Acoustic Correlates of Listener-Identified Boundaries in Spontaneous French Speech

Caroline L. Smith
University of New Mexico

1. Introduction

Prosodic structure provides an organization to spoken (and signed) language that signals the grouping of words into phrases, and the relative prominence of words and syllables (e.g., Shattuck-Hufnagel and Turk 1996). This organization is cued by articulatory and acoustic modifications to the signal. Different languages modulate different dimensions of the signal in different ways, but prosodic variation often includes durational lengthening or shortening, addition or removal of pauses or breaks in the flow of speech, changes in F0 patterns, expansion or shrinkage of gestural magnitude. These modulations facilitate the task of listeners by communicating the structure of the speech. Spontaneous speech, in particular, very often does not conform to conventional syntactic patterns, so listeners cannot rely on their knowledge of syntax to interpret what they hear.

Linguists have posited a hierarchy (Nespor and Vogel 2007) of prosodic units of different sizes, but it is not clear to what extent these different units are relevant for ordinary listening, nor do we know much about how listeners' perception of prosody differs among different languages. In order to investigate the place of prosodic units in listeners' grasp of a spoken message, studies have been done attempting to access the perceptions of naive listeners. These may well be different from the perceptions of linguists who bring theoretical expectations to bear when they analyze the structure of an utterance. This study takes listener perceptions as the basis for phrasal structure, and seeks to identify the acoustic properties of the speech signal that contribute to their perceptions. The first step was thus to obtain data on perceptions of prosodic structure, specifically the locations of boundaries between groups of words. The second step was to examine the speech signal immediately preceding the locations that were perceived as boundaries by a consensus of the listeners, since it is assumed that a listener needs to be able to detect a boundary from modulations of the signal before the boundary, if they are to interpret the speech stream successfully.

1.1. Phrasing in French

The structure of phrases in French has received considerable attention in the literature. Most authors agree that there are two levels of prosodic structure intermediate between the word and the utterance. While these have been named in diverse ways, I will refer to the smaller as an 'accent group' and the larger as an 'Intonational Phrase.' (For a helpful diagram comparing terminology used by different authors see Di Cristo 2005:152.) There seems to be general agreement that the last full (non-schwa) syllable of an accent group is more prominent than other syllables, and that this prominence is closely associated with the presence of a boundary (e.g., Di Cristo 2000, Mertens 2006). Post (2000) uses the term Phonological Phrase for what I am calling an accent group. She proposes a definition that is closer to how Nespor and Vogel (2007) define a Phonological Phrase, referring to

* Thanks to audiences at the LARP 5 conference, and at the Interface Discours-Prosodie (September 2009, Paris) and the Journée d'Etudes Linguistiques (June 2009, Nantes) where portions of this material were previously presented. A special thanks to all the speakers and listeners who participated in my study, to colleagues at the Université Lyon 2, Université Paris 3, and the ENS-LSH in Lyon who facilitated the work, and to Frédérique Bénard for editorial assistance with the transcriptions.

lexical categories. This differs from many other definitions that use more performance-based criteria, that is, the phrasal boundaries depend on how a sequence is produced in a specific production. The accent group or Phonological Phrase is normally no larger than a single lexical word and preceding clitics. Although different researchers use different criteria to define it, French speakers find these groups relatively straightforward to demarcate, as the present study will show.

The larger Intonational Phrase (IP) is more difficult to define. DiCristo (e.g., 1998, 2000) and Di Cristo and Hirst (e.g., 1997) treat the IP (their "Unité Intonative") as a unit for the assignment of tonal features, but describe it as also subject to syntactic factors and "contraintes sémantico-pragmatiques" (semantic-pragmatic constraints). D'Imperio et al. (2007:2) comment that IPs have been defined in "a rather fuzzy way as a unit showing … 'melodic cohesion'." This references the unit's physical (melodic) characteristics as the primary defining property, although these authors go on to mention the importance of semantic and pragmatic influences in delimiting IPs. Perhaps the clearest example of a definition of the IP comes from D'Imperio et al.'s (2007) finding of a difference between IP-final and non-final contours, a result distinct from the predictions of previous models. Their results suggest that in French, as in English and many other languages, the larger IP unit does affect production in dimensions that are perceptible and measurable. The present study seeks to identify some of the measurable properties that might be signaling IP or other phrasal boundaries to listeners.

1.2. Previous studies of naive listeners' perception of prosody

Studies in several languages have tested untrained listeners to elicit their perception of one or both of the key dimensions of prosodic structure, the prominence of words or syllables, or the locations of boundaries demarcating groups of words. Because these studies involve listeners without formal linguistic training, they do not define in linguistic terms the group of words that they are asking listeners to identify; thus, the nature of a "group of words" may differ considerably among the different studies, even beyond any differences among the languages that were investigated.

