Running head: Timing and meter in word perception

Effects of timing regularity and metrical expectancy on spoken-word perception

Hugo Quené

Utrecht institute of Linguistics OTS, Utrecht University

Robert F. Port

Department of Linguistics, Indiana University

7 March 2005

address for correspondence:

Dr. Hugo Quené

Trans 10, 3512 JK Utrecht, The Netherlands

hugo.quene@let.uu.nl

Abstract

Certain types of speech, e.g. lists of words or numbers, are usually spoken with highly regular inter-stress timing. The main hypothesis of this study (derived from the Dynamic Attending Theory; M.R. Jones (1976), Psych. Rev. 83, 323–355) is that listeners attend in particular to speech events at these regular time points. Better timing regularity should improve spoken-word perception. Previous studies have suggested only a weak effect of speech rhythm on spoken-word perception, but the timing of inter-stress intervals was not controlled in these studies. A phoneme monitoring experiment is reported, in which listeners heard lists of disyllabic words in which the timing of the stressed vowels was either regular (with equidistant inter-stress intervals) or irregular. In addition, metrical expectancy was controlled by varying the stress pattern of the target word, as either the same or the opposite of the stress pattern in its preceding words. Resulting RTs show a main effect of timing regularity, but not of metrical expectancy. These results suggest that listeners employ attentional rhythms in spoken-word perception, and that regular speech timing improves speech communication.

Keywords: timing; rhythm; meter; dynamic attending; stress; speech; spoken-word recognition

Effects of timing regularity and metrical expectancy on spoken-word perception

Introduction

Speech movements, speech sounds and speech percepts unfold over time. Breathing movements and articulatory movements are intrinsically periodic. Previous research has clearly demonstrated that speech production and the resultant speech signal are indeed rhythmically constrained (Port et al., 2002; Port, 2003). For example, in speech cycling (repeating the same phrase), speakers tend to place the stressed syllables non-randomly within the phrase, preferably at 1/2 or 1/3 or 2/3 of the phrase period. The resultant speech rhythm can be modelled as a dynamical system consisting of two coupled oscillators (Port et al., 2002). The slower oscillator produces the whole-phrase repetition, and the faster oscillator produces the time points that attract the onsets of stressed vowels (Allen, 1972).

As a next step, it has been hypothesized that not only the speaker's behavior is periodic, but that the listener too is sensitive to the resultant speech rhythms. "If auditory stimulus sequences normally occurring in the natural environment are characteristically organized in a certain way, e.g., they are rhythmic, then one might reasonably expect that perceptual mechanisms as they have evolved will be biased to listen for sounds organized in this certain way" (Sturges & Martin, 1974, p. 377). In other words, it seems likely that listeners pay more attention to those points in time where the rhythmic beats are strongest, i.e., when stressed syllables are expected to occur. Speech perception should be better (more accurate and/or faster) at these beats, relative to other off-beat points.

Focusing attention on the expected locations of stressed syllables would indeed be advantageous for listeners. Stressed syllables are more salient than unstressed ones: they have higher intensity, longer duration, and often a conspicuous pitch movement which aids in defining the spectral envelope (for a review of phonetic correlates of stress, see Beckman, 1986; Sluijter, 1995; Sluijter & Van Heuven, 1996). This makes stressed syllables particularly informative about the linguistic content of the speech signal. Indeed, stressed syllables contribute more than unstressed syllables to spoken-word perception (e.g. Van Leyden & Van Heuven, 1996; Quené & Koster, 1998; Cutler & Van Donselaar, 2001).

The main hypothesis in this study, then, is that listeners focus their attention to time points in the speech signal when salient, stressed syllables are expected to occur, with this expectancy derived from the timing (rhythmic, isochronous) properties of the speech signal heard so far.

This timing-expectancy hypothesis was first formulated many years ago (see Lehiste, 1973, 1977, 1980, and references given there). Various pieces of circumstantial positive evidence have been obtained since then. One such piece, for example, comes from a phoneme monitoring study by Meltzer, Martin, Mills, Imhoff, and Zohar (1976, Exp.2). Participants listened to sentences in which the target phoneme was on time, too early, or too late. The latter two conditions were obtained by splicing out a 100-ms fragment from the original tape, at 50 ms before the target phoneme; for the late condition a 200-ms interval of white noise was then spliced in at that point. Average reaction times (RTs) were faster for the on-time condition (700 ms) than for the early and late conditions (781 and 765 ms, respectively). Although this finding agrees with the above hypothesis, there are two reasons for caution. First, the timing of the carrier sentence (its tempo, rhythm, phrasing) was not controlled, so it is not clear what information listeners used as a reference for timing or for setting up attentional rhythms. Second, the acoustic distortions themselves might have had a negative effect on RTs, not due to their disrupting effect on the speech timing but due to their low-level acoustic artefacts.

More evidence comes from a study by Shields, McHugh, and Martin (1974). RTs to target phonemes were compared between accented (rhythmically predictable) vs. unaccented (unpredictable) syllables of nonsense words. Target words were embedded in a carrier sentence such as (1), where the target word occurs in the middle of the last phrase (capitals indicate stress).

(1) You will have to curtail any sightseeing plans,
as the plane to BENkik/benKIK leaves at noon.

Presumably, the carrier sentence would make listeners expect an initially-stressed target word (cf. London, Boston), with accent on its stressed initial syllable. The accented and unaccented conditions yield significantly different RTs, for target words in early (594–657 ms) and middle position (537–640 ms) of the last phrase. These differences cannot be due to acoustic differences between accented and unaccented targets, since the RT differences disappear if the words are excised and presented in a nonsense word list (671–687 ms). Caution is again required, however, because the carrier sentence was not kept constant across accent conditions, and because the timing of the carrier sentence was not controlled.

