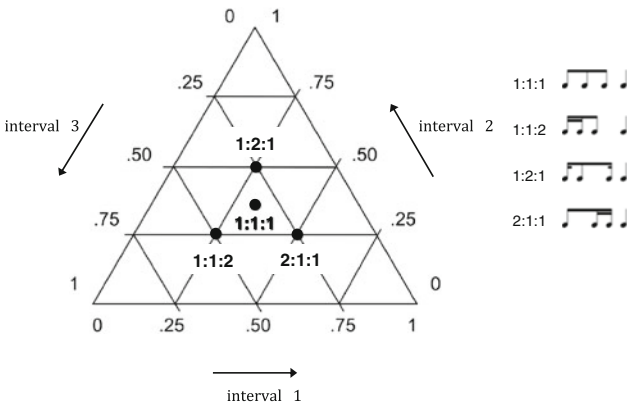

Fig. 4 Triangular representation of a three-interval sequence. Each axis plots the size of one of the three intervals, scaled between 0 and 1. The surface is continuous, so any configuration of three successive durations may be plotted. The dots mark the locations of four simple metrical configurations

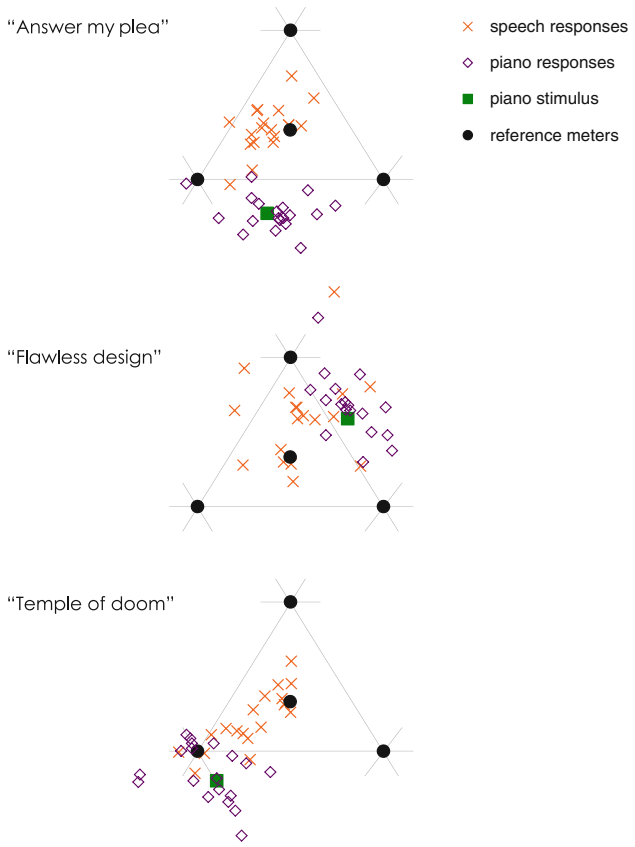

Fig. 5 Response graphs for three phrase pairs. Assimilation is visible in the speech responses

speech responses that approximate two instantiations of quadruple meter, *Yellowstone Park* (1:1:2) and *Eat macaroons* (2:1:1), as well as one of quintuple meter, *Sweep the canteen* (2:1:2). In the first two examples, speech points seem to have migrated toward the quadruple meter and away from regions containing cognitively more complex meters: quintuple (1:2:2) in the case of *Yellowstone Park*, and quintuple (2:2:1) and sextuple (3:2:1) in the case of *Eat macaroons*. Does this mean that listeners mold speech rhythms into simpler metrical categories? The cognitive salience of simple-ratio rhythms has been extensively documented (Barnes & Jones, 2000; Collier & Wright, 1995; Desain & Honing, 2003; Drake & Botte, 1993; Essens & Povel, 1985; Fitch & Rosenfeld, 2007; Fraisse, 1956; Grube & Griffiths, 2009; Hannon & Johnson, 2005; Keller & Repp, 2005; Martin, Deltenre, Hoonhorst, Markessis, Rossion & Colin, 2007; Palmer & Krumhansl, 1990; Povel, 1981; Hannon, 2009). But it remains unclear from the current results whether the effect is as pronounced and widespread in speech. The presence of the quintuple pattern 2:1:2 just mentioned might well be a viable metric template in the perception of speech rhythms. If so, the resulting diversity of rhythmic configurations leads to the conjecture that speech contains "mixed" meters. For