Dutch listeners with no special training (Streefkerk et al. 1997) were tested on their perceptions of the prominence patterns and phrasal groupings of a set of sentences read aloud, which had been constructed so the set contained all the phonemes of Dutch at least once. This kind of controlled sentences is likely to be less varied prosodically than spontaneous speech. Untrained listeners were shown to use linguistic knowledge in perceiving prosodic boundaries by Mettouchi et al. (2007). They asked both native speakers and non-speakers of Kabyle and Hebrew to mark boundaries in samples of speech (the native speakers only worked with their own language). Native speakers listened to speech that had been filtered to render segmental information unintelligible, so that they had to respond on the basis of prosody alone. Even so, the native speakers' responses were closer to an expert transcription than were the non-speakers' labeling, which was presumably also based on gross prosodic patterning.

A recent study undertaken in English (Cole et al. 2010 a, b) included a larger number of listeners, 97 in four groups. Cole et al. asked untrained listeners, native speakers of American English, to mark prominent words, or boundaries between groups of words, while listening to a sample of spontaneous, conversational speech. As they listened, they followed along on a printed, orthographic but unpunctuated transcription. Half the listeners marked prominence first for one set of materials, then boundaries on a different set, while the other half performed the tasks in the reverse order. The listeners indicated their responses by underlining a word they perceived as prominent, or by marking a slash between two words where they perceived a boundary between two "chunks" of speech. Cole et al. obtained high rates of agreement among their listeners, particularly for the marking of boundaries.

Studies with somewhat similar methodology have also been conducted in French. Both Pagel et al. (1995) and Obin et al. (2008) asked naive listeners to label accents and/or boundaries in recorded speech passages, but they do not report the responses of their listeners in any detail, as in both studies, the primary interest was in developing automated methods for prosodic labeling. A study by Portes (2000) compared prosodic labeling by 12 naive listeners and 5 experts. Her naive listeners were first given a brief explanation of their task but no feedback or training. They labeled boundaries, accented syllables and emphasized words or expressions. The criterion for identification of a boundary, accent or emphasis was that it was marked by a majority of the listeners. Portes found that the syllable preceding a boundary was marked as accented at 84% of the identified boundaries, supporting the view

that the end of an accent group is the most common location for a boundary. Portes notes that it is impossible to claim that the boundary locations identified by listeners correspond to a specific linguistic unit: the labeled boundaries demarcate chunks that vary greatly in length and syntactic content. She notes this particularly for the non-terminal (comma) boundaries identified by the expert labelers, but the concern applies to those labeled by the naive listeners as well, and to a lesser extent, to the boundaries identified as terminal (in punctuation, corresponding to a period).

Portes's study raises many interesting issues, but is somewhat limited in that only one speech sample was analyzed and relatively few listeners participated. The study reported here uses a methodology similar to that of Portes and Cole et al., but by comparing two types of speech materials, aims to uncover additional factors contributing to listeners' perceptions of the structure of spoken French. It also includes more listeners than Portes's study.

2. Method

2.1. Speech materials for the listening experiment

Two types of speech materials were used. One set of ten extracts was prepared from recordings of a map task experiment that had been previously recorded at a Paris university (Smith 2007). The speakers were ten female undergraduates from the Paris region. They were recorded individually in a task which required them to give directions over the telephone for using the Paris métro system to travel to different places around the city. They were speaking to an interlocutor who they could not see, but who asked questions and provoked discussion. These extracts consist of fairly informal, spontaneous task-directed speech. They include many disfluencies, repetitions, hesitations and other characteristics of spontaneous speech. The extracts were selected from portions of the conversations during which the one speaker had a relatively long conversational turn, without overlap by the interlocutor. These extracts varied from 13 to 24 seconds in length.

The second set of ten extracts was taken from a discussion/debate that was broadcast in December 2008 on a current affairs program on the France Info radio station, the subject of which was television advertising. These extracts also consist of single-speaker passages of spontaneous conversational speech, but the speakers are journalists and public figures. Their conversation was recorded for broadcast and thus illustrates a more formal register, although it also exhibits typical properties of spontaneous speech, such as repetitions and filled pauses. The selected samples include two extracts from each of the five speakers who participated in the discussion. The duration of these extracts is from 26 to 53 seconds.