In a modified replication of this last experiment, however, Pitt and Samuel (1990) found no support for the above hypothesis. In a phoneme monitoring study, a stress-neutral target word (e.g. *per.mit*) was embedded in a carrier sentence. Average RTs were only 7 ms faster in predicted-stressed conditions (trochee target *PER.mit* as noun, with initial-syllable target; iambic target *per.MIT* as verb, with final-syllable target) than in the opposite, predicted unstressed conditions. Again, timing in the carrier sentence was not controlled. In a follow-up experiment, Pitt and Samuel (1990) attempted to exert more control over the carrier sentence. This was done by embedding the target word as the fifth element in a word sequence, where it was preceded only by trochees (2a) or only by iambes (2b). The same token of the target word was used in both sequence conditions.

(2) a. OLive–VILLage–TISSue–KAyak–permit...
    b. obTUSE–creATE–abSCOND–eQUATE–permit...

Average RTs were 24 ms faster in the predicted-stressed conditions than in the predicted-unstressed conditions (606–570 ms for initial-syllable targets; 547–535 ms for final-syllable targets).

From these results, Pitt and Samuel (1990) conclude that rhythmic expectancy has only a small effect on English spoken-word perception. Caution is again required, however, because the perceptual effect of rhythm may have been underestimated in this study. Pitt and Samuel (1990) controlled the contents of their speech materials, and systematically varied *what* was heard in their stimuli. The relevant factor is thus the *metrical* sequencing of strong and weak syllables, and listeners' metrical expectancy. But like in the other studies reviewed above, there was no control over *when* the strong syllables were heard, i.e. over the actual timing of the speech context leading up to the target words. In their word sequence experiment, the "interword interval within a word sequence ranged from 500 to 800 ms" (Pitt & Samuel, 1990, p. 568). The hypothesis being tested was thus not about rhythmical expectancy, but about metrical expectancy. The considerable variation in inter-word time intervals (probably yielding even larger variation in inter-stress intervals) has made the actual timing of stressed vowels effectively random, yielding no regular timing at all. Listeners were unable to build any rhythmic expectancy about the upcoming speech signal, hence the absence of a strong effect of speech rhythm.

Buxton (1983) reports a phoneme monitoring study in which sentence timing was disrupted, by cross-splicing sentence fragments with matched and unmatched inter-stress intervals:

(3) a. The small girl was playing with the // red toy in the garden...
b. The small girl was playing with the // reddish toy in the garden...

If these sentence fragments are cross-spliced at the break point indicated, then slower responses in phoneme monitoring of /t/ are reported (although the paper provides no details). This delay is ascribed to the inappropriate duration of the affected inter-stress interval, which raises false expectations by the listener about when the following stressed syllable will occur.

Grabe and Warren (1995, p. 104) present some additional support for the timing-expectancy hypothesis, based on stress transcription data. Under certain conditions, word stress can shift back in time, usually to the initial syllable of that word. If such a stress-shifted target word is presented without context (e.g. IDeal in 4a), three trained judges report perceived stress-shift in 38% of their responses. If the preceding sentence context is provided (4b), stress shift is perceived in 67% of their judgments, rising to 100% if the whole sentence context is presented (4c).

(4) a. ideal

b. As they will never find their ideal partners

c. As they will never find their ideal partners,
they must learn to compromise.

This suggests that listeners notice stress shift better, if they hear more of the sentence context. In our interpretation, the timing patterns in the preceding sentence context lead to a timing (or rhythmic) expectancy that contributes to the perception of stress shift.

Stronger positive evidence for the timing-expectancy hypothesis comes from the Dynamic Attending Theory (Jones, 1976, 1990). In various experiments, Jones and colleagues found large effects of the temporal location of auditory non-speech stimuli, on various tasks requiring auditory attention. For example, Large and Jones (1999) found that listeners performed better in detecting timing discrepancies in target tone sequences, if there was less variation in inter-onset intervals (IOI) in the surrounding tone sequences. They speculate that the attention location was kept constant by an oscillator's attentional pulse; this oscillator can track slow changes in IOI but not fast changes.

In summary, then, there is some weak and controversial support for attentional rhythms from studies of speech perception, but stimulus timing was poorly controlled in these studies. There is stronger support from perception studies with stricter control over stimulus timing, but using non-speech materials and tasks. The experiment presented below attempts to fill the obvious gap between the above studies, by investigating the effect of stimulus timing on the perception of speech, using the phoneme monitoring paradigm (cf. Buxton, 1983; Pitt & Samuel, 1990;

Connine & Titone, 1996; Finney, Protopapas, & Eimas, 1996).

This effect can be assessed for English by varying the *regularity* of inter-stress intervals as a first factor. Does regular timing facilitate spoken-word perception? If words are spoken with constant inter-stress intervals, then perception of the target word should be facilitated, due to good entrainment of the attentional oscillator to salient time points in the auditory stimulus. If, however, the words in a sequence are spoken with irregular timing, with varying inter-stress intervals, then spoken-word perception should be hindered (cf. Large & Jones, 1999). Listeners' attentional pulse will then often miss the phonetically most salient part of the intended target word, or it will be too wide to make any perceptual difference. Hence, the Dynamic Attending Theory would predict a large effect of timing regularity on spoken-word perception.

By contrast, the *metrical expectancy* of the word sequence will constitute the second factor in the present experiment, as in previous studies (Shields et al., 1974; Meltzer et al., 1976; Pitt & Samuel, 1990). If metrical expectancy facilitates spoken-word perception, then a trochee should be perceived faster in a trochee sequence (same-meter) than in an iambes sequence (different-meter), ceteris paribus. The use of word sequences allows us to independently vary listeners' metrical expectancy about the stress pattern in the target word. If listeners are indeed sensitive to the linguistic content of the speech signal (i.e. to the strong–weak syllable distinction within words), then one would predict a large effect of metrical expectancy.