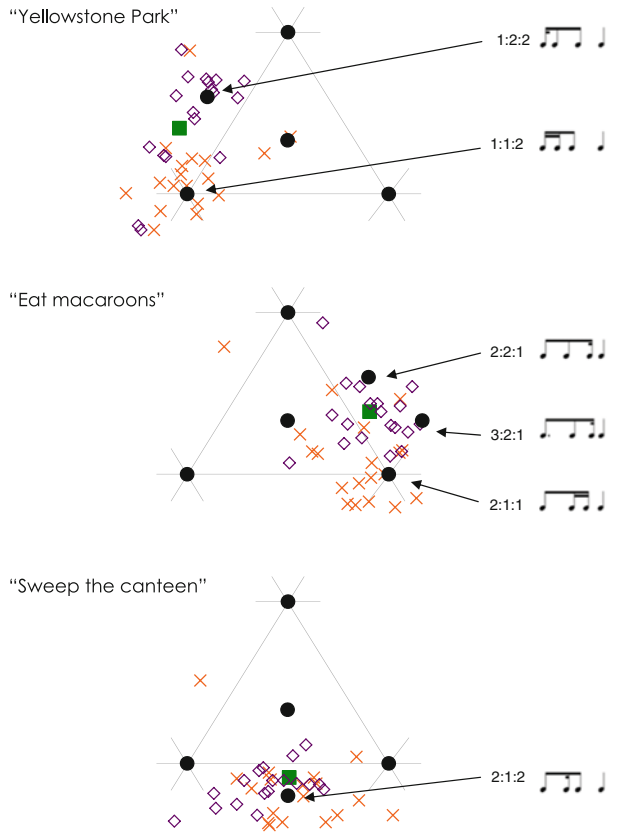

Fig. 6 Quadruple (*top two*) and quintuple (*bottom*) meters in speech responses

instance, a sequence of three trisyllabic feet might contain three different meters such as quadruple → triple → quintuple (Benadon, 2009). Mixed meters (also known as non-isochronous meters) are the norm in numerous non-Western musical traditions (London, 2004) and in the music of many twentieth-century composers.

The experiment results revealed that rhythmic assimilation was accompanied by high variability in locating syllable onsets. These seemingly contradictory trends (assimilation and variability) were likely due to the continuously evolving nature of the speech signal. That is, the assimilation bias may have functioned as a cognitive stabilizer during the somewhat intractable task of pinpointing syllable onsets in the speech sequence. Also, it is important to note that a perceptual tendency toward even ratios is not the same as the perception of an even rhythm in triple meter. Assimilation was strong enough to elicit different response trends in speech versus non-speech, and weak enough to preserve a varied set of speech response configurations. Assessing each participant's rhythmic skill served as a reminder that the complexities of speech rhythm perception are mediated by individual proficiency. While previous speech rhythm studies have commented on between-participant differences, these differences could not be attributed to individual rhythmic skill because it had not been gauged. The results of this study show that rhythmically skilled participants were more likely to assess syllabic rhythm according to vowel onset location, thereby reducing the variability of their responses. This finding adds a new dimension to the P-center problem by emphasizing that the perception of syllabic onsets is dependent on the listener's rhythmic aptitude.

Acknowledgments This study was supported by a Faculty Research Award from American University's College of Arts & Sciences. The author also wishes to thank Scott Parker, Zehra Peynircioglu, and Jun Lu for their feedback, and Eric James Schmidt for recording the speech stimuli.

Appendix

Phrase	ISI (ms)
Answer my plea[a]	621
Corporate desk	520
Delicate glass[a]	503
Diesel device	521
Dissonant note[a]	530
Don't blink an eye	616
Eat macaroons[a]	568
Flawless design	579
Gatorade rocks[a]	498
Gone with the wind[a]	547
Greeting his friend	588
Incoming mail[a]	524
Lenny's TV	602
Light the cigar[a]	413
Memory lane	500
Mexican food[a]	519
Mystical place	547
Ned is so brave[a]	615
Painless return[a]	656
Paint it all green	580
Panic set in[a]	593
Permanent bliss	574
Pick the right one	593
Rally the troops[a]	513
Silly but true	628
Sweep the canteen	653
Temple of doom[a]	479
Washing machine	553
Watch the cartoon[a]	609
Yellowstone Park	617

[a] List A; all others list B

References

Abercrombie, D. (1964). Syllable quantity and enclitics in English. In D. Abercrombie, et al. (Eds.), *In honour of Daniel Jones: Papers contributed on the occasion of his eightieth birthday, 12 September 1961*. London: Longman.

