Laura C. Dilley, Jessica Wallace & Christopher C. Heffner
(Lansing, USA)

Perceptual isochrony and fluency in speech by normal talkers under varying task demands

1 Introduction

The term *isochrony* refers to the concept that certain phonological units (i.e., stresses) occur at approximately equal time intervals (Lehiste 1977; Pike 1945). Over time, a sizeable body of research has largely failed to find evidence supporting the notion of acoustic isochrony, i.e., physical regularity of time intervals between stresses (Bolinger 1965; Classe 1939; Cummins 2003, 2005; Dauer 1983; Kohler 1986; Roach 1982; Shen and Peterson 1962; Uldall 1971). Instead, isochrony has for some time been recognized as primarily a perceptual phenomenon in which phonological units such as stresses are merely heard as occurring at regular intervals (Lehiste 1977). There has been controversy over whether listeners perceive more regularity in speech than in other kinds of auditory stimuli (e.g., tone sequences, music), as well as why such differences in perception might occur (Scott, Isard and de Boysson-Bardies 1985; White 2002). The present work will not deal with these controversies but rather will focus on quantifying how often the commonly-reported but little-studied phenomenon of perceptual isochrony occurs in spoken language. Quantifying the occurrence of perceptual isochrony is a critical issue given a newly-identified role for perceptual isochrony in contributing to processing of spoken language, discussed below.

Recent evidence suggests that perceptual isochrony plays a role in two important psycholinguistic processes underlying spoken word recognition: word segmentation and lexical access (Brown, Salverda, Dilley and Tanenhaus 2011; Dilley, Mattys and Vinke 2010; Dilley and McAuley 2008). This role for perceptual isochrony has been demonstrated using word lists ending with lexically ambiguous syllable sequences which could either form compound words (foot-note-book-worm → footnote bookworm, foot notebook worm, etc.; see Dilley and McAuley 2008), or else more subtly end-embedded words (/kraIsIstɝnIp/ → crisis turnip, cry sister nip; see Dilley et al., 2010) and even grammatical sentences with embedded lexical ambiguities (e.g., …that kidney do[ctor]… vs. …that kid knee Do[nald]…; see Brown et al. 2011). The task in several experiments was to report the final word in each experimental sequence containing the lexical ambiguity. When the F0 and duration characteristics of the words in the distal (i.e., nonlocal) context of to-be-judged material formed rhythmic patterns, these patterns influenced listeners' judgments of the identity of the final word (Dilley & McAuley 2008). The more rhythmically regular the distal context prosody based on the combination of F0 and duration cues – that is, the more perceptually isochronous the distal context – the larger the effects on word segmentation; the largest effects were observed when the rhythmic regularity was carried by both F0 and duration cues (Dilley et al. 2010; Dilley and McAuley 2008). The effect of the perceived rhythmic regularity of prosodic context has now been demonstrated with multiple methodologies, including a surprise test-recognition paradigm (Dilley and McAuley 2008), as well as on-line tasks including lexical decision using cross-modal identity priming (Dilley et al. 2010), and eye tracking (Brown et al. 2011).

We speculate that perceptual isochrony plays a role in helping listeners to efficiently process lexical content in speech: given a rhythmically regular (i.e., perceptually isochronous) context, listeners will attend more readily to specific, predicted time intervals, such as stressed syllables (Large and Jones 1999; McAuley 1995; Pitt and Samuel 1990). Indeed, rhythm has been suggested to be a 'guide' for speech processing (Barry 1981; Pitt and Samuel 1990), and listeners can readily adjust their speech to sound more perceptually isochronous, though not necessarily more acoustically isochronous (Studdert-Kennedy 1980). Perceptual isochrony, then, could help listeners identify where stressed syllables occur in upcoming speech material (e.g., Cutler and Norris 1988), and thus aid in speech processing, since stressed syllables are statistically likely to form word onsets in English (Cutler and Carter 1987). Stressed syllables also occur in predictable locations in a word in most languages (Kager 1995) such that the location of stressed syllables has been proposed as a reliable cross-linguistic cue to word segmentation (Norris, McQueen, Cutler, Butterfield and Kearns 2001). Rhythmic regularity among phonological units in speech (e.g., stressed syllables) is hypothesized to arise from endogenous oscillators with a given inherent period of oscillation which can adapt to (i.e., entrain to) timing in the environment (Barbosa 2007; Cummins and Port 1998; McAuley 1995; Port 2003; Saltzman, Nam, Krivokapic and Goldstein 2008).

As a result of the role of perceptual isochrony of context prosody in psycholinguistic processing, the question arises of how often perceptual isochrony is present in speech and thus the extent to which it could serve as a useful cue in spoken word recognition. The rhythm of speech for a given utterance is influenced by a variety of linguistic features, including syntax, semantics, and pragmatics; however, certain contexts, such as repetitions of simple phrases (Cummins and Port 1998) and lists (see Cummins 2003, 2005 for studies of acoustic timing of intervals in lists) anecdotally seem to foster the use of perceptual isochrony. We therefore investigated the extent to which simple word lists are produced in a perceptually isochronous manner, as a starting point to investigating lexical contexts where perceptual isochrony is particularly likely to be of use to listeners in processing spoken language.

To investigate production of perceptually isochronous speech under a range of conditions, we utilized a series of tasks which were expected to place various degrees of demand on speech planning and/or cognitive faculties. In each task, participants were asked to recite a word list that was short by design, consisting of five semantically related monosyllabic words. The first task was a simple reading task; reading was expected to place relatively low demands on memory and speech planning. The second task was a memorization condition in which participants studied the word list before reproducing it from memory. Because the lists were relatively short, this condition was expected to place low demands on memory and speech planning processes. Finally, a third, dual task procedure required participants to first memorize a word list, and then to recite the list while simultaneously processing a visually-presented sentence. This latter task was expected to place substantial speech planning, memory, and attentional demands on talkers, thereby interfering with normal fluent speech production processes (Livant 1963), and resulting in speech productions which less frequently sounded isochronous.