Method

Targets

All English plosive consonants /p, t, k, b, d, g/ were used as target phonemes. The target words used by Pitt and Samuel (1990) were not useable for the present study, because of their ambiguous stress pattern (e.g. *permit*). English target words selected for the present study must have unambiguous stress patterns, to allow unambiguous time alignment of a target word in a multi-word sequence. For real target stimuli, 36 disyllabic words were selected, with a plosive target consonant in the onset of the stressed syllable. Between the target words, *stress pattern* was also varied, to allow both trochaeic (e.g. *camel*) and iambic target words (e.g. *raccoon*) and preceding context words. In addition, 36 similar fillers contained the target phoneme in their unstressed syllable (e.g. *bacon, police*). All targets and fillers were either nouns, adjectives or adverbs. Verbs were excluded throughout, since these generally have an iambic stress pattern, and listeners seem to have statistical knowledge of the distinctive stress patterns of nouns vs. verbs (Kelly, 1988; Kelly & Bock, 1988). Targets and fillers are listed in Table 1.

Insert Table 1 about here

Word sequences

Each target word was presented in 4 conditions defined by the two within-word factors, viz. timing regularity and metrical expectancy. These 4 conditions were manipulated by varying the multi-word sequence in which the target or filler word was embedded. For each target word, the same token was used in all 4 conditions. As will be explained below, metrical expectancy was manipulated by varying the stress pattern of the other words in the sequence. Timing regularity was manipulated by varying the temporal alignment between the words in the sequence.

Metrical expectancy in word sequences. For real targets, each trial consisted of a sequence of words, with the length of the sequence ranging between 5 and 7 words. Real target words always occurred at the 5th serial position in the sequence (Pitt & Samuel, 1990). In the "same-meter" condition, all words in the sequence had the same stress pattern as the target word. In the "different-meter" condition, the words preceding and following the target words all had a stress pattern opposite from that of the target word.

For matching fillers, each trial consisted of a similar sequence, with the length of the sequence ranging between 4 and 7 words. Matching fillers occurred at either the 4th or 6th serial position in the sequence. For each participant, this arrangement yields targets at the 4th, 5th and 6th serial position in a 1 : 2 : 1 ratio, in order to avoid strategic effects. For each filler word, trials were constructed for both the "same-meter" and the "different-meter" conditions.

In addition, 36 foil sequences were constructed, balanced over the 6 target phonemes. These foils did not contain (a word with) the target phoneme. The length of foil sequences ranged between 4 and 7 words. Foil sequences were similar to those for real targets and matching fillers, in that half of the sequences were "same-meter" and the other half "different-meter".

In total, 380 additional words (again only nouns, adjectives or adverbs) were selected for constructing all multi-word sequences. Words were selected with the help of several dictionaries and word lists (a.o. Ferguson, 1985; Blockcolsky, Frazer, & Frazer, 1987).

A multi-word sequence containing a real target did not contain other words that rhyme with the intended target; otherwise, phonological priming might affect recognition of the target word (e.g. Slowiaczek, McQueen, Soltano, & Lynch, 2000). Pitt and Samuel (1990) excluded consonants that were phonetically similar (in manner of articulation) to the target phoneme, in the word immediately preceding the target word. This precaution was impossible in the present experiment, due to limitations of the word set, and due to all target phonemes being plosives. Voicing alternates of the target phoneme were excluded in the preceding word, however: if the target was /k/, for example, then the preceding word did not contain /g/ either.

All 452 words were listed in random order, and then recorded on digital audio tape (DAT) by a female native speaker of American English. The recordings were re-digitised at 22050 Hz (16 bits), and each word was excised and stored in a separate audio file.

In order to time-align or synchronize each word, its *alignment point* had to be determined. This point marks the onset of the stressed vowel (Allen, 1972); it is similar to the perceptual center (P-center) of a word (Morton, Marcus, & Frankish, 1976; Marcus, 1981), modified for disyllabic words. A word's alignment point was determined using the following algorithm (Scott, 1993; Cummins & Port, 1998; Quené & Port, 2002). First, the signal was de-emphasized with –6 dB/octave, to enhance the vowel region of the speech spectrum. Second, the intensity contour of the filtered signal was computed (window length 1/150 s, window shift 0.01 s), and thirdly the first-order derivative of the intensity contour was computed. For each word the intensity peak was determined, the time of which is denoted as t_{peak}. The point preceding this peak at which the intensity has its steepest rise (i.e., a peak in its derivative) is denoted as t_{rise}. Finally, the alignment point was calculated as the time point halfway between t_{peak} and t_{rise}, rounded to ms. All alignment points were verified manually by the first author. For real target trials, the alignment point always coincided with the offset of the immediately preceding target phoneme. In target words like ca.mel and ra.ccoon, for example, the onset of the stressed vowel coincided with the offset of the target phoneme /k/. The resulting alignments points in target words were in fact highly correlated with the P-centers as defined by Marcus (1981, with $a = 0.65, b = 0.25$, for trochees $r = .86$; for iambes $r = .97$). (The small differences between the two measures may be ascribed to differences in syllable structure: some target words had missing onset, missing coda, ambisyllabic consonants, etc.)

Timing regularity in word sequences. Multi-word sequences were concatenated automatically, using the pre-determined alignment points at the onset of stressed vowels. For regular sequences, the inter-stress interval was always 1.1 s, an optimal value established in pilot studies. For irregular sequences, the duration of this interval was sampled randomly from a uniform distribution, ranging between 0.6 s and 1.6 s (i.e. within 1.1 ± 0.5 s). Pilot studies showed that the resulting timing was indeed irregular or jittery. During concatenation, the appropriate inter-word interval was calculated from the required inter-stress interval, taking into account the speech durations after the alignment point in the preceding word, and before the alignment point in the following word. Each concatenated multi-word sequence was stored as a separate audio file.

