Perceptual organization in intonational phonology: A test of parallelism

Laura C. Dilley
Departments of Psychology and Linguistics
Ohio State University
Columbus, OH 43210
Email: dilley.28@osu.edu

J. Devin McAuley
Department of Psychology
Bowling Green State University
Bowling Green, OH 43403
Email: mcauley@bgnet.bgsu.edu
1. Introduction

A widespread characteristic of the world’s languages is that their prosodic structures exhibit *patterning*. Certain types of cross-linguistic prosodic patterning have received a great deal of attention in phonological theories and thus have been central to the development of theoretical arguments. For example, observations of widespread alternations of strong and weak syllables and phenomena such as stress retraction have provided motivation for metrical stress theory (e.g., Liberman and Prince 1977; Selkirk 1980, 1984; Halle and Vergnaud 1987, Hayes 1995), while reduplicative patterning have provided compelling arguments for correspondence theory within Optimality Theoretic frameworks (e.g., Prince and Smolensky 1993, McCarthy and Prince 1995). Other types of prosodic patterning, by contrast, have received less attention as well as no explanation within phonological theory. One such type of patterning concerns the occurrence of parallel pitch patterns on stressed and unstressed syllables, i.e., the repeated sequential co-occurrence of a high tone with a stressed syllable in combination with a low tone with an unstressed syllable, or vice versa (e.g., L+H* L+H* L+H* or L*+H L*+H L*+H). In spite of this, the existence of patterning in pitch has been noted to be a critical theoretical issue within intonational phonology (e.g., Ladd 1986, Pierrehumbert 2000).

The present paper fills a theoretical gap by developing an account for the occurrence of parallel pitch patterns, such as L+H* L+H* L+H*. Consistent with the Laboratory Phonology approach of seeking explanations for linguistic facts in terms of human perception, action and cognition, our account draws heavily on work done outside of linguistics proper, primarily within general auditory perception. The proposed explanation is tested in a perception experiment, and the results provide support for our
proposal. Overall the present research provides insight into a variety of facts in prosodic phonology and phonetics, including prosodic constituency and notions of the prosodic hierarchy, and may provide insight into word segmentation and language acquisition as well.

1.1 Repetition of accent sequences

It has widely been observed that intonational accents tend to form repeating sequences (Palmer 1922; Pike 1945; Schubiger 1958; Kingdon 1958; Schachter 1961; Crystal and Quirk 1964; Crystal 1969; Halliday 1967; Gibbon 1976; Chafe 1988; Ladd 1986, 1996; Pierrehumbert 2000; Dainora 2001). Ladd (1986) focused on the fact that pre-nuclear accents tend to be of like type in developing his proposal of recursion in prosodic phrasing. This is shown in Figure 1 (modified from Ladd 1986), which illustrates that a typical pattern is either for the prenuclear accents to be identical, or for the nuclear accent to be different. Such repetition has been noted for a variety of languages, including German, Bengali, Japanese, Spanish, Italian, Korean, and French (Beckman and Pierrehumbert 1986, Pierrehumbert and Beckman 1988, Hayes and Lahiri 1991; Jun 1993; Grice 1995; Prieto, van Santen, and Hirschberg, 1995; D’Imperio 2000, Welby 2003). However, neither Ladd (1986) nor other theorists have put forward an explanation for why intonational patterns might tend to repeat in the first place.

In spite of the dearth of accounts for intonational repetition, it is clear that this repetition is central to intonational phonological theory. This is pointed out by Pierrehumbert (2000), who notes that proposals within the autosegmental-metrical (AM) theory of intonation (e.g., Pierrehumbert 1980, Beckman and Pierrehumbert 1986) have
not adequately dealt with facts about intonational repetition. For example, she notes that
the intonational grammar proposed by Beckman and Pierrehumbert (1986) generates 36
accentual combinations for phrases with two accents, and 216 accentual combinations for
phrases with three accents. Pierrehumbert states (p. 27):

“…nothing like the full set [of accents] generated by the grammar has ever
been documented. For three accent phrases, the typical pattern is either to
use the same accent type in all three positions, or else to use one type of
accent in both prenuclear positions, and a different type in nuclear
position…To understand this observation, it is worthwhile to bear in mind
how sparsely languages in general sample the cross-product of the
available units… *we need to work out what factors cause gaps in the set of
intonation patterns observed.*” ¹

Several questions remain unanswered to date. First, why do speakers use repeated
intonation patterns? Second, why do such patterns appear to be so prevalent across the
world’s languages? Third, why is the final accenntual position of an utterance often
marked by a change in the accenntual sequence? Finally, what are the functional
consequences of intonational repetition, both for phonology and for speech
communication? In the following section we develop a proposal that addresses these
questions.

1.2 Some observations from work in auditory perception
In this section we review some of work here which we believe may be relevant to understanding perception of prominence and constituency in language. Research in auditory perception has identified a number of general principles of perceptual organization concerning how auditory events (such as non-speech tones or musical notes) are organized by listeners into patterns of prominence and grouping. One line of work in particular has examined the perceptual consequences of repetition and patterning of pitch and timing events on perceived grouping and accentuation of those events. It is now well-established that regularities in pitch, timing, and amplitude in auditory signals influence the perceived organization of auditory events in sequence (Bolton 1894; Woodrow 1909, 1911, 1951; Jones and Boltz 1989; Boltz 1993; Deutsch 1980; Thomassen 1982; Povel and Essens 1985; Jones and Pfordresher 1997). For example, Woodrow (1911) showed that for a sequence of isochronous pure tones of equal amplitude alternating between a fixed high and fixed low frequency as in (1), listeners tend to hear a binary grouping structure with accents on either the high or low tone, as in (2) or (3), with the accented element beginning the group:

(1) … H L H L H L H L H …
(2) …(H* L) (H* L) (H* L) (H* L) (H*…
(3) …H) (L* H) (L* H) (L* H) (L* H) …
(4) …((H* L H) (L* H L)) ((H* L H) …

This finding is interesting because listeners tend to hear the accentual and grouping structures in (2) or (3) even though there are no local cues to this structure in the signal
(i.e., each tone is equivalent in duration and amplitude, and tone onsets are isochronous). Moreover, listeners have preferences for structures like (2) and (3) over structures like (4), even though (4) is also logically possible. The phenomenon whereby listeners infer structure from repetition in auditory events has been termed “subjective rhythm” in the literature (Handel 1989).