Orthographic transcriptions of the extracts were prepared by the experimenter (a fluent non-native speaker of French), then edited by a native speaker with phonetic training. These transcriptions were prepared for use in the listening test by removing punctuation and line breaks except as necessary to fit on the page, in order to avoid providing any hints as to the structure. Disfluencies such as repeated or partial words were included in the transcripts but filled pauses ("euh…") were not indicated. Three additional extracts (two from map task conversations, one from a radio broadcast similar to the one used for testing) were also prepared to serve as practice samples.

2.2. Participants and testing procedure

Fifty-one listeners without advanced training in phonetics or prosody were recruited at higher education institutions in France. Most were undergraduate students in linguistics. In order to test listeners in groups for efficiency reasons, they were not screened for native language, and thus, a few were included who are non-native speakers of French. This was considered unlikely to be a problem because: (i) All participants are sufficiently fluent to participate in higher education (and informal conversation with the experimenter suggests all are very comfortable in French); and (ii) No more than 10% of the listeners tested are non-native users of French, so their influence on the overall results will be minimal. Different listeners participated in the experiment in different settings at the Université Paris 3, Université Lyon 2 and the École Normale Supérieure – Lettres et Sciences Humaines in Lyon: some were tested in groups of 5-17 in a classroom, others individually or in groups of two or three in a

sound-attenuated room. Since the goal of this study was to investigate how listeners perceive prosody under ordinary listening conditions, variation in the settings in which the listeners participated was considered to add to the validity of the study. Each listener was presented with a packet containing instructions and the printed transcriptions of the practice and test extracts. They marked all their responses on these print-outs.

A total of 26 listeners were asked to mark a vertical line between words at locations where they perceived a boundary between two phrasal units (*syntagmes*, defined as groups of words that form a single unit for both meaning and function).[1] All listeners heard the extracts in the same order, with brief pauses between extracts, controlled by the experimenter depending on the listener(s)' wishes. They practiced their task first on the two practice map task extracts, then responded to ten map task extracts, after which they practiced with the extra broadcast extract, and responded to the ten test extracts from the radio broadcast. The ten extracts from the map task were presented in the (random) order in which the speakers had been recorded. The radio broadcast extracts were presented in random order, not in the order in which they occurred originally in the program; no two extracts with the same speaker were presented consecutively.

One listener failed to follow directions, so that individual's responses were excluded from analysis, leaving a total of 25 listeners. All of their responses to the twenty test extracts were retained, and coded in Excel spreadsheets.

2.3. Statistical analyses

Most of the results reported here are counts, proportions and t-tests that were calculated using Excel. Agreement among listeners was assessed using a modified form of Cohen's Kappa. Kappa is a statistic that assesses agreement, taking into account the amount of agreement that can be expected by chance. Kappa values can vary between 0 and 1. The particular form of Kappa used here is based on Brennan and Predinger (1981); it is suitable for tasks with multiple raters in which the raters are not constrained as to how many items they assign to each category ("free marginals"). Calculations were made using the Online Kappa Calculator (Randolph 2008). Kappa values were determined for each extract, pooling across all the listeners.

A second type of analysis involved the calculation of a boundary score for each word, equal to the proportion of listeners who marked that word as being followed by a boundary. Those words marked by two-thirds or more of listeners (17 or more out of the 25) were considered to have "consensus" agreement. This criterion was arbitrary but indicates a substantial consensus.

3. Results

3.1. Agreement among listeners

The kappa statistic was used to assess agreement among listeners, as described above. Kappa values for each extract ranged from .75 to .88 with a mean of .83. Randolph (2008) suggests that for this form of kappa, .7 or above is "adequate", so the values obtained in this study are well above this proposed cut-off. Perhaps the most striking examples of agreement among listeners are the thirteen locations where every listener in the boundary-marking group marked a boundary at the same place. Three of these locations occurred in three different map task extracts, and ten in four of the broadcast extracts. These results suggest that there are some locations where speakers are extremely successful at communicating the presence of a boundary.

[1] The drawback to using the term *syntagme* in the instructions to the listeners is that it can be used like the English word "sentence", implying a syntactic unit. However, it also means "phrase" in both the syntactic and prosodic senses (*syntagme nominal, syntagme intonatif),* and there is no convenient term that distinguishes these concepts. (Thanks to Christopher Stewart for commenting on this.)

3.2. Distribution of boundaries in the extracts

As described in section 2.3, "consensus" markings were identified as those locations where at least 67% of the participants had marked a boundary. The number of locations identified in this way is given in Table 1, together with data on the number of words in the extracts, and the number of words between locations that were consensus boundaries (averaged for each extract).