Figure 1 summarizes the word sequences (along the horizontal axis) for each experimental condition (along the vertical axis, abbreviated in the left margin). A particular trochee target word was presented in four conditions: either preceded by trochees (same-meter, rows 1 and 2) or by iambes (different-meter, 3 and 4), with regular (1 and 3) or irregular timing (2 and 4). Similarly, a particular iambic target word was either preceded by iambes (same meter, rows 5 and 6) or by trochees (7 and 8), again with regular (5 and 7) or irregular timing (6 and 8).

Insert Figure 1 about here

Participants, design, and procedure

Participants were 32 native speakers of American English, with self-reported normal hearing. Most of them were students at Indiana University, in Bloomington, Indiana. They were paid $5 for participating in this study. Participants were divided at random into 4 groups of 8 listeners each.

The 4 (within-word) experimental conditions of the 36 multi-word sequences were rotated over the 4 listener groups according to a Latin square, counterbalanced by stress position of the target word. Each listener heard only one condition of a particular target sequence, and each listener heard a particular target word only once. In addition to the 36 target sequences, each listener heard the same 36 filler and 36 foil sequences.

All 108 multi-word sequences were blocked by their target phoneme, and ordered at random within each block. The order of blocks of target phonemes was rotated across listener groups.

Each subject was tested individually in a quiet room. Listeners were seated with a computer screen in front of them. The speech material was presented over closed headphones (Shure SM2). At the onset of a new block, the new target phoneme was presented visually on the computer screen, with two example words, e.g. (5a); the target phoneme was word-initial in the first (monosyllabic) example word and word-medial in the second (disyllabic) word. At the same time, an auditory cue specifying the new target phoneme with the first, monosyllabic example word was presented, e.g. (5b). These cues were spoken by the second author, i.e., by a voice different from the other materials. Example words were not used elsewhere in the experiment.

(5) a. –p– as in Please, comPass

b. now listen for P as in "please"

Listeners were seated with a button box under their dominant hand, and the computer keyboard under their other hand. They were instructed to press a button as quickly as possible (with their dominant hand), after hearing the pre-specified speech sound in a word in a sequence. It was pointed out that they should also respond to this sound if it does not appear in the spelling of the word, as for /k/ in account.

As an additional task, listeners had to make a semantic classification of the target-bearing word in a stimulus sequence. This was done to ensure post-lexical responses to the phoneme monitoring task (cf. Pitt & Samuel, 1990). If the word containing the target sound referred to a living object (human, animal or plant), then listeners should press the space bar of the computer keyboard (with their non-dominant hand). This additional response should be given only after pressing the reaction time button. For technical reasons, responses on this additional task were neither recorded nor analyzed.

Each item consisted of a multi-word sequence, preceded by a sinewave beep (880 Hz, 200 ms) and a 400 ms silent interval. Reaction times were computed during the experiment, using the button pressing event and the appropriate alignment point in the stimulus audio file for that item. After each item, there was a 2500 ms time window for responses, followed by a 500 ms "dead" interval before the next item.

Before the actual test session, listeners were presented with a practice session consisting of 12 items (multi-word sequences, both fillers and foils), distributed over the 6 target phonemes, so they heard all cue sentences during practice. They could then ask for clarification if necessary. The first 6 items of the actual test session were fillers allowing listeners to "warm up" to their task. These were excluded from further analysis. Total time of a test session was about 20 minutes.

Results

Listeners failed to detect the target phoneme in 42 instances. These misses mostly occurred with two target words (*coward, guitar*), which yielded miss rates over 75%. Inspection revealed that the target phoneme in these words was indeed barely audible, due to articulatory weakening. Responses for these words were removed from the data. In order to maintain a balanced experimental design, two other target words were also removed from the data (*captain, brigade*, yielding next-highest miss rates). Hence only 32 out of 36 target words remained for further analysis. In addition, a few of the hit responses were premature, i.e. initiated *before* the target phoneme was audible in the speech signal. These responses were also discarded from further analysis. For the 32 remaining target words, the pooled error rate of misses and premature responses was 7% (73/1024).

RTs were transformed to their logarithmic values, in order to remove the intrinsic positive skew and non-normality of their distribution (Keene, 1995; Limpert, Stahel, & Abbt, 2001). Log-transformed RTs were then analyzed by means of repeated measures ANOVAs by listeners and by words, respectively (Clark, 1973). For this purpose, missing observations were replaced by the log-RT average for the corresponding listener or target word, in the corresponding condition. Timing regularity and metrical expectancy were included as within-listener and within-word factors. Stress pattern was included as a within-listener, between-word factor. The main effect of metrical expectancy, between same-meter (561 ms) and different-meter sequences (569 ms) was not significant [$F_1(1, 31) < 1$, n.s.; $F_2(1, 28) < 1$, n.s.]; the interactions involving this factor were not significant either. This factor was therefore ignored in subsequent ANOVAs (effectively pooling the variances of the main effect and interactions with appropriate error variances), to improve the power of the statistical tests. Means and standard errors of the pooled log-transformed RTs are summarized in Figure 2.

Insert Figure 2 about here

The pooled ANOVAs showed an interaction between timing regularity and stress pattern, illustrated in Figure 2, which was significant only in the analysis by listeners, but not in the combined analyses [$F_1(1, 31) = 9.8, p = .004, {\eta_p}^2 = .241$; $F_2(1, 28) = 2.8, p = .106, {\eta_p}^2 = .091$; $minF'(1, 26) = 2.716, p > .1$]. The timing regularity yielded a consistent, medium-sized, and highly significant effect: irregular

598 ms vs. regular 534 ms [$F_1(1,31) = 42.7, p < .001, {\eta_p}^2 = .579$, Cohen $f = .28$; $F_2(1,28) = 10.1, p = .004, {\eta_p}^2 = .264$, Cohen $f = .15$; $minF'(1,41) = 8.146, p = .007$]. Although the effect of timing regularity varied for the two stress patterns, separate ANOVAs showed that the effect was significant both for trochees [$F_1(1,31) = 5.6, p = .025, {\eta_p}^2 = .152$; $F_2(1,12) = 21.2, p = .001, {\eta_p}^2 = .639$] and for iambes [$F_1(1,31) = 38.8, p < .001, {\eta_p}^2 = .556$; $F_2(1,15) = 12.2, p = .003, {\eta_p}^2 = .448$]. The main effect of stress pattern, between trochees (600 ms) and iambes (532 ms), was significant only in the analysis by listeners, but not in the combined analyses [$F_1(1,31) = 21.6, p < .001, {\eta_p}^2 = .411$, Cohen $f = .22$; $F_2(1,28) = 3.1, p = .089, {\eta_p}^2 = .100$, Cohen $f = .09$; $minF'(1,26) = 2.716, p > .1$]. Note that stress pattern was a between-words factor; the large variation in log-RT between target words [$F_2(31,919) = 5.8, p < .001$, $\eta^2 = .164$; Cohen $f = .42$] has reduced the power for this effect.