We hypothesized that perceptually isochronous speech would be more likely to occur when talkers are under relatively low task demands; under such conditions, focused attention should make it fairly easy to entrain to the temporal regularity of oscillators that are hypothesized to underlie productions of stressed syllable strings (Barbosa 2007; Cummins and Port 1998; McAuley 1995; Port 2003; Saltzman, Nam, Krivokapic and Goldstein 2008), leading to a well-coordinated, fluent sequence of stresses. In contrast, for more challenging task conditions, demands on memory recall and/or motor execution of syllables should result in localized delays (perturbations) to oscillator behavior, resulting in localized disfluencies. The possibility of eliciting disfluent speech through the differing demands of the task conditions for lexical sequences further permitted exploration of the relationships between perception of isochrony and perception of fluency. Though a core behavior of disfluent speech (e.g., stuttering) is sound prolongation (Eklund 2004; Guitar 2005; Van Riper 1982) – the main correlate of which is timing – the manner by which timing contributes to perception of fluent speech is not well understood (Amir and Yairi 2002;

Kawai, Healey and Carrell 2007). Thus, it is not obvious that localized disfluencies necessarily disrupt perception of global rhythmic regularity.

By carrying out a production experiment in which speech was elicited under distinct task conditions which were expected to range from quite easy to quite hard, we anticipated that we would obtain utterances which sounded perceptually isochronous and/or fluent to varying degrees. Specifically, we hypothesized that the dual task would induce the least perceptual isochrony as well as the least fluency. Overall, we predicted that speech that was perceived as isochronous would also be perceived as fluent, and that speech perceived as anisochronous would also be perceived as disfluent. It is by no means a foregone conclusion that such a close relationship between perception of isochrony and perception of fluency would be predicted. Summarizing a wide range of the phonetics literature, Patel (2008:159) states *"…languages have rhythm…[but] this rhythm does not involve the periodic recurrence of stresses, syllables, or any other linguistic unit"*. Based on observations about temporal variability in acoustic measurements of stresses in speech (Bolinger 1965; Classe 1939; Cummins 2003, 2005; Dauer 1983; Kohler 1986; Roach 1982; Shen and Peterson 1962; Uldall 1971), as well as the heterogeneity of rhythmic structures that characterize spoken language (as opposed to say, rap music; see Patel 2008), it's not clear a priori that the temporal variation introduced by local deviations from fluency would disrupt global utterance rhythm. Indeed, it's not clear that localized deviations from fluency would have any relationship whatsoever to perceived global speech rhythm, particularly given evidence that highly disfluent speech can show regularized, harmonic timing patterns (Tilsen 2006) and that relatively small rhythmic units (i.e., the interstress interval or prosodic foot) comprise the domain for compensatory adjustments in timing (Fant, Kruckenberg and Nord 1991).

Three perception experiments and an acoustic study were then conducted to assess the extent to which monosyllabic word lists sounded perceptually isochronous and/or fluent, as well as the relationship of these perceptions to acoustic timing. First, a perception study was carried out in which the speech was evaluated for perceptual isochrony by two prosodic analysts that had previously been trained in the use of the Rhythm and Pitch (RaP) prosody labeling system (Breen, Dilley, Kraemer and Gibson, in press; Dilley, Breen, Bolivar, Kraemer and Gibson 2006; Dilley and Brown 2005). Second, a perceptual experiment was conducted in which a group of naïve listeners evaluated the degree of rhythmicity of recordings produced under each task condition. Next, a perceptual experiment was conducted in which a separate group of listeners evaluated the degree of fluency of the same recordings across conditions. Finally, an acoustic study was carried out to assess how the perception of speech rhythm related to acoustic timing of intervals between vowel onsets of syllables.

2 Production experiment

To assess the relative frequency of perceptual isochrony in word lists, as well as the relationships among perceptual isochrony, perceived fluency, and speech timing, thirty lists of words were constructed. Each list consisted of five monosyllabic content words. The words were semantically related, consistent with the fact that words which occur in lists in everyday life typically share a meaning relationship (baking supplies to purchase at the grocery store, types of animals at the zoo, etc.). (See Appendix for lists in the study.) All lists contained syllables with C(C)VC(C) patterns in order to facilitate identification of vowel onsets for acoustical measurements. As the lists were relatively homogeneous, consisting solely of stressed monosyllabic words, it was expected there would be low inter-list variability in the variables measured. The lists were further counterbalanced across participants and conditions to ensure the lexical content of word lists did not affect final results. Participants in the production experiment were nine female talkers aged 22 to 26, all graduate students at Michigan State University. All had self-reported normal hearing and speech ability.

There were three within-subjects production conditions: a Reading condition, a Memory condition, and a Dual task condition; see descriptions of these conditions below. Word lists will be referred to here as items; items were counterbalanced across conditions as follows. First, the thirty items were divided into three groups of ten items (Groups A-C in the appendix) in a fixed within-group order. The order of these three groups of items was rotated using a Latin Square design; that is, the three groups of items were rotated in such a way that each group appeared in each position of presentation equally often (i.e., ABC, BCA, CAB); see Grant (1948) and McNemar (1951) for introductions to the use of Latin Squares in behavioral experiments. This resulted in creation of three different orders of groups. Moreover, three different orders of the task conditions were created, also using a Latin Square design. The three different orders of groups of items were paired with the three different orders of task conditions, yielding nine possible pairings of word lists with task conditions (i.e., three list orders times three task condition orders). Each of the nine participants was assigned to a different pairing of list and task condition order, with one participant per pairing.

In the Reading condition, participants read each string of words from a display on a computer. Participants were told to read each list aloud until they were ready to record, then they pressed the spacebar to begin recording. Instructions appeared prompting the participant to say the list, and a beep was played which notified the speaker that the recording device was on. When the participant was finished speaking, she pressed the spacebar. In the Memory condition, participants were given a maximum of 10 seconds to memorize the list. After the participant pressed the spacebar or 10 seconds elapsed (whichever was first), a prompt was displayed indicating that participants should recite the list of words from memory. The instructions then appeared, and the recording of the list was obtained as for the Reading condition.

In the Dual task condition, participants were given computer-based prompts as in the Memory condition, but were told that the experimenter would be controlling some of the spacebar presses. Participants had to memorize a list and then were presented with a complete sentence during the recording phase on the computer screen. Each sentence was 8-12 words in length and contained between 3 and 6 content words. All sentences had a simple grammatical structure with no relative clauses. The sentence was read silently by the participant while she recorded and recited the memorized list. The recording process was stopped by the experimenter pressing the spacebar. Then the participant was asked to answer a simple comprehension question about the silently read sentence after finishing the list. All recordings were subsequently placed into a database for later analysis.