The phenomenon of subjective rhythm has also been demonstrated for tones which alternate merely in their durational and/or intensity characteristics (Woodrow 1951; Povel and Okkerman 1981; Povel and Essens, 1985). If every other tone is lengthened in an otherwise identical sequence of tones, then listeners tend to hear tones grouped by twos with the lengthened tone in each pair ending the group (an iambic grouping). Similarly, if every other tone is increased in intensity then listeners similarly tend to hear tones grouped by twos, but with the increased intensity tone in each pair beginning the group (a trochaic grouping).

Lerdahl and Jackendoff (1983) codified the perceptual tendency to organize sequences of tones into parallel grouping and metrical structures in terms of two separate “parallelism preference rules”. They note on a number of occasions certain analogies in the ways that both music and language seem to use acoustic variation in the same ways in order to signal metrical and grouping (or constituency) structure. For example, they remark (p. 85):

“…[our] preference rules … are reminiscent of the principles governing prosodic features of language. It is well known (see for example Trubetzkoy 1939)… that there are a limited number of discrete ways in
which languages mark the distinction between strong and weak
syllables….That stress and length function as markers of metrical strength
in music as well as in language can hardly be a coincidence. Rather it
seems that we are dealing with a more general cognitive organization that
has manifestations in both musical and linguistic structure.”

Based on these findings from general auditory perception, there is good reason to
believe that repeated patterns in pitch and/or time might generate percepts associated with
grouping and accentuation for syllables, just as for tones. In other words, the findings
above suggest that pitch pattern repetition occurs because such repetition perceptually
generates precisely those structures which are associated with prosodic prominences and
constituent structures.

Building on these observations, we therefore propose a Parallelism principle for
language which combines the two parallelism rules from Lerdahl and Jackendoff’s
(1983) work on music. This principle is stated as follows:

*Parallelism Principle:* When two or more sequences of syllables can be construed as
parallel, they preferably form parallel parts of groups with parallel metrical structure.

According to this principle, syllable sequences which display parallel sequences in
pitch and/or in time will be preferentially perceived as having parallel metrical and
grouping structure. That is, the sequence of tones in (5) will be perceived as in (6) or
possibly (7), but probably not as in (8).
The Parallelism principle furthermore leads us to expect that not only regularities in pitch, but also regularities in time will lead to preferences for certain grouping and metrical structures. Therefore, we expect that the Parallelism principle should also apply to timing patterns in speech, leading listener to hear parallel metrical structure, all other factors being equal. Another way of stating this is that the syllables in speech will be heard as more parallel, the more perceptually regular their timing patterns. It has been often been claimed that speech sometimes sounds perceptually isochronous; Lehiste (1977) has proposed that this is because listeners hear the stresses of speech as more regular than they actually are. More recent work on this topic suggests that there are degrees of perceptual isochrony, ranging from very perceptually isochronous to highly anisochronous (McAuley and Dilley 2004; see also Dilley 1997).

If listeners are shown to perceive sequences of syllables in parallel ways based on pitch and/or time characteristics, it will support the proposed Parallelism principle as a (partial) explanation for pitch accent repetition. (A more complete explanation of accentual repetition based on general auditory perceptual principles is deferred to the Discussion section.) Moreover, a finding that listeners perceive parallel sequences of syllables with parallel structure will fill a critical gap in intonational phonology by providing an explanation for why accents fail to combine freely in sequence.
A perception experiment was conducted to test whether the proposed Parallelism principle provides a viable explanation for pitch accent repetition. This experiment investigated how a preceding prosodic context affected the perceived organization of syllable sequences such as *foot note book worm* which have ambiguous lexical structure. In the experiment, the acoustic characteristics of the final part of the syllable sequence were held constant while the pitch and/or timing characteristics of the prior context were varied. We predicted based on the Parallelism principle that listeners would perceive parallel sequences of pitch and/or timing structure from the contexts, and thus infer different word-level organizations for the final part of the sequence depending on the context.

2. Methods

2.1 Participants

One hundred thirty-eight native speakers of America English completed the experiment in return for credit in an introductory psychology course at the Ohio State University. Participants were at least 18 years of age with self-reported normal hearing and a range of a musical experience. They were randomly assigned to one of three prosodic context conditions: Pitch (n = 57), Duration (n = 40) or Pitch+Duration (n = 41).

2.2 Stimuli and conditions

Twenty eight-syllable target sequences were constructed consisting of two disyllabic trochaic words (e.g., *channel dizzy*) followed by a final four syllable string that could be organized into words in more than one way (e.g., *footnote bookworm* versus *foot notebook worm*, etc.); see Appendix A for the complete set of targets. For each target sequence, we were interested in whether listeners’ organization of the final syllables into
words would be influenced by expectations generated by the pitch and/or duration patterning of the first five syllables (e.g., *channel dizzy foot*). Under one possible organization of each target sequence, the final item was a disyllabic compound word (e.g., *bookworm*), whereas for the contrasting organization, the final item was a monosyllabic content word (e.g., *worm*). In addition to the twenty target sequences, forty filler sequences were constructed which consisted of a mixture of monosyllabic and disyllabic content words. Filler sequences varied in length from six to 10 syllables; the order of disyllabic and monosyllabic items varied across filler sequences. Unlike the target sequences, syllables in filler sequences could only be organized into words in one way (i.e., the lexical structure was unambiguous). Half of the filler sequences ended in a monosyllabic word, and half ended in a disyllabic word.

