What do listeners attend to in hearing prosodic structures? 
Investigating the human speech-parser using short-term recall

Annie C. Gilbert, Victor J. Boucher

Laboratoire de Sciences Phonétiques, Département de linguistique et traduction
Université de Montréal, Montréal, Canada
annie.gilbert@umontreal.ca, http://www.ling.umontreal.ca/phonetique

Abstract
This study examines how heard prosodic patterns are parsed by reference to a principle of focus of attention [1]. According to this principle, attention holds up to four items at once, and the same upper limit appears to apply to the number of syllables in rhythm groups [2]. On this basis it was predicted that in recalling heard prosodic structures, listeners would attend primarily to rhythm groups. 31 Ss were asked to recall the prosody of heard series of [pa] bearing various intonation groups and repetitive or varying rhythms. Exp. 1 showed how the focus of attention can shift when rhythm patterns are repetitive. However, Exp. 2 showed that listeners focus on rhythm when patterns vary (as in speech). The results bear implications on explaining the role of prosodic groups in speech.

Index Terms: speech prosody, parsing, short-term memory, attention.

1. Introduction

Generally, “prosodic structure” above the syllable designates two classes of marks. One class involves lexical marks, such as word stress, and another class involves non-lexical marks bearing on rhythm and intonation groups (RG and IGs; [3]). The fact that these latter groups often fail to match assumed word and phrase units stands as a fundamental paradox of linguistic theory and is at the center of bootstrapping issues [4]. One pivotal aspect of this mismatch bears on well known “size effects”. For example, IGs and RGs can extend across the subject and verb parts of utterances when constituents involve small numbers of syllables. The nature of such effects remains unexplained but point to frameworks that are independent of syntax. Consider the following sentences used by Juscyck [4] to illustrate size effect on prosodic groups (indicated by “/”):

(1) He ate/the cake
(2) Thomas/ate the cake
(3) The neighbor’s boy/ate the cake

There is likely to be fluctuation in the way speakers group (2): “Thomas ate/the cake” might also occur. But where the grouping becomes highly predictable is when one lengthens constituents as in (3)

A prosodic division after “ate” in this case would be unusual. Thus, in increasing the number of syllables of constituents, prosodic grouping becomes predictable, and this is basically what we found in an earlier study of French IGs [5]. We used French because, in a language like English, lexical stress can confound the observation of groups [6]. For example, in the above sentence (1), lexical stress overlaps temporal and tonal changes marking the ends of RGs and IGs. The absence of lexical stress facilitates the observation of prosodic grouping in different speech behaviors, and this is crucial to defining the nature and role of prosodic grouping in speech.

In fact, the literature on prosody fails to note that prosodic structure not only occurs in meaningful utterances. It also arises spontaneously when people recall lists like phone numbers or series of nonsense syllables. Such common behaviors show rather clearly that prosodic grouping is not dependent upon syntax, and bears functional links to processes of short-term memory (STM).

Yet, the idea that prosodic structure may bear on memory processes has not been the subject of much research. In psychology, it is known that the presentation of lists of digits or nonsense syllables in groups facilitates memorization, and that groupings of 3 or 4 have optimal effects on recall. However, memory research on “grouping effects” typically involves the presentation of series containing repetitive three-item groups usually separated by pauses. This offers few parallels with prosodic groups in speech that can vary in size and that involve simultaneous timing and tonal changes. For this reason, an earlier study [2] using French examined how grouping effects on recall conformed to prosodic groups in speech.

Experiment 1 of that study involved the recall series of nonsense syllables presented either with RGs of varying size or without any rhythm. The results showed that serial recall was enhanced by RGs of up to four syllables, but with longer groups, recall was significantly lower than baseline (series with no rhythm). Experiment 2 examined how the limit on grouping effects conformed to RGs in French utterances. Sentences containing noun-phrases of varying length were presented and subjects had to say the sentences and then repeat their utterances while pacing syllables with glottal stops (an echoing procedure). The finding was that, when noun-phrases exceeded four syllables, an internal lengthening was produced to create shorter RGs. In sum, size effects on RGs in utterances matched the size of groups that benefited recall.