Allen, G. D. (1972). The location of rhythmic stress beats in English: An experimental study. *Language and Speech*, *15*, 72–100, 179–195.

Attridge, D. (1982). *The rhythms of English poetry*. London: Longman.

Barnes, R., & Jones, M. R. (2000). Expectancy, attention, and time. *Cognitive Psychology, 41*, 254–311.

Benadon, F. (2009). Speech rhythms and metric frames. *Proceedings of the Second International Conference of the Society for Mathematics and Computation in Music*, New Haven (pp. 22–31).

Bent, T., Bradlow, A. R., & Smith, B. L. (2008). Production and perception of temporal patterns in native and non-native speech. *Phonetica, 65*(3), 131–147.

Block, R. A. (1978). Remembered duration: Effects of event and sequence complexity. *Memory & Cognition*, *6*(3), 320–326.

Clopper, C. G., Pisoni, D. B., & de Jong, K. (2005). Acoustic characteristics of the vowel systems of six regional varieties of American English. *Journal of the Acoustical Society of America, 118*, 1661–1676.

Collier, G. L., & Wright, C. E. (1995). Temporal rescaling of simple and complex ratios in rhythmic tapping. *Journal of Experimental Psychology: Human Perception and Performance, 21*(3), 602–627.

Cooper, W. E., & Eady, S. J. (1986). Metrical phonology in speech production. *Journal of Memory and Language, 25*, 369–384.

Cooper, A. M., Whalen, D. H., & Fowler, C. A. (1986). P-centers are unaffected by phonetic categorization. *Perception & Psychophysics, 39*, 187–196.

Couper-Kuhlen, E. (1993). *English speech rhythm: Form and function in everyday verbal interaction*. Amsterdam: John Benjamins Publishing Co.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics, 26*, 145–171.

Cureton, R. D. (1992). *Rhythmic phrasing in English verse*. London: Longman.

Cutler, A. (1990). Exploiting prosodic probabilities in speech segmentation. In G. T. M. Altman (Ed.), *Cognitive models of speech processing* (pp. 105–121). Cambridge: MIT Press.

Darwin, C. J., & Donovan, A. (1980). Perceptual studies of speech rhythm: Isochrony and intonation. In J. C. Simon (Ed.), *Spoken language generation and understanding* (pp. 77–85). Dordrecht: D. Reidel.

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

Desain, P., & Honing, H. (2003). The formation of rhythmic categories and metric priming. *Perception, 32*, 341–365.

Dilley, L. C. (1997). Some factors influencing duration between syllables judged perceptually isochronous. *Journal of the Acoustical Society of America, 102*, 3205–3206.

Donovan, A., & Darwin, C. J. (1979). The perceived rhythm of speech. *Proceedings of the Ninth International Congress of Phonetic Sciences*, Copenhagen, Vol. 2 (pp. 268–274).

Drake, C., & Botte, M. C. (1993). Tempo sensitivity in auditory sequences: Evidence for a multiple-look model. *Perception & Psychophysics, 54*, 277–286.

Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language, 36*, 202–225.

Essens, P., & Povel, D. (1985). Metrical and nonmetrical representations of temporal patterns. *Perception & Psychophysics, 37*, 1–7.

Fabb, N., & Halle, M. (2008). *Meter in poetry: A new theory*. Cambridge: Cambridge University Press.

Fant, G., Kruckenberg, A., & Nord, L. (1991). Stress patterns and rhythm in the reading of prose and poetry with analogies to music performance. In J. Sundberg, L. Nord, & R. Carlson (Eds.), *Music, language, speech and brain* (pp. 380–407). Stockholm: MacMillan Press.

Fitch, W. T., & Rosenfeld, A. J. (2007). Perception and production of syncopated rhythms. *Music Perception, 25*(1), 43–58.

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

Fox, R. A., & Lehiste, I. (1987). The effect of vowel quality variations on stress-beat location. *Journal of Phonetics, 15*, 1–13.

Fraisse, P. (1956). *Les structures rythmiques*. Louvain: Publications Universitaires de Louvain.

Gerken, L. A. (1991). The metrical basis for children's subjectless sentences. *Journal of Memory and Language, 30*(4), 431–451.