2.2.1 Pitch condition

In the Pitch condition, the final three syllables (syllables 6, 7 and 8) of each eight-syllable target sequence had a HLH tonal pattern, with one tone per syllable. These syllables were preceded by either a HLHLHL tonal sequence or a LHLHL tonal sequence; see Figure 1. In the case of the HL context, a two-tone (HL) falling sequence was realized on the fifth syllable, whereas in the case of the LH context, a single tone (L) was realized on the fifth syllable. In general, we expected the two pitch contexts to yield different expectations about the perceived grouping of the 7th and 8th syllables (i.e., of the prosodic constituent structure); the type of prosodic constituency expected to be generated is considered in the Discussion section. Specifically, the HL context was expected to generate a ‘HL expectancy’ and induce the perception of a monosyllabic final
word (e.g., *worm*), whereas the LH context was intended to generate a ‘LH expectancy’ and induce the perception of a disyllabic final word (e.g., *bookworm*).

### 2.2.2 Duration condition

In the Duration condition, the final three syllables (syllables 6, 7, and 8) plus the preceding five syllables had a flat F0 pattern. To manipulate the duration context, we varied the duration of the interval between vowel onsets (inter-onset-interval, IOI) of syllables 5 and 6 to correspond to one of two criterion values: a “lengthened” interval or a “shortened” interval, as described below. In general, we expected that that the shortened and lengthened intervals would produce different expectations about rhythmic alternation, inducing different perceptions of the final word. In the lengthened interval context, the IOI formed by syllables 5 and 6 (the “original IOI”) was lengthened to match the average of the two successive IOIs formed by syllables 1, 3, and 5; see Figure 2. In this case, we expected that increasing the IOI by this amount would cause syllables 5 and 6 to each be interpreted as rhythmically strong, which should lead to an expectation of a weak-strong alternation for syllables 7 and 8 (‘weak-strong expectancy’) and to the perception of a monosyllabic final word. In the contrasting shortened context, the IOI formed by syllables 5 and 6 was shortened to a duration corresponding to the average of the two successive IOI’s formed by syllables 6, 7, and 8. In this case, we expected that decreasing the IOI to this value would cause syllables 5, 6, 7 and 8 to be interpreted as strong-weak-strong-weak (‘strong-weak expectancy’) leading participants to perceive a disyllabic final word.

### 2.2.3 Pitch+Duration condition
The Pitch+Duration condition involved a combination of the Pitch and Duration manipulations described above. Specifically, a ‘HL plus strong-weak expectancy’ was generated by pairing the HL tonal context in the Pitch condition with the lengthened duration context in the Duration condition. Similarly, a ‘LH plus strong-weak expectancy’ was generated by pairing the LH tonal context in the Pitch condition with the shortened duration context in the Duration condition. If pitch and duration cues reinforce each other, then combining pitch and duration cues should yield stronger expectations about the organization of the final three syllables into words than in either the Pitch or Duration conditions by themselves. Thus, we expected that the ‘LH plus weak-strong expectancy’ context should more frequently generate disyllabic word responses than the ‘LH expectancy’ or ‘weak-strong’ expectancy contexts, although all of these contexts were predicted to give rise to predominantly disyllabic word responses.

**2.2.4 Stimulus construction details**

Target and filler sequences were read by the first author with a monotone F0 contour and recorded in a quiet room onto a TASCAM DA-30MKII DAT recorder at 16 kHz using an N/D308A ElectroVoice Cardioid microphone connected to a Yamaha MV802 mixer. The digitized utterances were then transferred to a PC for subsequent manipulation. All stimuli were synthesized using a PSOLA algorithm (Moulines and Charpentier, 1992) as implemented in Praat software (Boersma and Weenink, 2002). Within each prosodic context condition, the acoustic characteristics of the final three syllables of each target sequence were held constant. For the duration context manipulation, lengthening was accomplished by increasing the duration of the silence or frication associated with a stop or fricative at the consonantal coda and onset of syllables.
5 and 6, respectively. In the case of *sun/spot* and *friend/ship*, however, the duration was lengthened by splicing in additional pitch periods for the /n/ and increasing the duration of frication for syllables 5 and 6, respectively. All splicing occurred at zero crossings, and the result was checked for naturalness. On average, the original IOI was lengthened by 162 ms. Shortening, in contrast, was accomplished by splicing out part of the rhyme of syllable 5 by removing pitch periods from the vowel nucleus or coda (when voiced) or shortening the duration of the frication or closure (when voiceless). In some cases, the duration of the initial consonant or consonant cluster of syllable 6 was also shortened. Again, all splicing occurred at zero crossings, and the result was checked for naturalness. The average duration by which the IOI was shortened was 104 ms.

Moreover, each filler sequence was synthesized to have both a repeating HL pattern and a LH pattern (in the Pitch and Pitch+Duration conditions) or a flat pattern (in the Duration condition). This gave 80 fillers per condition (40 sequences x 2 intonation patterns x 1 repetition for the Pitch and Pitch+Duration conditions, and 40 sequences x 1 intonation pattern x 2 repetitions for the Duration condition).

2.3 Procedure

At the start of the experiment, participants were randomly assigned to one of the three intonation conditions (Pitch, Duration, or Pitch+Duration). Participants were instructed that they would hear sequences of syllables over headphones and that they should write down the last word they heard in each sequence; six filler sequences served as practice. Following practice, participants listened to 100 sequences (20 targets, 80 fillers) presented in a random order. Half of the targets were paired with a prosodic context that was predicted to induce the report of a disyllabic final word, while the remaining targets
were paired with a prosodic context that was predicted to induce the report of a monosyllabic final word. The specific target-sequence context pairings were counterbalanced across participants so that within a prosodic context condition (Pitch, Duration, or Pitch+Duration), each target sequence occurred in both monosyllabic and disyllabic contexts.