	Mean	Minimum	Maximum
Map task extracts			
Number of words in extract	55.9	23	92
Number of consensus boundary locations	3.5	2	5
Number of words between consensus boundaries	10.7	6.9	13.8
Broadcast extracts			
Number of words in extract	134.8	87	206
Number of consensus boundary locations	8.5	5	12
Number of words between consensus boundaries	12.3	10.9	14.7

Table 1. Number of words and number of consensus boundary locations in two sets of extracts.

The intervals between locations marked as boundaries might be expected to delimit chunks of speech that correspond to a phrase. Investigating the size of these chunks should shed light on the type of unit(s) that listeners are perceiving. The median interval between boundary markings was 9.7 words, and the mean 11.5 words, across all 20 extracts and 25 listeners. The range of variation was greater among different listeners (averaging across all extracts) than among different extracts (averaging across all listeners). The range of variation among the different listeners (averaging across all extracts) was from 5.4 to 25.1 for the map tasks, and from 5.4 to 27.4 for the broadcast ones. Although the overall ranges (averaged across listeners) were fairly similar for the two sets of extracts, as can be seen in Table 1, the two sets differed in that listeners tended to mark boundaries more often when listening to the map tasks than they did for the broadcast extracts. This difference suggests that listeners perceived slightly longer chunks in the speech of the journalists, which is plausible given the more formal speech style and more complex syntactic structures that they employed. The number of consensus boundaries within an extract varied from two to twelve. This closely correlates with the number of words in the extract ($r = 0.9$), suggesting that listeners' strategies did not differ greatly across different extracts.

The size of the intervals between boundaries marked by listeners is closer in size to IPs than to accent groups. They also seem to be smaller in scope than an Utterance, as defined by Nespor and Vogel 2007. As in Portes's (2000) study, the nature of spontaneous speech means that the stretches of speech between consensus boundaries seldom constituted a grammatical constituent. In many cases, the speech between two consensus boundaries does not even constitute a complete unit prosodically: speakers sometimes produced abrupt breaks and resets in F0 without a full terminal contour. Because the listener-identified boundaries seem to be closest to IP boundaries, further analysis of the boundary locations was based on this hypothesis.

3.3. Experimenter labeling of Intonational Phrase boundaries

For the remaining analyses, the experimenter identified boundaries of IPs following the criteria given by Nespor and Vogel (2007) insofar as possible. This labeling is somewhat speculative, since as noted above, definitions of IPs are far from explicit. The notion of "melodic cohesion", coupled with what might be called "rhythmic cohesion", were the main criteria used for deciding how to label IPs in these speech samples. A clear break and re-start in the intonation or perceived rhythm of the speech was taken as the boundary of an IP. Since the speech examined here is spontaneous, a relatively small proportion of it consists of grammatically complete sentences. When there was a complete sentence, it was taken as coinciding with the end of an IP only if there was also some phonetic evidence of finality at the same location, such as a pause, lengthening, glottalization or an abrupt break in F0. In order to

determine whether the end of each extract should be identified as an IP boundary, reference was made to the longer recordings from which these extracts were taken. All but one of the extracts did end at an IP boundary. Data on the relation between locations marked as IP boundaries and the locations that listeners perceived as boundaries are given in Table 2.

	map tasks	broadcast
total labeled as IP boundary by experimenter	82	175
total marked as boundary by at least $^2/_3$ of listeners	35	85
total labeled as IP boundary by experimenter and marked as boundary by at least $^2/_3$ of listeners	35	78
total labeled as IP boundary by experimenter but not marked as boundary by at least $^2/_3$ of listeners	47	97
total marked as boundary by at least $^2/_3$ of listeners but not labeled as IP boundary	0	7

Table 2. Number of locations marked as boundaries by more than two-thirds of listeners and locations identified as an IP boundary by the experimenter.

From the table it can be seen that far more IP boundaries were marked by the experimenter than there were consensus boundaries identified by the listeners. Except for seven cases in the broadcast extracts, the consensus boundaries are a subset of the IP boundaries. Even though many IP boundaries were not marked by a consensus of the listeners, the average boundary score (proportion of listeners marking a boundary) was 0.58 across all experimenter-labeled IP boundaries, far above the average score for the rest of the words in the extracts (0.03). A working hypothesis is that the locations that the experimenter identified as IP boundaries are potential sites for listeners to mark boundaries, and that those marked by a consensus of listeners are the most salient of these.