Recordings were made using a Countryman MHHP6HH05B head-mounted microphone connected to a MOTU preamp wired to a PC. E-prime 2.0.8.73 software created by Psychological Software Tools (Sharpsburg, PA), running under Windows XP was used for data acquisition. The sound was captured at a sampling rate of 22050 Hz and encoded with 32-bit quantization.

Instructions emphasized that lists should be read in a neutral manner by encouraging participants to read each list of words as if they were simply reciting a grocery list. In addition, each participant was told that producing all words in the list accurately and in the correct order was very important. They were instructed to produce the list again from its beginning if they made a mistake.

Each of the nine talkers had to produce ten word lists in each of three task conditions, leading to a total of 270 productions (9 talkers × 10 lists × 3 task conditions); one item by a single talker was not recorded due to participant error, for a total of 269 productions. Before the Reading and Memory trials, a block of practice trials containing six lists with five words each was run for the participant to get achieve some level of familiarity with the task before running the trial recordings. Due to its greater difficulty, the Dual condition had a longer practice block consisting of ten recordings with five words each. After all data had been collected, Audacity 1.3 was used to trim the beep signaling the start of the recording from the beginning of each digitized audio file; any silence recorded after the production of the list was also removed. In addition, any nonspeech sounds (e.g., laughter) and non-task related speech (e.g. saying *"or something like that"*) were cut off if they occurred after the list was fully recited.

2.1 Acoustic measurements

To quantify the level of acoustic isochrony in speech, intervals between vocalic onsets were measured. Vowel onsets have been identified as being quite close to the perceptual moment of occurrence of a syllable (the p-center) (e.g., Cummins and Port 1998; Marcus 1981). Compared to measures of timing such as the normalized pairwise variability index (nPVI) (Grabe and Low 2002) and coefficient of variation of consonantal intervals and the percentage of vocalic intervals (Ramus, Nespor and Mehler 1999), the measure of duration between vocalic onsets would appear to provide the most direct correspondence between perception of a sequence of syllables in a fixed order and its acoustic properties. To do this, a script written for PRAAT (Boersma and Weenink 2002) was first used to automatically detect and label the midpoints of large changes in intensity, which were taken as approximations to vowel onsets. These estimates were then hand-edited by shifting, adding, and removing labeled vowel onset estimates using waveform and spectrogram displays to give a final accurate estimate of the vowel onset of each syllable. In cases where the vowels abutted other sonorant consonants (e.g., approximants), formant frequency and intensity guided segmentation, especially those of the second and third formants.

2.2 Prosodic annotation

Productions were also analyzed by two undergraduate coders proficient in the RaP prosody labeling system (Breen et al., in press; Dilley et al. 2006; Dilley and Brown 2005). This system allows for coding of perceptual isochrony, pitch accents, and other prosodic events and has been shown to yield high inter-rater reliability (Breen et al., in press). The coders had gone through extensive training in the RaP system for a different research project and were considered proficient in its application (Breen et al., in press). Moreover, they were kept blind to the experimental conditions of the experiment and were naïve to the purpose of the study. Recordings were marked for perceptually isochronous sequences of syllables, as well as any disfluencies that were present within the list. Recordings for which most (i.e., 80 % or more) of the intervals between vowel onsets belonged to a stretch of perceptual isochrony, as labeled by both coders, were designated "perceptually isochronous". In contrast, recordings with fewer intervals labeled as perceptually isochronous, or without any at all, were designated "not isochronous". This allowed for the calculation of the proportion of recordings which were judged perceptually isochronous within each production condition.

2.3 Perceptual evaluation of rhythmicity

Twenty individuals evaluated the perceived rhythmicity of the produced speech in return for course credit. Rhythmicity was defined as how easily participants heard "a rhythm", that is, "a steady, regular pattern of beats in the speech recording." Participants were told that a rhythm could be fast or slow, as long as it is steady; they were encouraged to tap, clap, or otherwise move to the speech in an attempt to find the beat. Participants judged the rhythmicity of the recordings produced by each talker for each item on a Likert scale from 1 to 6; 1 was defined as "very non-rhythmic", 2 as "moderately non-rhythmic", 3 as "somewhat non-rhythmic", 4 as "somewhat rhythmic", 5 as "moderately rhythmic", and 6 as "very rhythmic".

A practice block containing six trials was presented at the beginning of the rhythmicity judgment experiment. These six trials corresponded to six recordings that had been made during the practice blocks of the production experiment. Three of the selected practice trials were presented as examples of highly nonrhythmic speech, while the other three practice trials were presented as examples of highly rhythmic speech; rhythmic and nonrhythmic practice trials lacked disfluencies or contained at least one disfluency, respectively. After each practice trial, the participant was told the intended range of scores for that trial's recording; in particular, participants were told whether the range of scores for the practice trial should have been "rhythmic" (i.e., scores in the range 4-6) or "non-rhythmic" (i.e., scores in the range 1-3). For the experiment itself, a single list of recordings was created by arranging the 269 productions into 9 blocks of trials for rhythmicity judgments; on none of the experimental trials was a recording that had been presented during a practice trial repeated. Within a block, each participant heard each item (i.e., list of words) just once, produced by a single randomly selected talker. Lists were presented in a fixed, random order within a block. Eight blocks contained 30 trials, while the final block contained 29 trials. A second list of recordings was created by reversing the order of the first list. Participants were randomly assigned to one of the two lists. The average rhythmicity judgment across the 20 listeners for each word list produced by a given talker was determined.

2.4 Perceptual evaluation of fluency

A different set of twenty individuals evaluated the perceived fluency of speech in a separate study in return for course credit; none of these had participated in the rhythmicity judgment study discussed above. 'Fluent' speech was defined for participants as speech which sounds connected and fluid, with words having just the right duration and pauses in the right places. "Disfluent" speech was defined as speech which sounds disconnected and not fluid, with words having the wrong length (too short or too long) and pauses in the wrong places. Participants were told that fluent speech could be either fast or slow, that there would be no recordings of disordered speech, and that both fluent and disfluent recordings would be presented from normal speakers only (i.e. not stutterers). Participants judged the fluency of the recordings produced by each talker on a scale from 1 to 6; 1 was defined as "very disfluent", 2 as "moderately disfluent", 3 as "somewhat disfluent", 4 as "somewhat fluent", 5 as "moderately fluent", and 6 as "very fluent".