2.4 Predictions

Overall, we predicted that ‘LH’, ‘strong-weak’, and ‘LH plus strong-weak’ expectations generated in the Pitch, Duration and Pitch+Duration conditions, respectively, would induce grouping of the 7th and 8th syllables of each target sequence and result in the report of a disyllabic final word. We will refer to these conditions collectively as disyllabic contexts. In contrast, we predicted that the corresponding ‘HL’, ‘weak-strong’, and ‘HL plus weak-strong’ expectations would induce a distinction between the 7th and 8th syllables of each target sequence and result in the report of a monosyllabic final word. We will refer to these conditions collectively as monosyllabic contexts. Finally, we predicted that combining pitch and duration cues in the Pitch+Duration condition would yield stronger expectations than in either the Pitch or Duration conditions independently.

3. Results

Figure 3 shows proportion of disyllabic final word responses in disyllabic and monosyllabic contexts for the three prosodic conditions (Pitch, Duration and Pitch+Duration). Supporting the Parallelism Principle, significantly more disyllabic final word responses were reported in disyllabic contexts than in monosyllabic contexts for all
three prosodic conditions \( p < 0.01 \). The observed differences in final word reports were smallest for the Duration condition, intermediate for the Pitch condition, and largest for the Pitch+Duration conditions.

To address the possibility that some participants might have shown biases for reporting mono- or disyllabic final words, we used standard signal detection methods to separate perceptual sensitivity to the prosodic context manipulation from any general tendency to report disyllabic or monosyllabic words (Macmillan and Creelman, 1991). For this analysis, a *hit* was defined as the report of a disyllabic final word for target sequences assigned a disyllabic context (i.e., those intended to generate a ‘LH expectancy’, ‘strong-weak expectancy’, or both) and a *false alarm* was defined as the report of a disyllabic final word for target sequences assigned a monosyllabic context (i.e., those intended to generate a ‘HL expectancy’, ‘weak-strong expectancy’ or both). Proportions of hits and false alarms to the target sequences were determined separately for each participant and then used to calculate, \( d' \), a measure of perceptual sensitivity, and \( c \), a measure of response bias.

Figure 4 illustrates average \( d' \) with standard error bars for the Duration, Pitch and Pitch+Duration conditions. In general, \( d' \) scores were highest in the Pitch+Duration condition \( (M = 1.59) \), next highest in the Pitch condition \( (M = 1.17) \) and lowest in the Duration condition \( (M = 0.70) \). Confirming these differences, a one-way between-subjects ANOVA revealed a main effect of prosodic context condition \([F(2,135) = 12.05, p < 0.01, \eta_p = 0.152]\). Pair-wise comparisons of the three conditions using Tukey’s HSD were all significant \( (p < 0.01) \). Moreover, all three conditions were significantly above chance \( (d' = 0) \).
Figure 5 shows the average response criterion, $c$, with standard error bars for the Duration, Pitch, and Pitch+Duration conditions. With this measure, larger values of $c$ correspond to more conservative responses. Based on our specification of hits and false alarms, ‘more conservative’ means there is less of a tendency to make a disyllabic final word response and thereby a greater tendency to make a monosyllabic final word response. In general, it was noteworthy that $c$ scores were relatively close to zero for all three conditions, indicating very little response bias. However, there was a slight tendency to make more monosyllabic final word response in the Duration condition ($c = 0.30$), than in either the Pitch condition ($c = -0.012$) or the Pitch+Duration condition ($c = -0.03$). This result was confirmed by a one-way between-subjects ANOVA on the response criterion, $c$, which revealed a main effect of prosodic context [$F(2,135) = 3.51$, $p < 0.05$, $\eta_p = 0.049$]. Pair-wise comparisons between the three conditions using Tukey’s HSD showed a significant difference between the Duration condition and the two other conditions ($p < 0.05$), but not a significant difference between them ($p = 0.99$).

4. Discussion

The primary theoretical motivation for this paper was to develop an account of widespread repetition of pitch accent sequences, e.g., L+H* L+H* L+H*. As pointed out by Pierrehumbert (2000), this patterning is not predicted by models of the tonal grammar proposed within the AM theoretical framework, which has previously assumed that pitch accents combine freely in sequence. To account for this patterning, we proposed a Parallelism principle which is based on general principles of auditory perception, in
which listeners were claimed to perceive similar syllable sequences as having the same grouping and metrical structure.

We tested the Parallelism principle in a perception experiment where participants listened to ambiguous syllable sequences and reported the final word they heard. Consistent with the Parallelism principle, prosodic context strongly influenced how listeners organized the final syllables into words. In particular, listeners’ reports indicated that they organized the prosodic characteristics of the final syllables so as to be congruent with the patterning established earlier in the sequence, thereby affecting the perceived word structure. Both preceding pitch manipulations (the Pitch condition) and duration manipulations (Duration condition) were effective in eliciting parallel judgments. Moreover, combined cues (Pitch+Duration) were the most effective of all, as shown by the fact that these combined cues produced greater differentiation of disyllabic and monosyllabic final words than either pitch or duration context cues alone.\(^3\) These results extend earlier work by Dilley and Shattuck-Hufnagel (1998, 1999), which have similarly demonstrated effects of perceptual grouping principles and expectation in speech.

In the following discussion, we explore the possibility that the auditory perceptual principles which are the basis for the proposed Parallelism Principle also have the potential to provide a unifying explanation for other phenomena in prosodic phonology which have previously received no coherent account. In the course of this discussion, we hope to also situate our experimental results more firmly within a broader theoretical context of prosodic phonology and its relationship to phonetics.

The remainder of the discussion is organized as follows. In Section 4.1 we elaborate on the perceptual basis for the Parallelism principle. Next, in Section 4.2 we consider the
broader implications of this work for intonational phonology. Finally, in Section 4.3 we consider implications of these perceptual mechanisms for understanding the nature of the phonological hierarchy of prosodic constituents and how it is realized in terms of phonetic structure.