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

Greenberg, S., Carvey, H., Hitchcock, L., & Chang, S. (2003). Temporal properties of spontaneous speech: A syllable-centric perspective. *Journal of Phonetics, 31*, 465–485.

Grube, M., & Griffiths, T. D. (2009). Metricality-enhanced temporal encoding and the subjective perception of rhythmic sequences. *Cortex, 45*, 72–79.

Halle, M., & Keyser, S. J. (1971). *English stress: Its form, its growth, and its role in verse*. New York: Harper and Row.

Halle, M., & Vergnaud, J. R. (1990). *An essay on stress*. Cambridge: MIT Press.

Halle, J., & Lerdahl, F. (1993). A generative textsetting model. *Current Musicology*, *55*, 3–23.

Hamill, B. W. (1976). A linguistic correlate of sentential rhythmic patterns. *Journal of Experimental Psychology: Human Perception and Performance, 2*(1), 71–79.

Hannon, E. E. (2009). Perceiving speech rhythm in music: Listeners classify instrumental songs according to language of origin. *Cognition, 111*, 403–409.

Hannon, E. E., & Johnson, S. P. (2005). Infants use meter to categorize rhythms and melodies: implications for musical structure learning. *Cognitive Psychology, 50*, 354–377.

Harsin, C. A. (1997). Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception & Psychophysics, 59*, 243–251.

Hayes, B. (1995). *Metrical stress theory: Principles and case studies*. London: University of Chicago Press.

Hayes, B. (2009). Textsetting as constraint conflict. In J.-L. Aroui & A. Arleo (Eds.), *Towards a typology of poetic forms: From language to metrics and beyond* (pp. 43–62). Philadelphia: John Benjamins Publishing Co.

Hoequist, C. E. (1983). Syllable duration in stress-, syllable- and mora-timed languages. *Phonetica, 40*, 203–237.

Hollander, J. (1989). *Rhyme's reason*. New Haven: Yale University Press.

Howell, P. (1988). Prediction of P-centre location from the distribution of energy in the amplitude envelope. *Perception and Psychophysics, 43*, 90–93.

Huron, D., & Ollen, J. (2003). Agogic contrast in French and English themes: Further support for Patel and Daniele (2003). *Music Perception, 21*, 267–271.

Jacewicz, E., Fox, R. A., & Salmons, J. (2007). Vowel duration in three American English dialects. *American Speech, 82*(4), 367–385.

Keller, P. E., & Repp, B. H. (2005). Staying offbeat: sensorimotor syncopation with structured and unstructured auditory sequences. *Psychological Research, 69*, 292–309.

Kelly, M. H., & Rubin, D. C. (1988). Natural rhythmic patterns in English verse: Evidence from child counting-out rhymes. *Journal of Memory and Language, 27*(6), 718–740.

Kessinger, R. H., & Blumstein, S. E. (1998). Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. *Journal of Phonetics, 26*, 117–128.

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

Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America, 54*, 1228–1234.

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

Liberman, M. (1975). *The intonational system of English.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge.

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

London, J. (2004). *Hearing in time: Psychological aspects of musical meter*. New York: Oxford University Press.

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

Martin, X. P., Deltenre, P., Hoonhorst, I., Markessis, E., Rossion, B., & Colin, C. (2007). Perceptual biases for rhythm: The Mismatch Negativity latency indexes the privileged status of binary vs non-binary interval ratios. *Clinical Neurophysiology, 118*, 2709–2715.

McAuley, J. D., & Dilley, L. C. (2004). Acoustic correlates of perceived rhythm in spoken English. *Journal of the Acoustical Society of America, 115*, 2397–2398.

Miller, J. L., Green, K. P., & Reeves, A. (1986). Speaking rates and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica, 43*, 106–115.

Miller, J. L., & Volaitis, L. E. (1989). Effect of speaking rate on the perceptual structure of a phonetic category. *Perception and Psychophysics, 46*(6), 505–512.

Morgan, J. L. (1996). A rhythmic bias in preverbal speech segmentation. *Journal of Memory and Language, 35*, 666–688.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (P-centers). *Psychological Review, 83*, 405–408.

Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication, 41*, 233–243.

Ornstein, R. E. (1969). On the experience of time. Harmondsworth: Penguin.

Palmer, C., & Kelly, M. H. (1992). Linguistic prosody and musical meter in song. *Journal of Memory and Language, 31*(4), 525–542.