The same set of practice trials were used for the practice block as for the experiment involving perceptual evaluation of rhythmicity (Section 2.3). After each practice trial, the participant was told the intended range of scores for that trial's recording; in particular, participants were told whether the range of scores for the practice trial should have been "fluent" (i.e., scores in the range 4-6) or "disfluent" (i.e., scores in the range 1-3). A single list of recordings was created by arranging the 269 productions into 9 blocks of trials for fluency judgments. Within each block, each participant heard each list of words just once, produced by a single randomly selected talker; lists were presented in a fixed, random order within a block. Eight blocks contained 30 trials, while the final block contained 29 trials. A second list of recordings was created by reversing the order of the first list. Participants were randomly assigned to one of the two lists. The average fluency judgment across the 20 listeners for each word list produced by a given talker was determined.

3 Results

3.1 Acoustic analysis

The mean duration of the inter-onset-interval (IOI) between successive vowels was 551 ms (s = 147 ms). The speech rate as given by average IOI duration was analyzed using a repeated measures ANOVA both by-subjects (F_1) and by-items (F_2); speech rate did not change significantly across task conditions [F_1 (2,16) = .410, *MSE* = .004, $p > .050$, $\eta_p^2 = .049$; F_2 (2,58) = .706, *MSE* = .008, $p > .050$, $\eta_p^2 = .024$]. To quantify the regularity of intervals within a given recording, the coefficient of variation (CoV) was calculated by determining the standard deviation of interval durations for a recording and dividing by that recording's mean duration. There was a significant difference in CoV across tasks [F_1 (2,16) = 5.583, *MSE* = .003, $p < .050$, $\eta_p^2 = .411$; F_2 (2,58) = 6.872, *MSE* = .009, $p < .010$, $\eta_p^2 = .192$]; see Figure 1. Paired samples *t*-tests with Bonferroni correction on subject data revealed a marginally significant difference between the Reading and Dual task conditions ($p < .050$), but no other difference. To determine whether these differences in acoustic similarity were attributable to lengthened and/or irregular intervals associated with disfluencies, intervals characterized as a disfluency by either prosodic analyst (see Section 2.2) were removed from the analysis and CoV was recalculated based on fluent intervals only. When this was done, there was no significant effect of task on CoV [F_1(2,16) = .345, *MSE* = .002, $p > .050$, $\eta_p^2 = .041$; F_2(2,58) = .733, *MSE* = .003, $p > .050$, $\eta_p^2 = .025$]; see Figure 2.

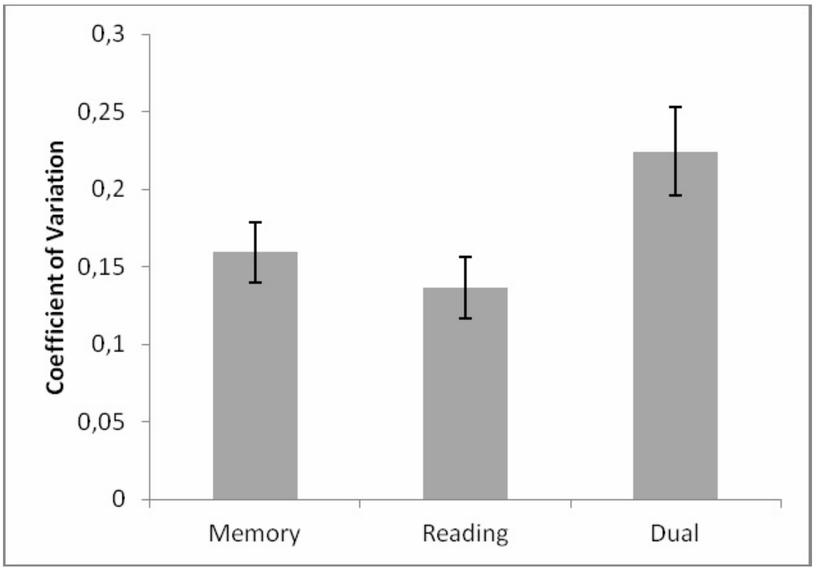

Figure 1: Coefficient of variation of inter-onset-intervals (IOIs) between vowel onsets across tasks in the production study.

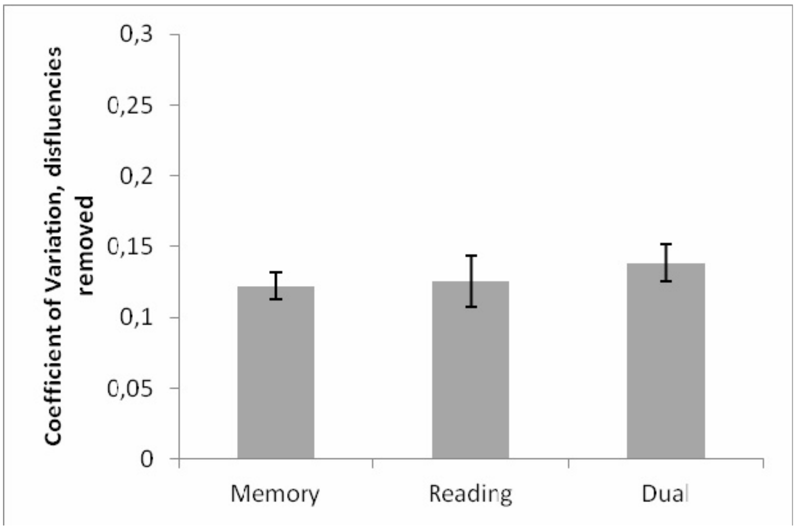

Figure 2: Coefficient of variation of inter-onset-intervals (IOIs) between vowel onsets across tasks in the production study, with disfluent intervals removed from calculations.

3.2 Prosodic labeling analysis

For each production condition we calculated the proportion of recordings (i.e., trials) which were judged perceptually isochronous by both prosody labellers, given the nine subjects in the production experiment and the 30 items produced by each subject in that experiment. A high proportion of trials in both the Memory condition and the Reading condition were found to sound perceptually isochronous by both analysts (M_{Memory} = 0.85, $M_{Reading}$ = 0.80), while the proportion of trials in the Dual task condition that sounded perceptually isochronous was substantially lower (M_{Dual} = 0.59); see Figure 3. A one-way ANOVA was conducted using the factor Task condition (Memory, Reading, Dual) on the mean proportion of trials judged perceptually isochronous by both labellers, both by-subjects (F_1, where here the subjects are the nine participants in the production study) and by-items (F_2). The effect of Task condition on the proportion of trials judged to sound perceptually isochronous was significant [F_1 (2,16) = 12.881, MSE = 0.26, p < .001, η_p^2 = .617; F_2 (2,58) = 13.890, MSE = 0.083, p < .001, η_p^2 = .324]. Post-hoc t-tests with Bonferroni correction confirmed the presence of a significant difference in the proportion of trials judged to sound perceptually isochronous between the Memory and Dual conditions and between the Reading and Dual conditions (p < .010), but no difference between the Memory and Reading conditions.