4.1 Perceptual Basis for the Parallelism principle

In the introduction to this paper, we described three related findings in the domain of auditory perception that formed the perceptual basis for a Parallelism principle and which we claim extend to understanding the prevalence of accentual patterning in language. The first finding is that listeners tend to hear structure – prominences and groupings - out of sequences of tones or musical notes. When tones showed a repeating pattern in pitch, amplitude, or durational characteristics, listeners tend to hear the same congruent structure of prominences and groupings on the basis of the repetition (Lerdahl and Jackendoff 1983; Parncutt 1994; Povel and Okkerman 1981; Woodrow 1951). However, as discussed in the introduction, listeners do not require that the tones be differentiated or that they show a repeating pattern in order for structure to be imposed perceptually – prominences and groupings are often perceived even when the tones are identical (Bolton, 1894).

The second finding is that listeners generate expectations about upcoming auditory events on the basis of prior ones. The Parallelism principle states that sequences of events (syllables) which are construed as parallel will be perceived with parallel structure. In order to detect parallel events in a speech stream, it is necessary first to perceive some events and then hold them in memory. Incoming events must then be compared with the
memory of earlier events. This can be thought of as a process of generating an expectation about incoming events. Indeed, there is a large body of evidence that listeners detect regularities in auditory sequences (such as tone sequences) and that they generate expectations about upcoming events on the basis of prior regularities (Jones, 1976; Large and Jones, 1999; McAuley and Jones, 2003).

A third related finding is that patterns in pitch and time have the potential to influence the allocation of attention. Specifically, listeners’ attention may be jointly guided by pitch and timing cues to accentuation. Regularities in pitch and time structure facilitate a rhythmic mode of attending whereby events that occur at expected points in time receive more attention than events that occur at unexpected points in time (Jones 1976; Jones and Boltz 1989; Large and Jones 1999; McAuley and Jones 2003). Supporting this view are studies showing that listeners are better able to detect/discriminate targets events that are on time with respect to an established rhythm compared to unexpected events (Jones, Boltz and Kidd, 1982; Jones, Moynihan, MacKenzie and Puente, 2002; Jones and Yee, 1997; Klein and Jones, 1996).

These general aspects of auditory perception provide some insight about why our experiment came out as predicted, and why, in effect, the Parallelism principle “works”. In our experiment, the preceding context strongly influenced how listeners organized the final syllables into words. In particular, listeners’ reports indicated that they organized the prosodic characteristics of the final syllables so as to be congruent with the patterning established earlier in the sequence, thereby affecting the perceived word structure. Our experiment provided clear support for this principle by showing that listeners’ word
reports reflected the predicted patterns of parallel metrical and grouping (constituency) structure given lexically ambiguous syllable strings.

It is worthwhile to note furthermore that in terms of the general perceptual processes described above, it can be inferred that repeating patterns are in some sense “special”. Repeating patterns in pitch and time are expected to lead to inferences about parallel grouping and prominence patterns. Moreover, repeating patterns in pitch and time are expected to lead to generation of expectations which will be accurate (since incoming patterns will match the expectation generated by previously-occurring patterns). Finally, repeating patterns in pitch and time are expected to lead to decreased attention (or increased habituation) over time, since the pattern will continue to be the same as before.

Next, we consider how these general principles of auditory perception may account for various other kinds of prosodic phenomena.

4.2 Implications for intonational phonology

The motivating observation for this paper was that accentual sequences tend to repeat, a fact which is not predicted by the grammar proposed within the AM theoretical framework and which has previously received no explanation. In this section, we describe in more detail why the Parallelism principle, together with the general perceptual mechanisms which underlie it, provide an account for the cross-linguistic prevalence of repeated pitch patterns. We then describe how the general principles underlying the Parallelism principle can furthermore account for the frequent observation that the last accent may be different. In so doing we fill a gap in the literature by providing an explanation within the AM framework for why pitch accents tend to repeat. Moreover, our account provides support for the original claim of Pierrehumbert (1980) that the same
inventory of accents occurred in nuclear and prenuclear positions. This was a critical feature of the theory which distinguished it from earlier proposals, such as Vanderslice and Ladefoged (1972) and the British intonational school (e.g., Halliday 1967), which proposed different theoretical status for prenuclear and nuclear accents.

According to the Parallelism principle and accompanying perceptual mechanisms, we have already established that repeated patterns in pitch and time are invoke particular biases for hearing grouping and metrical structure. This provides a direct and immediate account for the occurrence of intonational patterns. First, prosodic pattern repetition is expected to perceptually generate exactly the kinds of structures which are so important in phonology, namely, patterns of phrasing and accentuation. In other words, repeated patterns make it easier to perceive accents and constituents. Second, the perception of parallel structure should persist to influence the interpretation of subsequent events. Repetition of similar structures is known to strengthen the perceived organizational structure of events (e.g., Handel 1989). In other words, the Parallelism account suggests that repetition should strengthen the percepts associated with accentuation and constituency.

Another consequence of the Parallelism principle is to explain why certain patterns fail to occur. The parallelism principle predicts, for example, that $L+H^* L^*+H L+H^*$ is an impossible accentual sequence. We know of no counterexamples to this.\(^4\)

The Parallelism principle and associated general mechanisms of auditory perception can also explain the observation that the last (or nuclear) accent in a sequence tends to be different from the accents preceding it (e.g., Ladd 1986, Pierrehumbert 2000). The nuclear accent in an utterance is typically considered to be the strongest; moreover, this
accent is often reported to correspond to the most important information in an utterance (Bolinger 1989, Fernald and Simon 1984). To our knowledge, no explanation for these data has been put forward. However, one possible explanation for these findings is that a speaker’s use of repeated prosodic patterns generates expectations about the grouping and accentual structure of upcoming events in the mind of a listener. Then a change of these patterns is expected to draw a listener’s attention to the position in the sequence associated with the violation (Jones 1976). This suggests that the phonology licenses sequences which generate maximal attentional resources precisely at positions which are of the greatest structural and informational importance, namely at nuclear accentual position, through the mechanisms of expectancy generation and violation.

