Prosodic disambiguation of syntactically ambiguous sentences

Grace Kuo

National Taiwan University, Taiwan

graciakuo@ntu.edu.tw

Abstract

In this study, we present a sentence gating paradigm to examine Taiwanese listeners' use of prosodic phrasing to interpret ambiguous sentences. The correspondence between the phonetic attributes of the prosodic structures and the listeners' choice scores and confidence ratings in the gating task are also analyzed. The results show that sentences were disambiguated *only* at the disambiguation points and the most salient cues were final lengthening and final pitch declination. Multiple linear regressions further show that listeners used not only duration and f0 measures but also voice quality to make their judgments.

Index Terms: prosodic cues, sentence disambiguation, gating paradigm

1. Introduction

Prosodic phrasing and prosodic prominence are the two prosodic devices that we use to disambiguate sentences. We use prosodic phrasing to disambiguate *Someone shot the servant of the actress on the balcony*, whose ambiguity results from user's attachment preference in the level of syntactic structure. On the other hand, we use prosodic prominence to differentiate the two interpretations in *SOMEONE shot the servant* vs. *Someone shot the SERVANT*, which deals with the ambiguity at the level of information structure. This study focuses on ambiguous sentences that result from different prosodic phrasing.

Previous perception studies have shown that some global ambiguities and most of the temporary ambiguities can be resolved by prosodic structure. For instance, in *Someone shot the servant of the actress who was on the balcony*, more interpretations of the high attachment (i.e., relative clause as modifying *the servant* rather than *the actress*) were perceived when there was a prosodic boundary between *the actress* and the relative clause *who was on the balcony*.

It is known that *speakers* reliably manipulate the two pre-boundary cues, duration and f0, to signal interpretations of an ambiguous sentence. A great deal of data also indicates the *listeners* make use of this prosodic information in sentence comprehension. [1] and [2] found that durational differences were often associated with the ambiguous constituents of the sentence or the ambiguous boundaries of the constituents. [3] found that both duration and intonation had significant effects in changing perceived meanings, but intensity showed a significant effect only when it was combined with the other cues. [4] found that lengthening could signal the occurrence of a syntactic boundary, and thus change the perceived meaning of a sentence. [5] found that naïve listeners could reliably use duration and intonation to separate structurally ambiguous sentences. [6] indicated that prosody has an immediate influence on listeners' expectations about upcoming syntactic structures; the information of duration and pitch in particular was interactively processed in interpreting the ambiguous sentences. [7] and [8] showed that when the syntactic boundaries and prosodic boundaries are in conflict, prosody interferes with the syntactic parse; if the two boundaries coincided, the prosodic structure facilitated the comprehension of the syntactic structure. Based on these findings, the present study on parsing Taiwanese ambiguous sentences tests the hypothesis that the durational pattern and the pitch contour would be correlated with listeners' processing of the syntactic ambiguity. In addition, [11] demonstrated that voice quality plays an important role in identifying prosodic boundaries of different sizes (prosodic word vs. tone sandhi group vs. intonational phrase), voice measures that reveal creakiness and breathiness will also be examined in this study.

The gating paradigm was implemented to study listeners' processing of tonally identical sentences in Taiwanese. The gating paradigm is mostly used in spoken word recognition research and has been considered particularly useful in examining moment-to-moment recognition processes and in assessing the amount and location of acoustic-phonetic information needed for the correct identification of a word. Gating can also be used in sentence recognition except now the gates are as big as a syllable or a word, such as [9] and [10]. [11] conducted a perceptual study with the gating paradigm in order to see whether listeners could predict the length of an entire sentence at any point within the sentence, or whether they must hear the potentially last word of that sentence. For example, the sentence *Earlier my sister took a dip* could end on *dip* (+ 0 word), or could continue with *in the pool* (+ 3 words), or could continue with *in the pool at the club* (+ 6 words). They gave listeners parts of the sentence syllable-by- syllable, up to the potentially last word *dip*, and at each gate, the listeners had to decide whether there was more to come, and if so, how much more (the choices are +0/+3/+6). This was inspired by an earlier study by [12] where he had the listeners hear the sentence through the word *dip*, yet *dip* was presented in fragments of increasing duration. The result of this earlier study was that English listeners were very accurate at predicting how much material was missing and that their predictions got better as they progressed through the potentially last word *dip*. The result of the later study revealed that listeners estimated a longer length of the sentence as they progressed through all versions of the sentence, and a differentiation between the three ending choices (+0/+3/+6) was only found when the listeners heard the potentially last word *dip*. In the present experiment, a similar sentence-gating task was employed with Taiwanese ambiguous sentences [13].