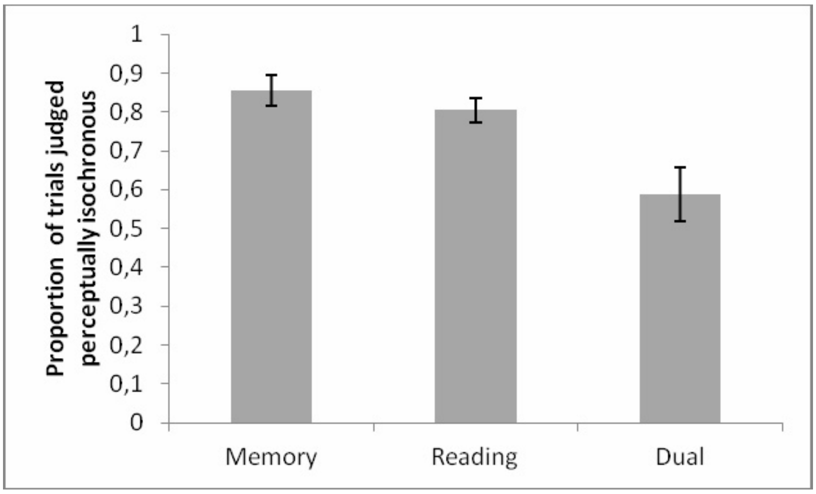

Figure 3: Proportion of trials in the production study which were judged to sound perceptually isochronous by prosodic analysts.

In addition, we determined the proportion of trials judged to be fluent by the prosodic analysts; a trial was defined as fluent if neither labeler judged it to contain a disfluency. Task type had an effect on rates of fluent recording production; the mean proportion of fluent recordings was high for the Reading and Memory conditions (M_{Memory} = 0.82, $M_{Reading}$ = 0.73) but lower for the Dual task condition (M_{Dual} = 0.41). This difference was significant [F_1 (2,16) = 8.016, MSE = 0.052, $p < .010$, $\eta_p^2 = .500$; F_2 (2,58) = 18.917, MSE = 0.074, $p < .001$, $\eta_p^2 = .395$]. Post-hoc *t*-tests with Bonferroni correction on subject data showed a significant difference between the Memory and Dual task conditions ($p < .010$) and no other significant differences.

3.3 Rhythmicity judgments

Mean ratings of rhythmicity varied by task (M_{Memory} = 4.39, $M_{Reading}$ = 4.52, M_{Dual} = 3.76); see Figure 4. These differences in mean rhythmicity ratings were significant [F_1 (2,16) = 11.390, MSE = .119, $p < .010$, $\eta_p^2 = .587$; F_2 (2,58) = 13.669, MSE = .373, $p < .001$, $\eta_p^2 = .320$]. Paired-samples *t*-tests with Bonferroni correction showed that the Memory and Dual task conditions were statistically different, as were the Reading and Dual task conditions, but the Memory and Reading task conditions were not significantly different.

Figure 5 shows that substantial proportions of recordings in the Memory task and in the Reading task were judged highly rhythmic (i.e., receiving a mean rating of 4.5 or greater) (M_{Memory} = 0.79, $M_{Reading}$ = 0.80), while the proportion of rhythmic recordings in the Dual task condition was much lower (M_{Dual} = 0.53). These differences in proportions were significant [F_1 (2,16) = 5.931, MSE = 0.034, $p < .050$, $\eta_p^2 = .426$]. Paired-samples *t*-tests with Bonferroni correction on subject data showed that the Memory and Dual task conditions were significantly different, while the Reading and Dual task conditions were marginally significantly different.

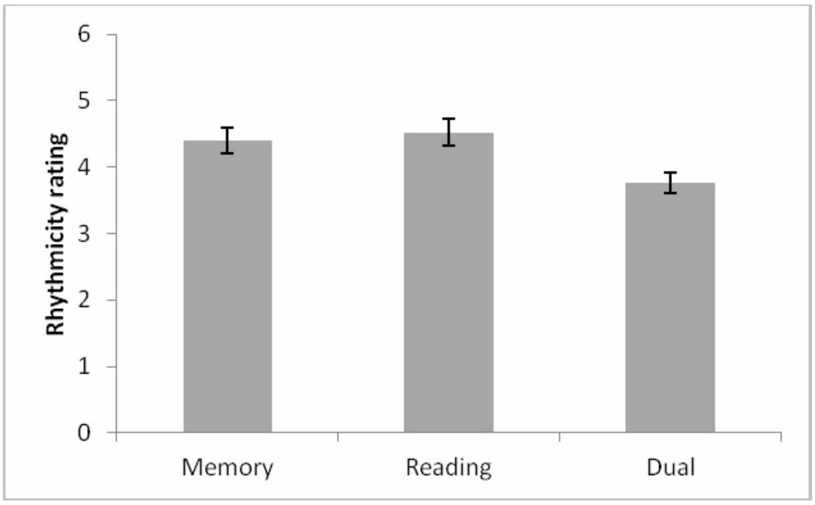

Figure 4: Rhythmicity ratings of recordings by task.

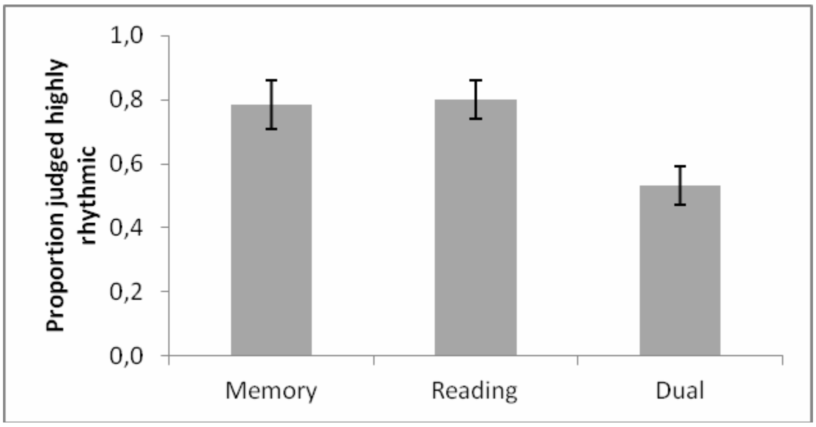

Figure 5: Average proportion of recordings judged highly rhythmic by naïve listeners.