The present work therefore fills a critical gap in the AM theoretical framework by describing one principle which appears to limit possible accentual sequences in language. A critical assumption of AM theory as originally proposed in Pierrehumbert (1980) was that the nuclear and prenuclear accent inventories are the same. Other theories (Vanderslice and Ladefoged 1972, Halliday 1967) had to resort to the assumption that nuclear and prenuclear elements had different theoretical status. By providing a perceptual reason for accentual sequences to show repeating patterns, we are able to provide support for the AM theoretic framework overall.

It is worthwhile briefly considering whether other phenomena within intonational phonology might be accounted for by the general perceptual mechanisms we have discussed above. For example, the cross-linguistic tendency of accents to drift down in sequence would seem to involve patterning and hence be expected to generate expectations for the trend to continue (Pierrehumbert 1980, Cooper and Sorensen 1981,
Prieto, van Santen, and Hirschberg 1995, Laniran and Clements 2003). The flip side of this observation is that a number of phenomena would also seem to be readily described in terms of a change from a pattern, such as the final accent in a sequence being lower than that expected based on the extrapolation (Liberman and Pierrehumbert 1984, Truckenbrodt 2004), accentual “upstep” (Truckenbrodt 2002), or pitch range reset (e.g., Ladd 1988). Full consideration of the possible applicability of these general perceptual principles is outside the scope of this paper.

Next, we consider how our results fit within a general framework of the prosodic hierarchy (e.g., Nespor and Vogel 1986, Beckman and Pierrehumbert 1986, Hayes 1989, Beckman and Edwards 1990, 1994). Moreover, we consider some implications of our results for understanding the phonetics-phonology interface as it regards this hierarchy.

4.3 Implications for the phonetic basis of phonological constituency

It is generally agreed among researchers that prosodic elements are organized into units which are structured in a hierarchy (Nespor and Vogel 1986, Beckman and Pierrehumbert 1986; see Shattuck-Hufnagel and Turk 1996 for a review). Furthermore, there is also good agreement that this prosodic hierarchy is heavily influenced by, but separate from, the syntactic hierarchy (e.g., Selkirk 1986, Hale and Selkirk 1987, Shattuck-Hufnagel and Turk 1996). Moreover, a number of researchers have found evidence in support of the constituency and prominence relations assumed within the Prosodic Hierarchy (Fougeron and Keating 1997; Turk and Sawusch 1999; Turk and White 1999; Cho 2002; Turk and Shattuck-Hufnagel 2000; Cole, Choi, Kim, and Hasegawa-Johnson, 2003; Cho and McQueen 2005).
How can these results be accommodated within theories of the prosodic hierarchy? At least three, and possibly more, layers of the prosodic hierarchy must be invoked to explain our results. The first layer of the hierarchy which is invoked is the Foot. Each of the crucial experimental stimuli in our study began with disyllabic words which were comprised of a single trochaic Foot (*channel, dizzy*). The second layer of the hierarchy which is invoked is the *Prosodic Word*. The crucial experimental stimuli each ended in a sequence of monosyllabic content words which could also form compound words (*foot, note, book, worm*). Each syllable in these sorts of compound words is itself assumed to be stressed and is therefore a foot. Moreover, it is agreed among researchers that the two syllables together comprise at least a single Prosodic Word, but some theorists have maintained that they form two Prosodic Words nested under a higher Prosodic Word (see Selkirk 1980, 1984, 1995; Nespor and Vogel 1986; Hayes 1989; Shattuck-Hufnagel and Turk 1996; and Turk and White 1999 for relevant discussion). The third layer of the hierarchy which is invoked is the *Intonational Phrase* (or possibly the *Intermediate Intonational Phrase*). This is due to the fact that this experiment used word lists together with intonation patterns which were consistent with and appropriate for such lists. It is widely held that lists are obligatorily comprised of Intonation Phrase-sized constituents (Schubiger 1958, Suci 1967, Shattuck-Hufnagel and Turk 1996, Beckman and Ayers-Elam 1997).  

What seems clear from our study is that the *structural status of the final syllables in our materials with respect to the Prosodic Hierarchy was perceived to change, depending on the prior prosodic context*. Crucially, the prosodic status of these materials changed in a manner consistent with the Parallelism principle, so that listeners heard
words with prosodic structures which were *parallel* to the preceding context. The role of the Parallelism principle in determining the perceived structure within the hierarchy can be understood using a metrical grid, as represented in (9). In this grid, we assume no organization of the final four monosyllables into Prosodic Words, due to their lexical ambiguity.

(9)

\[
x \quad x \quad x \quad x \quad x \quad x \\
x \quad x \quad x \quad x \quad x \quad x \quad x \\
\square \quad \square \quad \square \quad \square \quad \square \quad \square \\
\square \quad \square \quad \square \quad \square \quad \square \quad \square \quad \square \\
\text{[[channel]}_F]\text{pwd} \quad \text{[[dizzy]}_F]\text{pwd} \quad \text{foot}_F \quad \text{note}_F \quad \text{book}_F \quad \text{worm}_F
\]

We can first understand the role of Parallelism in determining the prosodic structure which listeners heard by considering each of the two prosodic contexts in our Pitch condition. For these two contexts, different autosegmental associations were made between pitch accents, phrasal tones and the first five grid positions, as shown in (10) and (11). Note that the intonational specification across the initial syllables served to group a H* and a L% in initial part of (10), but a L* and a H% in (11). The final acoustic-phonetic material in each case was the same. The addition of an accent boosts grid levels for each accent, following Beckman and Edwards (1994). Following standard assumptions regarding constraints on the nature of the prosodic hierarchy, specifically, Headedness and Layeredness (e.g., Selkirk 1995), foot becomes both an Intonational Phrase and a Prosodic Word in (10), but is neither in (11).

(10)

\[
x \quad x \quad x \quad x
\]
We have proposed that Parallelism mediates the perception of phonological structure through making events which are construed as similar have parallel metrical and grouping structure. We assume here that Parallelism affects meter and grouping (or constituency) through operations of beat deletion, silent beat insertion and/or defooting (Selkirk 1980, 1984; Hayes 1989, 1995; Giegerich 1985). The effects of Parallelism are illustrated below.