Such observations present an intriguing correspondence with a principle of STM defended by Cowan [1]. Briefly, Cowan’s main claim is that people cannot attend to multiple stimulus channels at once, so there is a focus of attention that can hold up to four items at once. This essentially constitutes a parsing principle. As an illustration, Cowan [1] mentions the case of phone numbers where groupings of no more than three or four digits indicates how many items can be kept in the focus of attention at once, and it is speculated that two or three groups could be held active (perhaps in sensory memory) so as to form “chunks” in long-term memory. Though no relation is drawn with respect to prosody (and notions of “items” and “chunks” remain controversial), there
is a striking parallel with respect to limit patterns on RGs and IGs in speech [2, 5].

By reference to the principle of focus of attention, the question arises of what listeners attend to in recalling structural prosody. In the tests by Boucher [2], the recall task required that subjects recall syllables. One could explain the findings by suggesting that, as long as RGs matched the focus of attention, recall was enhanced. But when the groups exceeded four syllables, representing unusual patterns in speech, attention swayed and recall performance fell significantly. However, if attention is not directed at recalling syllables, say in the case where listeners are asked to recall series of [pa] syllables with IGs and RGs, would listeners focus on rhythm or intonation? The present study examined this question with the purpose of evaluating the concept of focus of attention as a speech-parsing principle. Two experiments are used, involving complementary predictions:

(1) (Exp. 1 and 2) If in hearing series with prosodic patterns listeners attend primarily to RGs, then recall of RGs should not differ significantly when presented with or without intonation.

(2) Recalling repetitive RGs (Exp. 1) is fairly easy and would thus require less attention than the recall of varying RGs (Exp. 2). In the repetitive-RGs condition, non-matching IGs would create a distraction and shift the focus of attention away from RGs, affecting their recall. By contrast, for varying (non-repetitive) RGs, listeners would naturally pay more attention to rhythm, so their focus of attention would not be influenced by non-matching IGs and recall would not be affected.

2. Experiment 1: repetitive RGs

2.1. Method

2.1.1. Participants

The subjects were 31 native speakers of French aged from 19 to 36 years of age (mean of 24.5 yrs). These individuals had no history of speech or hearing disorders and no formal training in music or dance.

2.1.2. Stimuli

These were sequences of 9 and 12 [pa] syllables containing lengthening marks to create three equivalent RGs of either 3 or 4 syllables. The sequences also contained F0 resets marking IGs (see Table 1). The series were elaborated in two steps.

Table 1: Sets of stimuli used in Exp. 1 (arrow = IG)

<table>
<thead>
<tr>
<th>test conditions</th>
<th>stimuli</th>
</tr>
</thead>
<tbody>
<tr>
<td>Matching RG &amp; IG</td>
<td>RG1</td>
</tr>
<tr>
<td></td>
<td>RG2</td>
</tr>
<tr>
<td></td>
<td>RG3</td>
</tr>
<tr>
<td>Non-matching RG and IG</td>
<td>RG1</td>
</tr>
<tr>
<td>1</td>
<td>RG2</td>
</tr>
<tr>
<td></td>
<td>RG3</td>
</tr>
<tr>
<td>Non-matching RG and IG</td>
<td>RG1</td>
</tr>
<tr>
<td>2</td>
<td>RG2</td>
</tr>
<tr>
<td></td>
<td>RG3</td>
</tr>
<tr>
<td>Whole-utterance IG</td>
<td>RG1</td>
</tr>
<tr>
<td></td>
<td>RG2</td>
</tr>
<tr>
<td></td>
<td>RG3</td>
</tr>
<tr>
<td>Control</td>
<td>RG1</td>
</tr>
<tr>
<td></td>
<td>RG2</td>
</tr>
<tr>
<td></td>
<td>RG3</td>
</tr>
</tbody>
</table>

First, a pre-recorded syllable [pa] spoken by a native speaker of French (44.1 kHz sampling rate) was duplicated and assembled in series. To ensure the perception of rhythm in otherwise isochronous syllables, software functions (CSL model 4400, Kay Elemetrics) were used to adjust durations by reference to P-center ratios [7]. With these adjustments, intervals between the P-centers were equivalent except for long syllables marking the boundaries of RGs. In this latter case, the P-center interval between a “stressed” and a following “non-stressed” syllable presented a ratio of 1.7, reflecting the typical ratio of stress/non-stress syllables in French speech [8, 9]. With these timing adjustments, stimuli durations were 3 seconds for 9-syllable sequences and 3.84 seconds for 12-syllable sequences.