Discussion

The results above clearly show that timing regularity contributes significantly to spoken-word perception. If spoken words in a list are temporally aligned to regular inter-stress intervals, then words in such a list are relatively easy to perceive. If aligned to an irregular timing pattern, then words are more difficult to perceive. Hence, results confirm the main prediction for this study, and they provide further support for the Dynamic Attending Theory (Jones, 1976, 1990) and for the perceptual-isochrony hypothesis (Lehiste, 1980). The main effect of timing regularity in spoken-word perception is strikingly similar to the same effect in temporal discrimination in tone sequences (Large & Jones, 1999), which tested the same theory. The present results also provide empirical support for the PolySP model (Hawkins, 2003), in which speech perception is guided by rhythmic and temporal relations in the incoming speech signal.

According to the Dynamic Attending Theory, listeners' attention is focused on periodic time points, with entrainment between the dominant stimulus period and the attentional period. In the regular-timing conditions of the present experiment, the attentional period was entrained to the inter-stress interval. Consequently, listeners' attentional pulse coincided with (and focused on) the stressed syllable, yielding faster RTs. In the irregular-timing conditions, by contrast, such entrainment to the auditory stimulus was not possible. Hence, listeners' attentional pulse, if there even was one, tended to miss the most salient and informative part of a spoken word. This significantly reduced listeners' speed of perceiving the target word. Listeners had more trouble in perceiving words spoken with irregular inter-stress timing.

Interestingly, no effect of metrical expectancy was observed. It does not matter for spoken-word perception whether the stress pattern of the target word is the same as, or different from the preceding words in a list. This strongly indicates that in itself, the sequencing of stressed (strong) and unstressed (weak) syllables is not relevant for spoken-word perception. Anomalies in an iambus–iambus–iambus or trochee–trochee–trochee sequence have no detrimental effect on the speed of perception of the target word, contrary to the findings reported by Pitt and Samuel (1990). Listeners' attention appears to be focused not by the linguistic content of the speech signal (i.e., not by the strong–weak pattern within words), but clearly by the timing of the speech signal, i.e., by when the strong syllables are spoken.

Within a stimulus sequence, the preceding context words all had identical meter, while the meter of the subsequent target word was either the same as or different from those preceding words. Hence, metrical expectancy was manipulated to be either strongly towards (same meter) or strongly against (different meter) the target word. The timing expectancy, by contrast, was manipulated to be either strongly towards the target word (regular timing), or neutral (irregular timing). Given the weaker expectancies induced by the timing manipulations, one might expect that the timing effect would be smaller than the metrical effect. In fact, however, the opposite was observed: a large effect of timing regularity, and no effect of metrical expectancy.

Although the main effect of stress pattern was not significant, this was likely to have been due to low power, as argued above. The observed stress effect probably reflects a true RT advantage of iambic target words over trochaeic words in this study. This advantage is opposite to the usual pattern for English, where trochees are reported to be perceived faster and easier than iambes (Taft, 1984; Grosjean & Gee, 1987; Cutler & Norris, 1988; Gow & Gordon, 1993; Mattys & Samuel, 2000). In the present study, however, RTs have been measured from the offset of the target phoneme, and hence by definition from the onset of the stressed vowel. This temporal alignment point fell in the first syllable for trochees (mean 95 ms after word onset), but in the second syllable for iambes (mean 330 ms after word onset). This should yield a trivial difference in RT. For iambes, word perception is initiated by the first syllable, and already underway by the time the second, stressed syllable arrives, yielding shorter RTs for iambes (but cf. Mattys & Samuel, 2000). In this way, the observed RT difference between trochees and iambes corroborates that listeners in this study were indeed engaged in higher-level spoken-word perception, and not only in lower-level phonetic processing.

Even though timing regularity has a significant overall effect on RTs, this effect appears to be stronger for iambes than for trochees. In our interpretation, these interactions are caused by the listeners' metrical strategy in spoken-word perception.

For English listeners, trochee words are generally easier to perceive than iambes. This difference has been ascribed to a so-called 'metrical' strategy, by which English listeners assume that words begin with a stressed syllable (Taft, 1984; Grosjean & Gee, 1987; Cutler & Norris, 1988; Gow & Gordon, 1993; Mattys & Samuel, 2000). This assumption is indeed correct in 90% of English content words (Cutler & Carter, 1987). The strong syllable at word onset is used to drive spoken-word perception, by activating a tentative set of appropriate candidates in the mental lexicon (Norris, McQueen, & Cutler, 1995). Stressed syllables are more effective than unstressed syllables for this purpose, because their vowels are less reduced and hence more salient than the vowels in unstressed syllables, at least in English and in Dutch (Kager, 1989; Sluijter, 1995; Sluijter & Van Heuven, 1996; Quené & Koster, 1998).