For the grid in (10), note is defooted to produce a single foot-level constituent notebook, while a beat is deleted from the column for book. Silent beat insertion operates to separate foot from note at the lowest grid level, as well as to generate parallel metrical structure for worm. Finally, the H L H sequence at the end of the string is reanalyzed with parallel metrical status and constituency at the levels of the pitch accent, Prosodic Word and Intonational Phrase. This produces the output shown in (12).
For the grid in (11), *foot* and *book* are first each respectively defooted to produce the single foot-level constituents *footnote* and *bookworm*. Moreover, a beat is deleted from each of *note* and *worm*. Next, a beat is added to *book* to give it comparable strength as parallel positions in earlier constituents. Finally, the H L H sequence at the end of the string is reanalyzed with parallel metrical status and constituency at the levels of the pitch accent, Prosodic Word and Intonational Phrase. This produces the maximally parallel output shown in (13).

\[\text{(12)}\]
\[
\begin{array}{cccccccc}
\times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times \\
\end{array}
\]

\[
[[\text{channel}]_{F}P_{wd}]_{IP}[[\text{dizzy}]_{F}P_{wd}]_{IP}[[\text{foot}]_{F}P_{wd}]_{IP}[[\text{notebook}]_{F}P_{wd}]_{IP}[[\text{worm}]_{F}P_{wd}]_{IP}
\]
\[
[H^{*}L^{-}]_{IP}[H^{*}L^{-}]_{IP}[H^{*}L^{*}]_{IP}[H^{*}L^{-}]_{IP}[H^{*}H^{*}]_{IP}
\]

\[\text{(13)}\]
\[
\begin{array}{cccccccc}
\times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times \\
\times & \times & \times & \times & \times & \times & \times & \times \\
\end{array}
\]

\[
[[\text{channel}]_{F}P_{wd}]_{IP}[[\text{dizzy}]_{F}P_{wd}]_{IP}[[\text{footnote}]_{F}P_{wd}]_{IP}[[\text{bookworm}]_{F}P_{wd}]_{IP}
\]
\[
[L^{*}H^{*}]_{IP}[L^{*}H^{*}]_{IP}[L^{*}H^{*}]_{IP}[L^{*}H^{*}]_{IP}
\]

Next, we briefly describe the analysis for stimuli in our Duration condition. For this condition, the F0 was flat and only the temporal interval between the fifth and sixth syllables was manipulated. In one context, this interval was lengthened (what we called
the ‘Weak-Strong’ context). In the contrasting context, this interval was shortened (what we called the ‘Strong-Weak’ context). We again propose that silent beat insertion, beat deletion, and defooting operate to generate parallel structures. For the ‘weak-strong’ expectancy condition, lengthening the interval generates a silent beat, so that the more isochronous rhythm sets up a parallel structure and that of (12) emerges, but without the tones. Similarly, for the ‘strong-weak’ expectancy context, the shortening of the interval had the effect of generating the grid structure in (13).

Our analysis as laid out in (9-13) makes at least two predictions. First, compound notebook as in (13) which reduces in our analysis to a single foot should have similar durational properties as a disyllabic foot like channel. This proposal finds phonetic support in studies of durational lengthening which have compared compounds like notebook with items like channel (Turk and White 1999, White 2002). Second, listeners should hear note and book as more closely linked in (12) than in (13). We have found support for this analysis in a followup word recognition study (Dilley and McAuley, in preparation). In a separate experiment, we exposed participants to all stimuli in this experiment in two phases: (a) a phoneme monitoring task, and (b) a subsequent word recognition task. Participants who heard notebook when it is presented in the context as in (12) more accurately responded that they had heard this item than when it was presented in a context like (13), even though all participants in fact did hear the sequence note followed by book. Our analysis captures this finding by proposing that participants who heard contexts as in (12) grouped note with book while participants who heard contexts as in (13) grouped note and book with different words rather than together and thus failed to recognize that they had in fact heard the word.
At first glance, the outcome of our experiment, as well as our phonological analysis above, might seem to be difficult to reconcile with the prosodic hierarchy. This is because much work within this framework assumes that prosodic boundaries are marked by local acoustic cues. (See e.g. Turk and Shattuck-Hufnagel 2000, White 2002 for reviews.) In our stimuli, whatever the local acoustic cues to word boundaries and/or other prosodic constituents present in the original stimuli, listeners readily reorganized the sequences so as to overcome those cues to generate new prosodic structures.

Rather than providing evidence against this hierarchy, we believe that our results provide new support in favor of the hierarchy. In particular, our results suggest that there is a perceptual basis for the hierarchy by demonstrating that global prosodic patterns can generate percepts associated with constituency (as well as prominence) relations. This suggests that global prosodic patterns can be an additional way in which prosodic structure manifests itself, in addition to local acoustic modifications such as final and accentual lengthening (Oller 1973, Cooper and Paccia-Cooper 1980, Wightman, Shattuck-Hufnagel, Ostendorf, and Price 1992, Cambier-Langeveld and Turk 1999). Moreover, we note that our results are neutral as to whether there is also an articulatory basis for the hierarchy, as has been claimed elsewhere in the literature (e.g., Beckman and Edwards 1990, 1994, Byrd 2000, Byrd and Saltzman 1998, 2003).

In addition, our results suggest a reason for why conflicting results have been found with respect to the acoustic marking of prosodic boundaries. A number of studies have found durational correlates of word and/or foot boundaries, but these effects have been shown to be inconsistent across studies and not clearly delimited by prosodic constituent boundaries. (See e.g., results in Beckman and Edwards 1990, Turk and White 1999, Turk
and Shattuck-Hufnagel 2000). Our findings suggest one reason why pinning down local acoustic correlates of constituency might prove to be elusive. That is, prosodic constituent boundaries may sometimes be marked through global prosodic characteristics, either in addition to, or perhaps in place of, local acoustic modifications or strengthenings.