In a second step, re-synthesis routines (4SL model 5104, Kay Elemetrics) were used to create F0 contours reflecting four experimental conditions: where contours matched the separate RGs (“matching” RGs condition), overlapped two RGs (“non-matching” RGs condition), overlapped all three RGs (“who-utterance” condition) or was flat (“control” condition) and the latter patterns were balanced as indicated in Table 1. The perceptual boundary of IGs was created by adjusting the reset of the F0 contour at the start of an IG by 11 semitones.

Following the above adjustments, all series were downsampled to 22.05kHz. These operations gave series devoid of clicks with perceptible rhythms and tonal contours, that were perceived as “natural sounding” in pre-test listening by non-participants.

2.1.3. Procedure

All series were preceded by a warning tone 1.5 second before the onset of the test sequence. Each type of sequence was presented twice in random order along with other series. In the test, the subjects sat in a noise-attenuating booth with a head-mounted omni-directional microphone placed near their lips (AKG C477 WRL). The stimuli were delivered via a loudspeaker at an intensity of 74 dBa at the subjects’ ears. Subjects were simply asked to recall as accurately as possible what they would hear, and there were no further instructions on what the point of the task was. This particular procedure was intended to observe subjects fluctuating attention with respect to varying prosodic groups. The subjects’ oral recall was recorded using a 16-bit external sound card set at a 44.1 kHz sampling rate (CSL model 4400, Kay Elemetrics). The use of an omni-directional microphone allowed the recording of both the stimuli and oral responses during the test.

2.1.4. Analysis of responses

Using the aforementioned analysis system, RGs in recorded responses were measured by reference to lengthened syllables in displayed waveforms. “Correct recall” was scored in terms of the production of a lengthened syllable corresponding to the location of RGs in the presented series.

As for the recall of IGs, the analysis used the peak-detection routines of the above CSL system. To ensure reliable measures, three pitch analyses were overlaid reflecting different settings of frame lengths and frame advances (these were length 20 ms/20 ms advance, length 25ms/20 ms advance, and length 25ms/25 ms advance). The accuracy of the recall was based on the presence and location of intonation resets, regardless of its range (Hz). To qualify as a successful recall, the number of resets in a response had to match the number of resets in a presented stimulus, regardless
of whether or not the subjects produced the right number of syllables. Figure 1 presents an example of a submitted stimulus along with a subject’s response (in this case, there was inaccurate recall of RGs and IGs).

![Figure 1](image1.png)

**Figure 1: Example of a stimulus and S’s response.** (frequency axis is logarithmic)

### 2.2. Results

As expected, subjects were highly successful in recalling repetitive RGs, as can be seen in Fig.2. Since assumptions of normality could not be met, non-parametric Friedman tests were applied to verify the effects of tonal conditions. The test statistics showed that the conditions did create some significant difference (for 9-syll. series: $\chi^2=18.529$, $n=31$, $p<0.001$; for 12-syllable series: $\chi^2=29.581$, $n=31$, $p<0.001$). Further analysis revealed that these differences were no longer significant when the results on the non-matching condition were removed ($\chi^2=2.8$, $n=31$, $p=0.247$; $\chi^2=3.263$, $n=31$, $p=0.196$, for 9- and 12-syll. series respectively). Thus, it appears that recall of repetitive RGs is not significantly different when series are presented without intonation (the monotone “control” condition) or with “matching” and “whole-utterance” contours. Differences in recall of RGs only occur with non-matching IGs.

![Figure 2](image2.png)

**Figure 2: Average recall of symmetric RGs.**

As for IGs, Fig. 3 shows that subjects were slightly less successful in recalling these groups compared to repetitive RGs, though Wilcoxon tests showed that recall of RGs was significantly superior to the recall of IGs only in the non-matching condition (for 9-syll. series: $Z=2.986$, $n=31$, $p=0.003$; for 12-syll. series: $Z=2.399$, $n=31$, $p=0.016$). This latter statistic may simply suggest that, even though on-matching IGs may have influenced subjects’ attention, recall of repetitive patterns is easier than the recall of varying IGs.