At seven locations a consensus of listeners marked a boundary, but the experimenter did not mark an IP boundary. These occurred at the end of a clause or major syntactic phrase, where there was a continuous intonation contour and no pause, lengthening, glottalization or other interruption to the rhythm. In all but one of these, there was a substantial pitch rise just before the location that listeners marked as a boundary. Impressionistically, these occurred when the speaker was trying hard not to lose the floor. Most likely listeners marked a boundary because they noticed the syntactic boundary, but no IP boundary was marked by the investigator because of the absence of prosodic indicators. Given the importance that syntactic structure has in determining prosodic structure (Shattuck-Hufnagel and Turk 1996), it is not surprising that listeners' boundary marking was influenced by syntax.

3.4. Acoustic characteristics of speech preceding consensus boundary locations

In order to determine what measurable properties might contribute to explaining listener perceptions, a variety of measures was taken of the words and syllables immediately preceding all locations of interest: these included all experimenter-labeled IP boundaries, plus the seven consensus boundary locations in the broadcast extracts that were not marked as IP boundaries. The total number of locations measured was 82 in the map task extracts and 182 in the broadcast extracts. Only a subset of the measures that were tested are reported here.

3.4.1. Pauses

The first property investigated was whether boundaries were marked at the location of pauses. All pauses with duration greater than 150 ms were identified. (This duration has been used as the minimum pause duration in a number of studies such as Stirling et al. (2001) as it exceeds the likely duration of silence due to an epenthetic glottal stop, for example.) Intervals labeled as pauses included periods of silence, filled pauses, breathing, or a combination of these. Locations with pauses were

much more likely to be marked as consensus boundaries than locations without pauses, as can be seen in Table 3. This difference was significant for both sets of extracts, with $\chi^2 < .001$ in both.

Map tasks	Pause	No pause
Locations identified as consensus boundary by listeners	31	2
Experimenter-labeled IP boundaries not identified as boundary by listeners	12	37
Broadcast extracts	Pause	No pause
Locations identified as consensus boundary by listeners	61	24
Experimenter-labeled IP boundaries not identified as boundary by listeners	36	61

Table 3. Count of locations (IP boundaries plus all consensus boundaries) with or without pauses.

Recall that pauses were not a necessary or sufficient condition for marking an IP boundary; 11 locations where pauses occur were not marked as IP boundaries. Looking at all pauses, not just those that coincide with IP boundaries, the average boundary score at the locations of pauses was 0.69 for the map tasks and 0.62 for the broadcast extracts, compared to average boundary scores of 0.12 and 0.10, respectively, over all words. Thus, locations where pauses occur are favored as locations for boundaries, even though many IP boundaries that coincided with pauses were not marked as boundaries by the listeners. For listeners, pauses are strong but not sufficient cues for a boundary.

3.4.2. Changes in F0

Another acoustic characteristic that was examined at IP boundaries is the F0 contour during the word preceding the IP boundary. In French, pitch rises are commonly found at the end of an accent group that is non-final in the utterance, with the final high anchored to the last full syllable in the phrase (Delattre 1966, Welby 2006). Words that are final in a statement will usually end with a fall. Because both rises and falls can occur at the end of accent groups, it is unlikely that all the IPs in these recordings would end with the same type of contour. However, each of the extracts being analyzed was taken from a single conversational turn. Within a turn, speakers may end an utterance with a rise to signal that they are holding the floor. Thus in these data, the majority of IPs end with a rise on the phrase-final word. The question is whether the presence of a rise favored the perception of a boundary.

F0 values during the word preceding the IP or consensus boundary location were obtained automatically using a script in Praat. The values output by the script were hand-corrected as necessary (by measuring the duration of a period in the waveform). The F0 contours were classified as fall or rise depending on whether the maximum F0 preceded or followed the minimum F0. If the maximum preceded the minimum, the contour was classified as a fall, and as a rise for the reverse pattern. This procedure ignores the possibility of multiple peaks and/or valleys in the contour; pilot analyses were conducted which took these into account, but found to give virtually identical results to the much simpler binary classification. The numbers of locations with rises and falls are given in Table 4..[2]

As can be seen in Table 4, the pattern differed between the two sets of extracts. For the map task extracts, rises and falls did not differ in the likelihood that a boundary would be perceived. There were more rises than falls overall, and there was no significant difference in which occurred at locations that were or were not perceived as a boundary. In contrast, in the broadcast extracts, rises were found at consensus boundary locations significantly more frequently than their overall rate of occurrence would predict ($\chi^2 < .001$).

[2] In the broadcast extracts, there is one fewer data point than for pauses because F0 was unmeasurable in one IP-final word due to irregular phonation throughout the word.