Finally, we speculate that our results have implications for understanding how listeners segment words from the speech stream. The vast majority of research on speech segmentation has focused on the contributions of local acoustic cues to the isolation of words from continuous speech. For example, stressed syllables are likely to be heard as word-initial (e.g., Cutler and Norris 1988, Cutler and Butterfield 1992). Moreover, listeners use the degree of coarticulation at word boundaries as a cue to word segmentation (Mattys 2004; Johnson and Jusczyk 2001; Mattys, White and Melhorn 2005). Recent studies have suggested that listeners use local repeated pitch patterns to segment speech (Welby 2003, S. Kim 2004). To our knowledge, the present study is the first to show an effect of prior context on adult word segmentation. While the kinds of contextual demonstrated here seem unlikely to play a dominant role in word segmentation by adults, it seems possible that such biases may be more important for word segmentation by infants, for two reasons. First, the kinds of biases which seem to have been demonstrated in the present study are thought to be innate; thus, it is conceivable that such biases could serve as useful tools which would aid infants in the segmentation process and the gradual development of a lexicon. Second, repetitive pitch patterns are common in infant-directed speech (e.g., Fernald and Mazzie 1991),
suggesting that biases for developing expectations about structure based on innate grouping principles could be a useful segmentation strategy.

5. Summary

The motivation for this paper was to provide an explanation for widespread usage of repeated, parallel pitch patterns across languages. The occurrence of such patterns has not previously received an account within AM theory. We proposed an explanation in terms of general processes of auditory perception. In particular, we proposed that a Parallelism principle governs perception of sequences of parallel accents and timing patterns. We found strong empirical support for this proposal: listeners’ judgments about the identity of the final word in lexically ambiguous syllable strings were strongly influenced by the preceding prosodic context, even though the acoustics of the final syllables were held fixed. This upholds a critical assumption of Pierrehumbert’s original (1980) proposal that the inventory of nuclear and prenuclear accents is one in the same. Moreover, these results have further theoretical significance by suggesting that the Prosodic Hierarchy might have a phonetic basis in perception, not just production. In particular, our results suggest the influence of global prosodic structure on the perception of prosodic boundaries and accents. Finally, our results potentially provide grounds for understanding potentially conflicting findings within work on phonetics of the Prosodic Hierarchy by suggesting that in some cases, boundaries and accents may not be realized acoustically in terms of local acoustic characteristics, but rather in terms of global ones. These proposals open up a number of new lines of inquiry into the nature of the phonetics-phonology interface as regards prosodic structure.
6. References


Journal of Experimental Psychology: Human Perception and Performance, 29(6), 1102-1125.


Appendix A

1. channel dizzy foot note book worm
2. chocolate lyric down town ship wreck
3. worthy vinyl life long hand shake
4. victim fragile sun spot light house
5. skirmish princess side kick stand still
6. sonar baggy air field work day
7. wallet ruthless back space suit case
8. prelude charcoal touch down play boy
9. rumor habit dream boat yard stick
10. texture nozzle coat tail gate post
11. swivel witness thread bare foot bridge
12. shower cannon pass word play pen
13. genius planet work horse whip lash
14. splendor radish friend ship board room
15. climate humble wise crack pot hole
16. brandy curtain ear drum head piece
17. border taxi life boat house top
18. pennies handy half back fire wood
19. quaker trophy step son/(Sun) day break
20. cherry vapor car pool side walk
I wanted to read it to Julia.

Figure 1. Repeated pitch patterns as described in Ladd (1986).

channel  dizzy  foot  note  book  worm  ‘HL expectancy’

channel  dizzy  foot  note  book  worm  ‘LH expectancy’

channel  dizzy  foot  note  book  worm  ‘weak-strong expectancy’

channel  dizzy  foot  note  book  worm  ‘strong-weak expectancy’

Figure 2: Example of stimulus used in the Duration condition.
Figure 3: Mean proportion of disyllabic final word responses with standard error bars for disyllabic and monosyllabic contexts for the three prosodic conditions.
Figure 4: Mean perceptual sensitivity, $d'$, with standard error bars for the Duration, Pitch, and Pitch+Duration conditions.
Figure 5: Mean response criterion, $c$, with standard error bars for the Duration, Pitch, and Pitch+Duration conditions.
For a small subset of stimuli in the Duration and Pitch+Duration conditions, the consonantal onset of the third syllable from the end was shortened. However, the vocalic portions of the last three syllables, as well as the entirety of the final two syllables, were never modified across any conditions.

An alternative explanation of participants’ behavior in the Duration condition was that lengthening the interval between syllables 5 and 6 in the ‘weak-strong’ expectancy context introduced a major prosodic boundary between these syllables. According to this explanation, the boundary caused listeners to interpret the subsequent syllables as being organized into a disyllabic word followed by a monosyllabic word. We agree that listeners may have heard a prosodic boundary between the 5th and 6th syllables as a result of acoustic modifications. However, this cannot fully explain the responses of participants. In particular, we believe participants could just as well have reported a monosyllabic word under these conditions, particularly since compounds which differ only in the location of a word boundary have been shown to be realized on occasion with similar durational characteristics.

It may be that given sufficient material between accents, non-parallel accentual sequences can occur. The Parallelism principle seems to operate over a relatively short time span.

The term Intonation Phrase is recognized as corresponding to what in earlier literature had been referred to as either a tone group or a breath group.

We have chosen to adopt the simpler notations L% and H%, following Gussenhoven (2004), in lieu of the more complex L-L% and H-H% notation often used within the AM framework, primarily to aid readability.

We see no need to assume that grid columns are boosted to a height 4 in list intonation contexts, in contrast with Beckman and Edwards (1994).

We are grateful to Heinz Giegerich for suggesting an analysis in terms of defooting.

The final syllable is necessarily heard as the location of an Intonation Phrase. Since this syllable ends in a high tone, we infer that the boundary tone is a H%.