Palmer, C., & Krumhansl, C. L. (1990). Mental representations for musical meter. *Journal of Experimental Psychology: Human Perception and Performance, 16*, 728–741.

Pariyadath, V., & Eagleman, D. (2007). The effect of predictability on subjective duration. *PLoS One*, *2*(11), e1264.

Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition, 87*, B35–B45.

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. *Journal of the Acoustical Society of America, 119*, 3034–3047.

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

Port, R. F. (2003). Meter and speech. *Journal of phonetics, 31*, 599–611.

Povel, D. J. (1981). Internal representation of simple temporal patterns. *Journal of Experimental Psychology: Human Perception and Performance, 7*(1), 3–18.

Prince, A. (1989). Metrical forms. In P. Kiparsky & G. Youmans (Eds.), *Phonetics and phonology: Rhythm and meter* (pp. 45–80). San Diego: Academic Press.

Quené, H., & van Delft, L. E. (2010). Non-native durational patterns decrease speech intelligibility. *Speech Communication*, *52*(11-12), 911-918.

Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech resynthesis. *Journal of the Acoustical Society of America, 105*(1), 512–521.

Repp, B. H. (2008). Metrical subdivision results in subjective slowing of the beat. *Music Perception, 26*(1), 19–39.

Repp, B. H., London, J., & Keller, P. E. (2005). Production and synchronization of uneven rhythms at fast tempi. *Music Perception, 23*(1), 61–78.

Repp, B. H., Windsor, W. L., & Desain, P. (2002). Effects of tempo on the timing of simple musical rhythms. *Music Perception, 19*(4), 565–593.

Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (Ed.), *Linguistic Controversies* (pp. 73–79). London: Arnold.

Rodríguez-Vázquez, R. (2010). Text-setting constraints: A comparative perspective. *Australian Journal of Linguistics, 30*(1), 19–34.

Rothermich, K., Schmidt-Kassow, M., & Kotz, S. A. (2012). Rhythm's gonna get you: Regular meter facilitates semantic sentence processing. *Neuropsychologia, 50*, 232–244.

Schiffman, H. R., & Bobko, D. J. (1974). Effects of stimulus complexity on the perception of brief temporal intervals. *Journal of Experimental Psychology*, *103*(1), 156–159.

Scott, D. R. (1982). Duration as a cue to the perception of a phrase boundary. *Journal of the Acoustical Society of America, 71*(4), 996–1007.

Scott, S. K. (1993). *P-centres in speech: An acoustic analysis*. Unpublished doctoral dissertation, University College, London.

Scott, S. K. (1998). The point of P-centres. *Psychological Research, 61*, 4–11.

Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and in French. *Journal of Phonetics, 13*, 155–162.

Selkirk, E. O. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge: The MIT Press.

Thomas, E. A. C., & Brown, I. (1974). Time perception and the filled-duration illusion. *Perception and Psychophysics, 16*, 449–458.

Turk, A. E., & Shattuck-Hufnagel, S. (2000). Word-boundary-related duration patterns in English. *Journal of Phonetics, 28*, 397–440.

Wallin, J. E. W. (1901). Researches on the rhythm of speech. In E. W. Scripture (Ed.), *Studies from the Yale Psychological Laboratory* (Vol. 9). New Haven: Yale University.

Wearden, J. H., Norton, R., Martin, S., & Montford-Bebb, O. (2007). Internal clock processes and the filled-duration illusion. *Journal of Experimental Psychology: Human Perception and Performance, 33*(3), 716–729.

Weismiller, E. R. (1989). Triple threats to duple rhythm. In P. Kiparsky & G. Youmans (Eds.), *Phonetics and phonology: Rhythm and meter* (pp. 261–290). San Diego: Academic Press.

Wenk, B. J. (1987). Just in time: On speech rhythms in music. *Linguistics, 25*(5), 969–982.

Wenk, B. J., & Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics, 10*, 193–216.

Williams, B., & Hiller, S. M. (1994). The question of randomness in English foot timing: A control experiment. *Journal of Phonetics, 22*, 423–439.

Winn, M. B., & Idsardi, W. J. (2008). Musical evidence regarding trochaic inversions. *Language & Literature, 17*(4), 335–349.

Witten, I. H. (1977). A flexible scheme for assigning timing and pitch to synthetic speech. *Language and Speech, 20*(3), 240–260.