![Figure 3](image3.png)

**Figure 3: Average recall of IGs over symmetric RGs.**

### 2.3. Discussion

In terms of our predictions, recall of repetitive RGs is not affected by the presence of whole-utterance F0 contours or by the absence of contours. In fact, the recall of RGs is only affected when IGs overlap only two of the three RGs. The recall of repetitive groups represents a relatively easy task (see the scores of Fig. 2) that did not require much attention. In these conditions, listeners’ focus could be influenced by the more demanding task of recalling varying IGs. However, in speech, RGs vary and with such patterns it was the prediction that listeners’ attention would not be influenced by non-matching IGs.

### 3. Experiment 2: varying RGs

#### 3.1. Method

##### 3.1.1. Subjects and procedure

The subjects of the first experiment also participated in the second experiment. The stimuli were randomly ordered and presented with those that served in Exp. 1.

##### 3.1.2. Stimuli and analysis of responses

The sets of stimuli of Exp. 2 reflected the intonation conditions of Table 1 except that RGs 1, 2, 3 contained 2 to 5 syllables. These latter groups of varying length were balanced in series of 9 and 12 syllables as summarized in Table 2. All the stimuli were elaborated following the same procedures as those of Exp. 1 and response analysis was unchanged.

<table>
<thead>
<tr>
<th>length of series (syll.)</th>
<th>no. of syllables</th>
</tr>
</thead>
<tbody>
<tr>
<td>RG1</td>
<td>2</td>
</tr>
<tr>
<td>RG2</td>
<td>4</td>
</tr>
<tr>
<td>RG3</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>12</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
</tr>
</tbody>
</table>

### Table 2: Combinations of varying RGs used.
3.2. Results

Compared to Exp. 1, subjects were less successful in recalling varying RGs across conditions (see Fig. 4).

![Average recall of varying RGs.](image)

Figure 4: Average recall of varying RGs.

Also, there are clear effects of length suggesting a greater degree of difficulty in the recall of RGs of different size. In this context, Friedman tests showed that the intonation conditions had no significant effect on the recall of RGs (for 9-syll series: $X^2=6.1$, $n=31$, $p=0.107$; for 12-syll series: $X^2=1.352$, $n=31$, $p=0.717$). Thus, there is a central difference between these results and those of Exp. 1. In the present test, recall of varying RGs was a more difficult task, which required a greater degree of attention. The finding of non-significant differences, regardless of the intonation condition, suggests that subjects’ attention was not influenced by varying intonation patterns and remained focused on heard rhythms.

Considering the recall of IGs, Fig 5 shows that subjects’ recall of these groups follows a pattern quite similar to that of intonation recall of Exp.1 (Fig. 3). In fact, Wilcoxon tests showed that recalling IGs in the non-matching condition is significantly harder than in the matching condition (for 9-syll. series: $Z=-2.581$, $n=31$, $p=0.01$; for 12-syll. Series: $Z=-2.237$, $n=31$, $p=0.025$) or whole-utterance condition (for 9-syll. series: $Z=-1.580$, $n=31$, $p=0.114$; for 12-syll. series: $Z=-2.597$, $n=31$, $p=0.009$).

![Average recall of IGs over varying RGs.](image)

Figure 5: Average recall of IGs over varying RGs.

In short, when IGs contained varying numbers of RGs (non-matching condition) recall of IGs was significantly affected. Such effects would not have occurred if listeners had focused on intonation rather than rhythm.

4. General discussion and conclusion

Exp. 1 involved series with repetitive RGs, which were easily recalled, except when intonation presented different patterns. Such results can be explained by considering that repetitive patterns do not require a focusing of attention and, in these conditions, changing intonation may draw listeners’ attention away for rhythm and thus affect their recall of rhythm. In expecting these effects on attention, Exp. 2 used varying RGs, -- which more nearly resembles the situation of speech. In this case, the results show that listeners’ attention focused on rhythm and did not fluctuate with varying intonation patterns. Thus, on the question of what do listeners attend to when they hear prosodic patterns, the observations suggest that they attend primarily to rhythm. It should be added that recall of IGs remained high across the experiments and represented a relatively easy task (judging from the recall scores). This is not incompatible with the parsing principle of the focus of attention. Sensory memory [1] may well apply to patterns of intonation, but the recall of heard series of syllables or different timing patterns likely requires that listeners focus their attention on given extents of speech. If further evidence confirms that attention is limited to a given number of syllables that match the size of rhythm groups, as we suspect, then there would be a working rationale for the role of prosodic structures in speech.

5. Acknowledgements

This research was supported by an SSHRC Master’s Scholarship awarded to the first author.

6. References