This metrical strategy would be effective in this English study for trochee words. Listeners apparently have employed this strategy in perceiving trochee target words, yielding a ceiling effect which leaves little room for further improvement by timing regularity. For iambic target words, however, listeners need to ignore the metrical strategy. Listeners then apparently used the timing regularity as an aid in word perception. If present, such regularity provided a strong and helpful timing expectancy about the most salient and informative part of the word. If absent, however, as in irregular sequences, listeners did not know when to expect the target word, and this slowed down their word perception considerably. Together, the interaction effects suggest that English listeners' metrical strategy may outrank timing regularity, for the purpose of spoken-word recognition. Further studies are required to investigate how these two factors interact in other languages, like French, Spanish, Catalan, or Japanese, whose listeners have a weaker or no tendency for basing word perception on anticipated stress cues (e.g. Cutler, Mehler, Norris, & Segui, 1986; Dumay, Frauenfelder, & Content, 2002; Sebastián-Gallés, Dupoux, Seguí, & Mehler, 1992; Soto-Faraco, Sebastián-Gallés, & Cutler, 2001; Otake, Hatano, Cutler, & Mehler, 1993; Otake, Hatano, & Yoneyama, 1996).

Skeptics might object to the main effect of timing regularity reported here, claiming that the word sequences in this experiment are highly artificial, and that effects reported for this "chanted poetry" do not generalize to connected speech (Björn Merker, personal communication). This generalizability question can be addressed easily, by repeating the present study with connected-speech stimuli instead of word sequences. Rhythmic relations between inter-phrase and inter-stress intervals should then be taken into account (at least in English, Dutch, and related languages; Buxton, 1983; Hawkins, 2003; Kohler, 2003). But many types of speech are in fact similar to the word lists used in this study: reading out a number string, a cook listing the ingredients for a recipe, taking a roll call, etc. The effects reported here would certainly generalize to these real-life types of speech behavior.

In conclusion, then, listeners' expectancy of the timing of stressed syllables does indeed contribute to spoken-word perception, as foreseen by Lehiste (1980) and others. English words with stressed syllables aligned at regular time points are perceived better than the same words in an irregularly timed sequence. These results support the Dynamic Attending Theory (Jones, 1976, 1990; Large & Jones, 1999): speech timing helps the listener to build temporal expectancies about when important phonetic information is likely to occur. Orators since classical times have known that speech rhythm contributes to the success of speech communication (e.g. M. Tullius Cicero; *De Oratore*, Book 3, 173–181; see May and Wisse (2001)). Regular timing of stressed syllables, or speech rhythm, helps to get a spoken message across to the listener, because such regularity helps the listener to attend to the speech signal and to its linguistic content.

References

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

Beckman, M. (1986). Stress and non-stress accent. Dordrecht: Foris.

Blockcolsky, V. D., Frazer, J. M., & Frazer, D. H. (1987). 40,000 selected words / organized by letter, sound, and syllable (2nd ed.). Tucson, AZ: Communication Skill Builders.

Buxton, H. (1983). Temporal predictability in the perception of English speech. In A. Cutler & D. R. Ladd (Eds.), Prosody: Models and measurements (pp. 111–121). Berlin: Springer.

Clark, H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. J. Verbal Learning and Verbal Behavior, 12, 335–359.

Connine, C. M., & Titone, D. (1996). Phoneme monitoring. Language and Cognitive Processes, 11(6), 635–645.

Cummins, F., & Port, R. (1998). Rhythmic constraints on stress timing in English. Journal of Phonetics, 26(2), 145–171.

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

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. Journal of Memory and Language, 25, 385–400.

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

Cutler, A., & Van Donselaar, W. (2001). 'Voornaam' is not a homophone: Lexical prosody and lexical access in Dutch. Language and Speech, 44, 171-195.

Dumay, N., Frauenfelder, U. H., & Content, A. (2002). The role of the syllable in lexical segmentation in French: Word-spotting data. Brain and Language, 81(1-3), 144–161.

Ferguson, R. (1985). The Penguin Rhyming Dictionary. Hammondsworth, U.K.: Penguin.

Finney, S. A., Protopapas, A., & Eimas, P. D. (1996). Attentional allocation to syllables in American English. Journal of Memory and Language, 35(6), 893–909.

Gow, D. W., & Gordon, P. C. (1993). Coming to terms with stress: Effects of stress location in sentence processing. Journal of Psycholinguistic Research, 22(6), 545–578.

Grabe, E., & Warren, P. (1995). Stress shift: do speakers do it or do listeners hear it? In B. Connell & A. Arvaniti (Eds.), Phonology and phonetic evidence: Papers in laboratory phonology iv (p. 95-110). Cambridge: Cambridge University Press.

Grosjean, F., & Gee, J. (1987). Prosodic structure and spoken word recognition. *Cognition*, *25*, 157–188.

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics*, *31*, 373–405.

Jones, M. R. (1976). Time, our lost dimension: toward a new theory of perception, attention, and memory. *Psychological Review*, *83*, 323–355.

Jones, M. R. (1990). Learning and the development of expectancies: an interactionist approach. *Psychomusicology*, *9*(2), 193–228.

Kager, R. (1989). *A metrical theory of stress and destressing in English and Dutch.* Dordrecht: Foris.

Keene, O. N. (1995). The log transformation is special. *Statistics in Medicine*, *14*, 811–819.

Kelly, M. H. (1988). Rhythmic alternation and lexical stress differences in English. *Cognition*, *30*(2), 107-137.

Kelly, M. H., & Bock, K. (1988). Stress in time. *Journal of Experimental Psychology: Human Perception and Performance*, *14*, 389-403.

Kohler, K. J. (2003). Domains of temporal control in speech and language: From utterance to segment. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 7–10). Barcelona.

Large, E. W., & Jones, M. R. (1999). The dynamics of attending: How people track time-varying events. *Psychological Review*, *106*(1), 119–159.

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.

Lehiste, I. (1980). Phonetic manifestation of syntactic structure in English. Annual Bulletin, Research Institute of Logopedics and Phoniatrics, University of Tokyo, 14, 1–27.