Map tasks	Rise	Fall
Location identified as consensus boundary by listeners	23	10
Experimenter-labeled IP boundaries not identified as boundary by listeners	31	18
Broadcast extracts	Rise	Fall
Location identified as consensus boundary by listeners	61	24
Experimenter-labeled IP boundaries not identified as boundary by listeners	38	58

Table 4. Count of locations (IP boundaries plus all consensus boundaries) with F0 rise or fall.

A further analysis of the F0 contours investigated whether a greater magnitude of F0 excursion affected the likelihood of perception of a boundary. The difference in Hertz between the maximum and minimum F0 values during the phrase-final word was calculated. T-tests were calculated separately for rises and falls, to compare the magnitude of the excursion at locations that were perceived as boundaries with those that were not. Recall that for the map task conversations, the perception of a boundary was not associated with a difference in the frequency of occurrence of rises and falls. However, the magnitude of the F0 excursion in rises and falls did pattern differently for locations that were or were not perceived as boundaries. When F0 rose, the extent of the rise was significantly greater at locations perceived as boundaries ($t=2.02$, $p<.01$). Likewise, when F0 fell, the fall was greater in magnitude at locations that were perceived as boundaries ($t=2.13$, $p<.01$). For these map task extracts, there is an association between a larger change in F0 and the perception of a boundary following the word over which F0 rise or fell. But the direction of this change does not seem to be associated with any difference in perception.

Map tasks	Rise	Fall
Locations identified as consensus boundary by listeners	111 Hz	120 Hz
Experimenter-labeled IP boundaries not identified as boundary by listeners	73 Hz	58 Hz
Broadcast extracts	Rise	Fall
Locations identified as consensus boundary by listeners	121 Hz	50 Hz
Experimenter-labeled IP boundaries not identified as boundary by listeners	91 Hz	58 Hz

Table 5. Mean difference in Hertz between maximum and minimum F0 during phrase-final words.

The pattern was different in the broadcast extracts. In these, the magnitude of rise in F0 was significantly greater for locations perceived as boundaries ($t=1.99$, $p<.01$), but for words in which there was an F0 fall, the magnitude did not differ between locations that were or were not perceived as boundaries ($t=2.01$, ns). This result is counter to the pattern observed in the map task extracts, although (as can be seen by comparing the values in the right column of Table 5) the mean magnitude of F0 fall at locations not perceived as boundaries was identical for both types of extracts.

3.4.3. Final lengthening

Lengthening is widely recognized as typical of the final full (non-schwa) syllable of an accent group (see, e.g. Di Cristo & Hirst 1997, Pasdeloup 1990, Post 2000). Since any larger prosodic unit will be composed of accent groups, it might be expected that lengthening would cue a unit boundary for listeners. In spontaneous speech such as was studied here, it is not possible to do a controlled analysis of duration, so the analysis presented here must be treated with caution. The comparison of interest is between the durations of IP-final words that preceded consensus boundary locations, versus IP-final words that were not perceived as preceding a boundary. A methodology was devised to enable comparison of the durations of different words by taking into account the words' lengths, in number of

segments. This method does not, however, take into account differences in intrinsic duration among different segments, or differences in speech rate among the different speakers.

The method used was as follows. For each lexical word at the end of an IP or preceding one of the seven consensus boundaries that had not been labeled as an IP, the duration of that word was measured in Praat, then divided by the number of segments in the word as it had been produced in the recorded data. This gave a mean duration per segment. This was then compared between locations that were consensus boundaries and those that were not, with the prediction being that the mean duration would be longer for words preceding consensus boundaries. For the map task conversations, the mean duration was very slightly longer preceding consensus boundaries (124 ms vs 117 ms for other IP boundaries), but the difference was not statistically significant. For the broadcast extracts, the mean duration was significantly shorter preceding consensus boundaries (87 ms vs 116 ms, $t(124) = -4.29$), suggesting that listeners were more likely to have perceived a boundary when segment durations were shorter.

This puzzling result can be at least partially understood by closer examination of the data. While it was not surprising that the three shortest mean segment durations occurred at locations that listeners did not perceive as a boundary, the same was true of the 23 longest mean durations. This pattern is broadly compatible with Morel et al.'s (2006) finding that syllabic lengthening in excess of about 200% was perceived as hesitation, not lengthening, by their expert transcribers. In the present study, all but two of the 14 words whose mean segment duration exceeded 180 ms were found to involve a disfluency or a hesitation, such as repeated words or filled pauses. 180 ms is 175% of the mean segment duration of 103 ms, a slightly smaller percentage increase than what Morel et al. found to cue perception of hesitation. This finding suggests that even the naive listeners in this study were able to take disfluencies and hesitations into account in their decisions on boundary perception.