3.4 Fluency judgments

The mean ratings of fluency reported by participants were higher in the Memory and Reading tasks than in the Dual task (M_{Memory} = 4.78, $M_{Reading}$ = 4.91, M_{Dual} = 4.18); see Figure 6. These differences were significant [$F_1(2,16) = 11.390$, $MSE = .119$, $p < .010$, $\eta_p^2 = .587$; F_2 $(2,58) = 13.461$, $MSE = .339$, $p < .001$, $\eta_p^2 = .317$]. Paired-samples *t*-tests with Bonferroni correction on subject data confirmed significant differences between the Memory and Dual task conditions and the Reading and Dual task conditions, but not between the Memory and Reading conditions. Moreover, the proportion of recordings judged fluent (receiving a rating of 4.0 or greater) was affected by task [$F_1(2,16) = 10.717$, $MSE = .010$, $p < .010$, $\eta_p^2 = .573$]. Paired-samples *t*-tests with Bonferroni correction using subject data showed that the Memory and Dual task conditions differed ($p < .010$), as did the Reading and Dual task conditions ($p < .010$), but the Memory and Reading task conditions were not different.

A correlational analysis was conducted for the relation between fluency judgments and rhythmicity judgments for each recording. The result is shown in Figure 7. Pearson's correlation coefficient was $R = 0.922$; this correlation was significant at $p < .0001$.

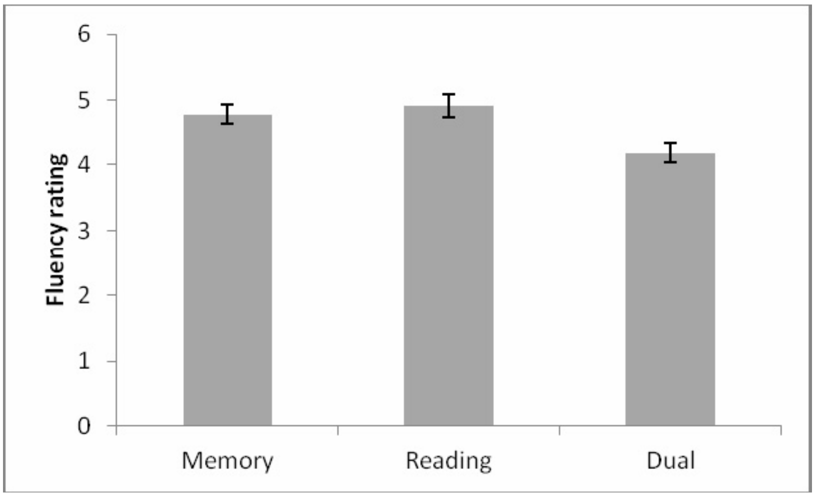

Figure 6: Average fluency ratings by naïve listeners of productions.

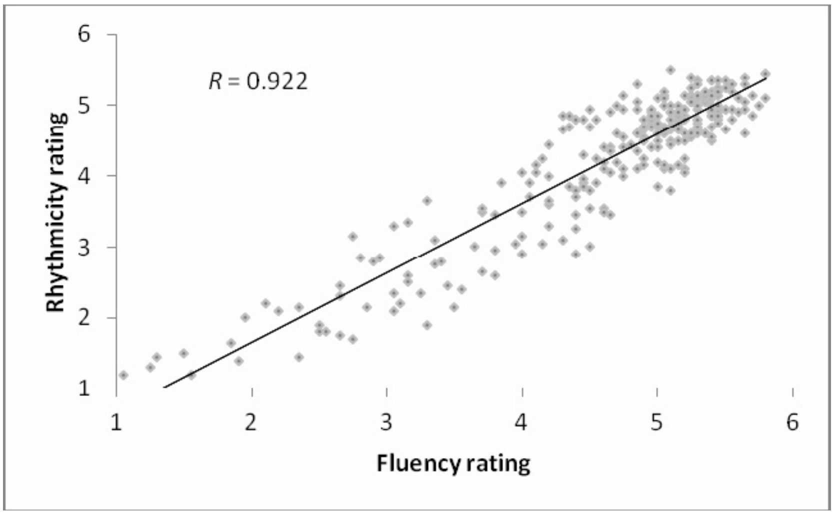

Figure 7: Fluency ratings plotted as a function of rhythmicity ratings for each recording; ratings of each recording were carried out by different groups of listeners. Pearson's correlation coefficient, *R*, is shown, along with the line of linear regression.

4 Discussion

This research builds on recent findings suggesting that perceptual isochrony plays a role in word segmentation and lexical access processes of spoken word recognition (Brown et al. 2011; Dilley et al. 2010; Dilley and McAuley 2008). The present study aimed to quantify the extent to which speech is perceived as rhythmically regular (i.e., perceptually isochronous) and the extent to which such perceptual cues could therefore be useful to perceivers in normal speech communication conditions. Though we do not claim this pattern is exhibited in speech across the board, it does show that perceptual isochrony is frequent (indeed, the norm) for at least one commonly-occurring type of lexical format, that is, monosyllabic content words in lists. This research also explored the relationship between perception of isochrony, acoustic timing, and perception of fluency, which presently is not well understood (Amir and Yairi 2002; Eklund 2004; Guitar 2005; Kawai et al. 2007; Van Riper 1982). In the present production study, participants spoke monosyllabic word lists under three different experimental conditions designed to elicit a range of fluency and acoustic timing characteristics. The attributes of the speech were assessed in a variety of ways, including perceptual judgments of rhythmicity and fluency, prosodic annotation, and acoustic measurements.

Measurement of other aspects of speech, such as pitch characteristics, or using alternative measures of timing, like nPVI (Grabe and Low 2002) or measures of consonantal variability and vocalic duration (Ramus et al. 1999), may prove to be fruitful ways of examining acoustic correlates of perceptual judgments of rhythmicity and fluency in the future.