Limpert, E., Stahel, W. A., & Abbt, M. (2001). Lognormal distributions across the sciences: Keys and clues. Bioscience, 51(5), 341–352.

Marcus, S. (1981). Acoustic determinants of perceptual center (p-center) location. Perception & Psychophysics, 30(3), 247—256.

Mattys, S., & Samuel, A. G. (2000). Implications of stress-pattern differences in spoken-word recognition. Journal of Memory and Language, 42(4), 571–596.

May, J. M., & Wisse, J. (Eds.). (2001). Cicero: On the ideal orator (De Oratore). Oxford: Oxford University Press.

Meltzer, R. H., Martin, J. G., Mills, C. B., Imhoff, D. L., & Zohar, D. (1976). Reaction time to temporally-displaced phoneme targets in continuous speech. Journal of Experimental Psychology: Human Perception and Performance, 2(2), 277–290.

Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual centers (p-centers). 83, 405—408.

Norris, D., McQueen, J. M., & Cutler, A. (1995). Competition and segmentation in spoken-word recognition. Journal of Experimental Psychology: Learning Memory and Cognition, 21(5), 1209–1228.

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or syllable - speech segmentation in japanese. Journal of Memory and Language, 32(2), 258–278.

Otake, T., Hatano, G., & Yoneyama, K. (1996). Speech segmentation by japanese listeners. In T. Otake & A. Cutler (Eds.), Phonological structure and language processing: Cross-linguistic studies (pp. 183–201). Berlin: Mouton de Gruyter.

Pitt, M. A., & Samuel, A. G. (1990). The use of rhythm in attending to speech. Journal of Experimental Psychology: Human Perception and Performance, 16(3), 564–573.

Port, R. F. (2003). On periodically produced speech. Manuscript.

Port, R. F., De Jong, K., Kitahara, M., Collins, D., Leary, A., & Burleson, D. (2002). Temporal attractors in rhythmic speech. Manuscript.

Quené, H., & Koster, M. L. (1998). Metrical segmentation in Dutch: Vowel quality or stress? Language and Speech, 41(2), 185–202.

Quené, H., & Port, R. F. (2002). Rhythmical factors in stress shift. In M. Andronis, E. Debenport, A. Pycha, & K. Yoshimura (Eds.), CLS 38: Papers from the 38th meeting of the Chicago Linguistic Society. (Vol. 1: Main Session, pp. 549–562). Chicago: Chicago Linguistic Society.

Scott, S. K. (1993). P-centers in speech: An acoustic analysis. PhD thesis, University College London.

Sebastián-Gallés, N., Dupoux, E., Seguí, J., & Mehler, J. (1992). Contrasting syllabic effects in catalan and spanish. Journal of Memory and Language, 31(1), 18–32.

Shields, J. L., McHugh, A., & Martin, J. G. (1974). Reaction time to phoneme targets as a function of rhythmic cues in continuous speech. Journal of Experimental Psychology, 102(2), 250-255.

Slowiaczek, L. M., McQueen, J. M., Soltano, E. G., & Lynch, M. (2000). Phonological representations in prelexical speech processing: Evidence from form-based priming. Journal of Memory and Language, 43(3), 530–560.

Sluijter, A. M. C. (1995). Phonetic correlates of stress and accent. Dordrecht: Foris.

Sluijter, A. M. C., & Van Heuven, V. J. (1996). Acoustic correlates of linguistic stress and accent in Dutch and American English. In Proceedings of the International Conference on Spoken Language Processing (Vol. 2, p. #602). Philadelphia.

Soto-Faraco, S., Sebastián-Gallés, N., & Cutler, A. (2001). Segmental and suprasegmental mismatch in lexical access. Journal of Memory and Language, 45(3), 412–432.

Sturges, P. T., & Martin, J. G. (1974). Rhythmic structure in auditory temporal pattern perception and immediate memory. Journal of Experimental Psychology, 102(3), 377–383.

Taft, L. (1984). Prosodic constraints and lexical parsing strategies. PhD thesis, University of Massachusetts.

Van Leyden, K., & Van Heuven, V. J. (1996). Lexical stress and spoken word recognition: Dutch vs. English. In C. Cremers & M. d. Dikken (Eds.), Linguistics in the Netherlands 1996 (pp. 159–170). Amsterdam: John Benjamins.

Author Note

Hugo Quené, Utrecht institute of Linguistics OTS, Utrecht University; Robert F. Port, Department of Linguistics, Indiana University.

The research reported here was supported by the Fulbright Visiting Scholar program. Our thanks are due to Theo Veenker and Paul Styles for technical assistance, to Tiffany Kershner and Rose Vondrasek for reading stimulus materials, to Karin Wagenaar for invaluable assistance in running the experiment, to Jaap Wisse for advice on Cicero's work, and to Sieb Nooteboom, Ilse Lehiste, Lee Wurm, Mark Pitt, and Randy Diehl for helpful comments on previous versions of this paper.

Correspondence concerning this article should be addressed to: Hugo Quené, Utrecht institute for Linguistics OTS, Trans 10, 3512 JK Utrecht, The Netherlands, e-mail hugo.quene@let.uu.nl.