4. Discussion

The purpose of this study was to identify some of the measurable phonetic properties that lead listeners to perceive a phrasal boundary. The focus on acoustic properties is not intended to deny the role of syntax in prosodic structure; rather, given that spontaneous speech contains few syntactically complete sentences, it seems worthwhile to explore the acoustic factors that may be used by listeners to interpret a speech stream that cannot be segmented straightforwardly into syntactic phrases.

The approach taken was to compare the acoustic characteristics of locations in the speech stream that listeners perceived as boundaries with other locations that they did not. Although the listeners were free to mark a boundary between any two words in the extracts, the vast majority of spaces between words were not marked as a boundary by any listener. In other words, only a small subset of locations were in fact possible boundary locations. The sparseness of potential boundaries means that comparisons between words preceding consensus boundary locations and all other words in the extracts would be uninformative, as most words would be unlikely to have any characteristic of a pre-boundary location. To make the comparison of acoustic properties more informative, it was limited to locations that had the potential to be boundaries. These were taken to be the set of locations identified as IP boundaries by the experimenter.

The extracts that the listeners heard were single-speaker extracts. They were taken from conversations in which two speakers (for the map task extracts) or five (for the broadcast extracts) were participating, but in the portion presented to the listeners, only one speaker was heard. The conversation from which the broadcast extracts were taken was quite heated, and the speakers interrupted each other frequently. In order for one speaker to hold the floor for the 30 seconds or more that constituted an extract, he would almost certainly have used some explicit cues, prosodic or otherwise, to discourage interruption. This might imply that most or all of the boundaries within the broadcast extracts would have been perceived as non-final boundaries, that is, locations where a group of words ends, but the overall flow of speech is continuing. Non-final boundaries in French are usually cued by a rise in pitch over the last full syllable, which is compatible with the finding here that rises in F0 were strongly associated with the perception of boundaries in the broadcast extracts. The same was not true of the map task extracts, which might relate to the different nature of the speech task

represented in those extracts. The speakers in the map task were explaining travel routes to a conversational partner who was asking for information. This meant the speaker was the more dominant participant in the conversation, and the interlocutor seldom spoke until the speaker explicitly indicated that she had completed the requested explanation. In the map task extracts, the locations where listeners perceived boundaries might thus in many cases have been produced as utterance-final boundaries. This is consistent with the result that larger magnitude F0 changes (more typical of utterance-final boundaries than non-final boundaries) were more likely to be perceived as boundaries for both rises and falls.

From these results, it appears that F0 was quite important in contributing to listeners' responses. The different patterns observed in the two sets of extracts suggest that listeners' interpretation of F0 modulation was quite nuanced, and influenced by speakers' use of intonation in a particular speech sample. The occurrence of a pause was a reliable indicator in both sets of extracts, as expected. In contrast, there is no clear evidence that listeners used duration to decide on their boundary perceptions, as far as can be concluded from the imperfect analysis that was possible with these data. They did, however, seem to be capable of discounting hesitations in their interpretation of duration. Thus this study has shown that even naive listeners, under considerable time pressure, can make sophisticated evaluations of acoustic patterns when they are interpreting cues to prosodic structure.

References

Brennan, Robert & Dale Prediger (1981). Coefficient Kappa: Some uses, misuses and alternatives. *Educational and Psychological Measurement* 41, 687-699.

Cole, Jennifer, Yoonsook Mo & Mark Hasegawa-Johnson (2010a). Signal-based and expectation-based factors in the perception of prosodic prominence. *Laboratory Phonology* 1, 425-452.

Cole, Jennifer, Yoonsook Mo & Soondo Baek (2010b). The role of syntactic structure in guiding prosody perception with ordinary listeners and everyday speech. *Language and Cognitive Processes* 25, 1141-1177.

Delattre, Pierre (1966). Les dix intonations de base du français. *The French Review* 40, 1-14.

Di Cristo, Albert (1998). Intonation in French. Hirst, Daniel & Albert Di Cristo (eds.), *Intonation Systems. A Survey of Twenty Languages*. Cambridge University Press, Cambridge, pp.195-218.

Di Cristo, Albert (2000). Vers une modélisation de l'accentuation en français (seconde partie). *Journal of French Language Studies* 10, 27-44.

Di Cristo, Albert (2005). Éléments de prosodie. Nguyen, Noël, Sophie Wauquier-Gravelines & Jacques Durand (eds.), *Phonologie et phonétique : forme et substance*, Lavoisier, Paris, pp. 117-157.