These results suggest that in normal speaking situations, i.e., when talkers are not under stringent performance demands on attention and/or memory, monosyllabic content word lists are perceptually isochronous a high proportion of the time (i.e., approximately 80 % of the time or higher) based on judgments of both naïve listeners and trained prosodic analysts. These findings indicate that at least in lists of monosyllabic content words under low task demands, perceptual isochrony is a cue which is often present in speech and would be expected to affect processes of spoken word recognition. Moreover, the present research confirms previous results showing that speech which is perceptually isochronous is not acoustically isochronous to any significant degree (Bolinger 1965; Classe 1939; Cummins 2003, 2005; Dauer 1983; Kohler 1986; Roach 1982; Shen and Peterson 1962; Uldall 1971), as shown by the relatively high value of coefficient of variation in interval timing.

By comparing speaking conditions differing in performance demands, the present research also revealed a previously unknown relationship between perception of isochrony and perception of fluency. A core behavior of disfluent speech, including clinical stuttering, is that speech sounds are perceived as inappropriately lengthened (Eklund 2004; Guitar 2005; Van Riper 1982), yet exactly how acoustic timing of speech contributes to perception of disfluency is unclear. The present data helps to shed light on these issues by demonstrating a relationship between perceptual isochrony and fluency through several converging measures. First, the production condition which entailed the least perceptual isochrony (i.e., the Dual task condition, which involved simultaneous silent reading and reciting of a memorized word list) also showed the least fluency; this pattern was demonstrated both by prosodic annotation by trained analysts, as well as by judgments of naïve listeners. Moreover, the relationship between perception of isochrony and fluency was revealed through the high correlation coefficient ($R = 0.922$) between judgments of rhythmicity by one group of naïve listeners and judgments of fluency by a different group of listeners. This provides insight into listeners' intuitions concerning the nature of the two concepts. Based on our experimental data, the degree of perceived rhythmicity was closely linked with the degree of perceived fluency. The strikingly high correlation is remarkable given that perceptual judgments of rhythmicity and fluency were made by separate groups of participants, and given that these participants were not given feedback about gradations of the judgments they should supply. However, the fact that participants were presented with the same set of six practice trials as examples of fluent or rhythmic speech, on the one hand, or disfluent or non-rhythmic speech, on the other hand, likely increased the extent of correlation observed across the two sets of perceptual judgments. Nevertheless, it is important to note that none of the practice recordings were repeated during experimental trials, so that all responses by participants during the experimental trials of both experiments reflected generalization about the definitions of rhythmicity and fluency that had been provided during practice. Thus, it is all the more striking that a high correlation was observed across the two sets of perceptual judgments. Finally, knowledge about the relationship between fluency and perceived rhythmic regularity was enhanced by the fact that the dual task condition showed significantly less acoustically regular timing (as measured by coefficient of variation) before, but not after, disfluent intervals were removed.

In summary, the present research demonstrates that certain types of lexical sequences, namely monosyllabic lists of content words, are produced with perceptual isochrony a high proportion of the time under production conditions typical of normal communication situations (i.e., producing lists of items from memory and reading). Moreover, a novel finding regarding the relationship between perception of isochrony and perception of fluency was revealed, findings which provide a useful starting point for further investigations of the acoustic underpinnings of speech produced with fluency. The present results, in conjunction with recent findings demonstrating a role for perceptual isochrony in spoken word recognition (Brown et al. 2011; Dilley et al. 2010; Dilley and McAuley 2008), thus indicate that perceptual isochrony is an attribute of speech which is common in certain types of lexical sequences and which can be fruitfully exploited in a variety of communicative listening situations. Moreover, maintaining perceptually isochronous speech timing appears to be important in order for certain types of lexical sequences to be perceived as fluent.

Overall, the present paper contributes to the themes of this edited volume (*context, function, communication*) by illustrating that at least one type of speech material, namely monosyllabic word lists, frequently contains prosodic cues to perceptual isochrony. Such distal (i.e., *context*) prosodic cues associated with perceived regular rhythm have been shown in our previous work to contribute to spoken word recognition (i.e., *communication*) via serving a role (i.e., *function*) in word segmentation and lexical access. The present studies help to illustrate that distal prosodic cues to regular rhythm are likely to be frequently available, and therefore potentially useful for spoken word recognition, in at least some speech materials and speaking situations, notably read and recited monosyllabic wordlists.

5 Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 0847653.

6 References

Amir, O. and E. Yairi (2002): The effect of temporal manipulation on the perception of disfluencies as normal or stuttering. *Journal of Communication Disorders* ***35***, 63-82.

Barbosa, P. (2007): From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication* ***49***, 725-742.

Boersma, P. and D. Weenink (2002): Praat, a system for doing phonetics by computer (Version 4.0.26). *Software and manual available online at http://www.praat.org.*

Bolinger, D. (1965): Pitch accent and sentence rhythm. In: I. Abe, T. Kanekiyo (eds): *Forms of English: Accent, Morpheme, Order* (pp. 139-180). Cambridge, MA: Harvard University Press.

Breen, M., L. C. Dilley, J. Kraemer and E. Gibson. (in press): Inter-transcriber agreement for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch). *Corpus Linguistics and Linguistic Theory.*

Brown, M., A. P. Salverda, L. C. Dilley and M. K. Tanenhaus (2011): Distal prosody influences lexical interpretation in on-line sentence processing. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society. Austin, USA.*

Classe, A. (1939): *The Rhythm of English Prose.* Oxford: Blackwell.

Cummins, F. (2003): Rhythmic grouping in word lists: competing roles of syllables, words and stress feet. *Proceedings of the 15th International Congress of Phonetic Sciences, Barcelona, Spain,* 325-328.

Cummins, F. (2005): Interval timing in spoken lists of words. *Music Perception* ***22,*** 497-508.

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

Cutler, A. and D.M. Carter (1987): The predominance of strong initial syllables in the English vocabulary. *Computer Speech and Language* ***2***, 133-142.

Cutler, A. and D.G. Norris (1988): The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance* ***14***, 113-121.