Table 1

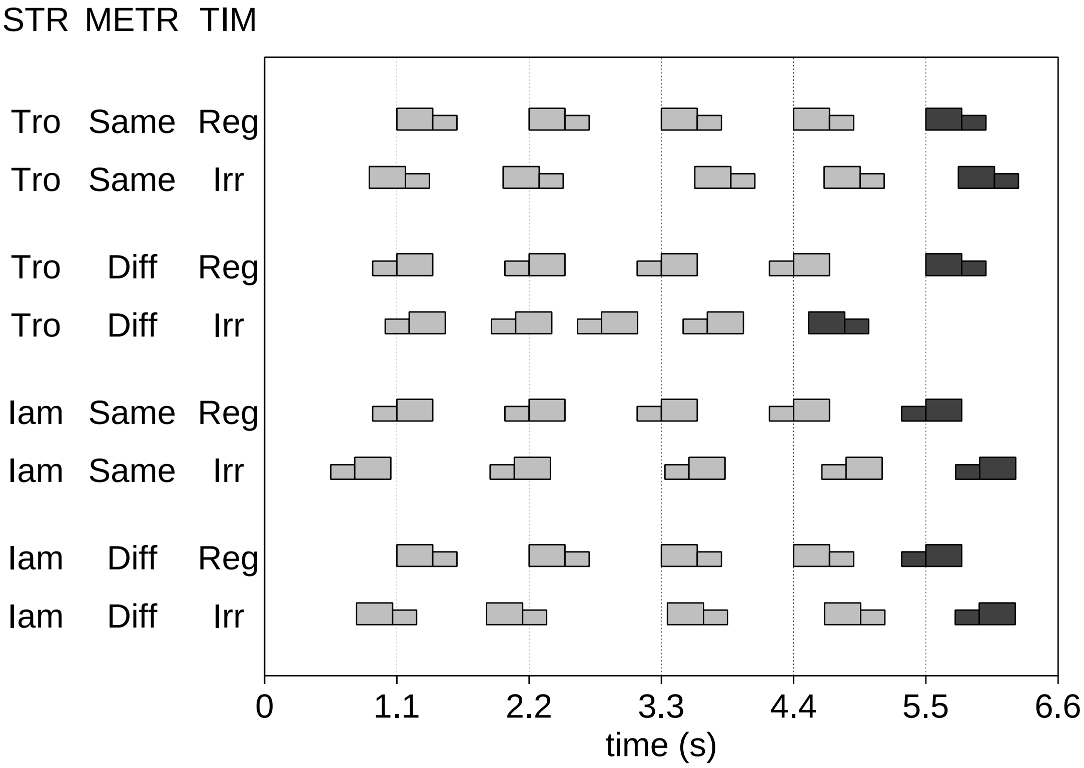
List of real target words and matching fillers, broken down by stress pattern. Target phonemes are indicated by corresponding uppercase characters.

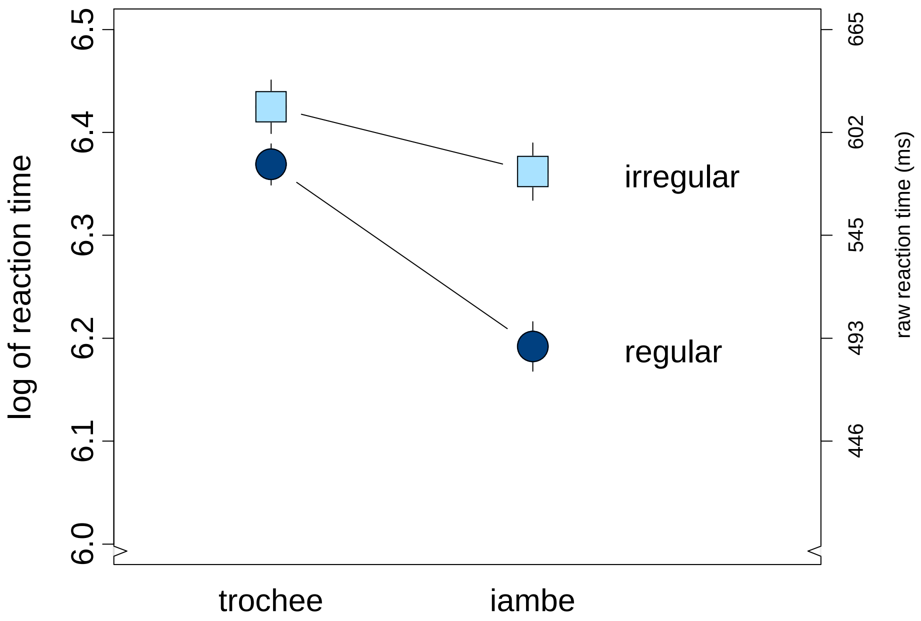
type	stress	words
real target	trochee	Pirate, Pocket, Timber, Transit, Camel, Cotton, Captain, Cousin, Coward, Bandit, Bargain, Biscuit, Bounty, Diaper, Distance, Garden, Gospel, Gutter
	iambus	camPaign, harPoon, traPeze, canTeen, carToon, guiTar, plaToon, ponToon, rouTine, raCCoon, caBoose, oBese, taBoo, tromBone, caDet, orDeal, laGoon, briGade
matching filler	trochee	asPect, camPus, carPet, trumPet, curTain, cusTom, neuTral, baCon, blanKet, poKer, roCKet, amBer, caBin, neighBor, burDen, vanDal, suGar, waGon
	iambus	Parade, Parole, Police, Precise, Today, Typhoon, Career, Cassette, Corvette, Balloon, Bazaar, Beret, Buffoon, Disease, Dragoon, Gazelle, Gazette, Grenade

Figure Captions

Figure 1. Schematic overview of the stimulus conditions, broken down by stress pattern (STR, trochaeic vs. iambic target words), metrical expectancy of target word (METR, same vs. different from preceding context) and timing regularity (TIM, regular vs. irregular). Each box represents a syllable; larger boxes correspond to stressed syllables; target words are marked here by darker shade. Context words following the target word (in some sequences) are left out for clarity. Words were aligned to the onset of its stressed vowel.

Figure 2. Means and standard errors of log-transformed reaction times, broken down by stress pattern (trochee vs iambe) and timing regularity (regular vs irregular).


STR METR TIM
Tro Same Reg
Tro Same Irr
Tro Diff Reg
Tro Diff Irr
Iam Same Reg
Iam Same Irr
Iam Diff Reg
Iam Diff Irr
0
1.1
2.2
3.3
4.4
5.5
6.6
time (s)


log of reaction time
6.5
6.4
6.3
6.2
6.1
6.0
665
602
545
493
446
raw reaction time (ms)
irregular
regular
trochee
iambe