Di Cristo, Albert & Daniel Hirst (1997). L'accentuation non-emphatique en français : stratégies et paramètres. Perrot, Jean (ed.), *Polyphonie pour Ivan Fónagy*. L'Harmattan, Paris, pp. 71-101.

D'Imperio, Mariapaola, Roxane Bertrand, Albert Di Cristo & Cristel Portes (2007). Investigating phrasing levels in French: Is there a difference between nuclear and prenuclear accents? Camacho, José, Nydia Flores-Ferrán, Liliana Sánchez, Viviane Déprez & María José Cabrera, *Selected Papers from the 36th Linguistic Symposium on Romance Languages* (LSRL). John Benjamins, New Brunswick, pp. 97-110.

Mertens, Piet (2006). A predictive approach to the analysis of intonation in discourse in French. Kawaguchi, Yuji, Ivan Fónagy & Tsunekazu Moriguchi (eds.), *Prosody and Syntax*. John Benjamins, Amsterdam, pp. 64-101.

Mettouchi, Amina, Anne Lacheret-Dujour, Vered Silber-Varod & Shlomo Izre'el (2007). Only prosody? Perception of speech segmentation in Kabyle and Hebrew. *Nouveaux cahiers de linguistique française* 28, 207-218.

Morel, Michel, Anne Lacheret-Dujour, Chantal Lyche & François Poiré (2006). Vous avez dit *proéminence* ? *Actes des XXVI*es *Journées d'Études sur la Parole*, Dinard, 183-186.

Nespor, Marina & Irene Vogel (2007). *Prosodic phonology: with a new foreword*. Walter de Gruyter, Berlin.

Obin, Nicolas, Xavier Rodet & Anne Lacheret-Dujour (2008). French prominence: a probabilistic framework. *Proceedings of ICASSP 2008* (Las Vegas), pp. 3993-3996.

Pagel, Vincent, Noël Carbonell, Yves Laprie & Jacqueline Vaissière (1995). Spotting prosodic boundaries in continuous speech in French. Elenius, Kjell & Peter Branderud (eds.), *Proceedings of the XIIIth ICPhS, Stockholm*, Vol. 4, 308-311.

Pasdeloup, Valérie (1990). *Modèle de règles rythmiques du français appliqué à la synthèse de la parole*. PhD dissertation, Université de Provence.

Portes, Cristel (2000). *Approche du rôle de la prosodie dans la structuration du discours oral en français*. DEA thesis, Université de Provence.

Post, Brechtje (2000). *Tonal and phrasal structures in French intonation*. PhD dissertation, Katholieke Universiteit Nijmegen.

Randolph, Justus (2008). Online Kappa Calculator. http://justus.randolph.name/kappa.

Shattuck-Hufnagel, Stefanie & Alice Turk (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research* 25, 193-247.

Smith, Caroline (2007). Prosodic accommodation by French speakers to a non-native interlocutor. Trouvain, Jürgen & William J. Barry (eds.), *Proceedings of the 16th ICPhS*, Saarbrücken, Germany, 1081-1084.

Stirling, Lesley, Janet Fletcher, Ilana Mushin & Roger Wales (2001). Representational issues in annotation: Using the Australian map task corpus to relate prosody and discourse structure. *Speech Communication* 33, 113-134.

Streefkerk, Barbertje, Louis Pols & Louis ten Bosch (1997). Prominence in read aloud sentences, as marked by listeners and classified automatically. *IFA Proceedings* 21, Institute of Phonetic Sciences, University of Amsterdam, pp. 101-116.

Welby, Pauline (2006). French intonational structure: Evidence from tonal alignment. *Journal of Phonetics* 34, 343-371.

Selected Proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology

edited by Scott M. Alvord

Cascadilla Proceedings Project Somerville, MA 2011

ISBN 978-1-57473-449-2 library binding

Ordering information

Orders for the library binding edition are handled by Cascadilla Press.
To place an order, go to www.lingref.com or contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, fax: 1-617-776-2271, sales@cascadilla.com

Web access and citation information

This entire proceedings can also be viewed on the web at www.lingref.com. Each paper has a unique document # which can be added to citations to facilitate access. The document # should not replace the full citation.

This paper can be cited as:

Smith, Caroline L. 2011. Acoustic Correlates of Listener-Identified Boundaries in Spontaneous French Speech. In *Selected Proceedings of the 5th Conference on Laboratory Approaches to Romance Phonology*, ed. Scott M. Alvord, 142-152. Somerville, MA: Cascadilla Proceedings Project. www.lingref.com, document #2643.