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

Dilley, L.C., M. Breen, M. Bolivar, J. Kraemer and E. Gibson (2006): A comparison of inter-transcriber reliability for two systems of prosodic annotation: RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices). *Proceedings of Interspeech 2006, Pittsburgh, USA.*

Dilley, L.C. and M. Brown (2005): The RaP (Rhythm and Pitch) Labeling System, Version 1.0: *Available at http://tedlab.mit.edu/rap.html.*

Dilley, L.C., S.L. Mattys and L. Vinke (2010): Potent prosody: Comparing the effects of distal prosody, proximal prosody, and semantic context on word segmentation. *Journal of Memory and Language **63***, 274-294.

Dilley, L.C. and J.D. McAuley (2008): Distal prosodic context affects word segmentation and lexical processing. *Journal of Memory and Language **59***, 294-311.

Eklund, R. (2004): *Disfluency in Swedish human-human and human-machine travel booking dialogues*. Linköping Universitet.

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

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

Grant, D. A. (1948): The Latin square principle in the design and analysis of psychological experiments. *Psychological Bulletin **45***, 427-442.

Guitar, B. (2005): *Stuttering: An Integrated Approach to its Nature and Treatment* (3rd ed.). Baltimore: Lippincott Williams & Wilkins.

Kager, R. (1995): The metrical theory of word stress. In: J.A. Goldsmith (ed.): *The Handbook of Phonological Theory* (pp. 367-402). Cambridge, Mass.: Blackwell.

Kawai, N., E.C. Healey, and T.D. Carrell (2007): Listeners' identification and discrimination of digitally manipulated sounds as prolongations. *Journal of Acoustical Society of America **122***, 1102-1110.

Kohler, K.J. (1986): Invariance and variability in speech timing: from utterance to segment in German. In: J. Perkell, D.H. Klatt (eds): *Invariance and Variability in Speech Processes* (pp. 268-289). Hillsdale, New Jersey: Lawrence Erlbaum.

Large, E.W. and M.R. Jones (1999): The dynamics of attending: How people track time-varying events. *Psychological Review **106***, 119-159.

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

Livant, W.P. (1963): Antagonistic functions of verbal pauses: filled and unfilled pauses in the solution of additions. *Language and Speech **6***, 1-4.

Marcus, S. M. (1981): Acoustic determinants of perceptual center (P-center) location. *Perception and Psychophysics **30***, 247-256.

McAuley, J.D. (1995): *Perception of time as phase: toward an adaptive-oscillator model of rhythmic pattern processing.* Unpublished Ph.D. Dissertation, Indiana University.

McNemar, Q. (1951): On the use of Latin squares in psychology. *Psychological Bulletin **48***, 398-401.

Norris, D., J.M. McQueen, A. Cutler, S. Butterfield and R. Kearns (2001): Language-universal constraints on speech perception. *Language and Cognitive Processes **16***, 637-660.

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

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

Pitt, M. and A.G. Samuel (1990): The use of rhythm in attending to speech. Journal of Experimental Psychology: *Human Perception & Performance* ***16***, 564-573.

Port, R.F. (2003): Meter and speech. *Journal of Phonetics* ***31***, 599-611.

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

Roach, P. (1982): On the distinction between 'stress-timed' and 'syllable-timed' languages. *Linguistic Controversies*, 73-79.

Saltzman, E., H. Nam, J. Krivokapic and L. Goldstein (2008): A task-dynamic toolkit for modeling the effects of prosodic structure on articulation. *Proceedings of the 4th International Conference of Speech Prosody, Campinas, Brazil*, 175-184.

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

Shen, Y. and G.G. Peterson (1962): Isochronism in English. *Studies in Linguistics, Occasional Papers: University of Buffalo*, 1-36.

Studdert-Kennedy, M. (1980): Speech perception. *Language and Speech* ***23***, 45-66.

Tilsen, S. (2006): Rhythmic coordination in repetition disfluency: a harmonic timing effect. *UC-Berkeley Phonology Lab Annual Report*, 73-114.

Uldall, E. T. (1971): Isochronous stresses in R.P. In: L.L. Hammerich, R. Jakobson, E. Zwirner (eds): *Form and Substance: Phonetic and Linguistic Papers Presented to Eli Fischer-Jorgensen* (pp. 205-210). Copenhagen: Akademisk Forlag.

Van Riper, C. (1982): *The Nature of Stuttering* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

White, L. (2002): *English speech timing: a domain and locus approach.* The University of Edinburgh.

Appendix

Group A

dog, fish, fox, frog, hawk
wool, fur, felt, lace, quilt
skirt, vest, shirt, belt, hat
car, bus, boat, jet, plane
lip, leg, hand, heart, chin
stalk, stem, leaf, bud, root
red, green, gold, black, white
wolf, worm, shark, bat, bird
ball, top, dice, cards, doll
pan, pot, tank, sack, jar

Group B

thumb, skull, rib, neck, lung
ship, truck, van, bike, train
drum, horn, harp, gong, flute
can, cup, box, bowl, dish
bun, bread, corn, plum, ham
bug, bull, cat, cow, crab
brass, brick, stone, glass, wood
square, line, point, curve, sphere
gulf, beach, hill, cliff, pond
wisk, fork, spoon, knife, plate

Group C

scarf, shawl, cap, dress, coat
twill, silk, cloth, thread, yarn
lamb, moth, pig, ram, rat
house, barn, shed, yard, farm
fin, claw, tusk, tail, fang
drill, nail, wrench, tool, screw
bank, mall, park, school, church
sun, moon, night, star, cloud
cheese, fruit, salt, milk, pork
bark, thorn, bush, fern, grass

List of Index Terms

Acoustic isochrony
Anisochrony
Disfluencies
Entrainment
Fluency
Intonation (of lists)
Memory demands
Monosyllables
Perceptual isochrony
Prosody
Rhythmicity
Speech rhythm
Spoken word recognition
Stuttering
Vowel onsets
Word segmentation

Short Portraits

Dilley, Laura

Assistant Professor in the Department of Communicative Sciences and Disorders, Michigan State University. Research interests: role of prosody in word segmentation and lexical access; phonological representations in children and adults.

Wallace, Jessica

B.S. in Genomics and Molecular Genetics and B.A. in Linguistics from Michigan State University. Graduate student in the Cognition and Cognitive Neuroscience program, Department of Psychology, Michigan State University. Research Interests: first language acquisition, speech prosody, speech perception, music cognition, genetic basis and evolutionary roots of language.

Heffner, Christopher C.

Undergraduate in Linguistics (B.A.) and Psychology (B.Sc.) programs at Michigan State University; Research interests: neurolinguistics of speech processing, especially word segmentation, and phonology.