(1) a. early boundary condition

i33 chin33 gau33 kong55 kou51 | su33-lang33 ...

he very good at tell folklore private …

b. late boundary condition

i33 chin33 gau33 kong55 kou51-su33 | lang33 …

he very good at tell story the …

These sentences, as exemplified in (1), were considered ambiguous in that the sentences in each pair were phonetically similar (i.e. their sequences of surface tones were identical), and listeners would need to rely on their knowledge about Taiwanese prosodic phrasing in order to resolve the ambiguity. The pipe symbols indicate how the sentences are divided into tone sandhi group domains. In (1a), the first tone sandhi group domain ends at *kou51*, and *su33* is the onset of the second tone sandhi group (hence, *early boundary condition* hereafter); in (1b), *kou51-su33* is one prosodic word, and the first tone sandhi group domain ends at *su33* (hence, *late boundary condition*).

2. Gating

2.1. Participants

Fourteen native Taiwanese-speaking undergraduates were recruited. Four participants were excluded for not being able to disambiguate all temporarily ambiguous sentences in the end.

2.2. Materials

Eight pairs of sentences were created. The sequence of surface tones in each sentence pair were identical, yet the ambiguity could be resolved by the end of each sentence, leaving only one interpretation. A male speaker and a female speaker (both from Southern part of Taiwan) were recorded reading these sentences according to the interpretations given.

2.3. Procedure

All the listeners first passed a pre-test before participating in the actual experiment. They were asked to read all the test sentences (written in Chinese characters) verbally to the experimenter. The purpose of this pre-test was to (a) make sure the listeners used the same lexicon as they would hear in the recording, and (b) familiarize the listeners with the test sentences.

The experiment is a two-alternative forced-choice design. During the experiment, the two alternative sentences in each trial were visually presented with Chinese characters side-by-side on the laptop screen, counterbalanced for appearance on the right and left sides of the screen. The listeners saw the sentences first, and then heard a sentence gate. Their task was to listen to each gate and to determine after each presentation as quickly and accurately as possible whether the sentence gate that had been presented came from the sentence on the right or the sentence on the left. They were asked to click on their answer, and to indicate how certain they felt about their choice with a slider confidence rating scale. The listeners had to give a confidence rating for each trial in order to proceed to the next trial. The listeners saw a slider whose left end and right end were labeled "very unsure" and "very sure" respectively. There were no numbers labeled on the slider, but the listeners were explicitly told that the scale is gradient, not binary or categorical; they could give ratings anywhere on the scale. The gates were presented in an increasing word-blocked fashion; that is, all the gates with only one word were presented first, followed by all the gates with two words, three words, and so on. The last gate for each sentence corresponds to the entire sentence

2.4. Analyses

Two dependent variables were measured in this experiment. The first was listeners' choice of response. The listener's choice was given a score of '1' if the listener chose the *late boundary condition* for the answer and '0' if the listener chose the *early boundary condition*. In other words, a score close to '1' for the *late boundary condition* sentences and a score close to '0' for the *early closure condition* sentences indicate higher correctness.

The other dependent variable was listeners' confidence rating regarding each gate. This is to examine how confident they felt about their decision. Listeners' ratings were given using a slider bar, but the program converted bar positions to values on a 1-100 scale.

In order to study the major phonetic cues that the listeners use to disambiguate sentences, we measured the rime duration, f0 mean, f0 median, f0 range, f0 slope, H1*-H2*, HNRs and CPP of all the syllables in the sentences; values were obtained from VoiceSauce [14]. These acoustic measure of duration, pitch and voice quality were considered useful to distinguish prosodic boundaries of different sizes in Taiwanese.

3. Results and Discussion

Because the sentence pairs have different numbers of gates preceding the disambiguation points, the "choice scores" and the "confidence ratings" of only six selected gates are presented here, namely "the first gate", "the last gate", the disambiguation point in the *early boundary condition* (henceforth: "the early DP gate"), the disambiguation point in the *late boundary condition* (henceforth: "the late DP gate"), the gate that preceded the early DP gate (henceforth: "the pre-early gate") and the gate that immediately follows the late DP gate (henceforth: "the post-late gate"). Take syllables in (1a) and (1b) for example. We examined the choice scores and confidence ratings of *i33* ("the first gate"), *kong55* ("the pre-early gate"), *kou51* ("the early DP gate"), *su33* ("the late DP gate"), *lang33* ("the post-late gate"), and the last gate in each sentence.

3.1. Choice Score

The choice scores were entered for repeated measures ANOVA in R with the factors of the 6-level "gate", and the 2-level "boundary condition" (i.e., *early boundary condition* and *late boundary condition*).

The statistical results and the average choice scores across all listeners and sentences are shown in Table 1 and Figure 1. Significant effects were found for "boundary condition" and the interaction between "boundary condition" and "gate".

Table 1: Repeated measures ANOVA for Choice Score

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BoundaryCondition	1	16.02	16.02	134.94	< .01 *
Gate	5	0.26	0.05	0.45	0.817
BoundaryCondition × Gate	5	9.14	1.83	15.41	< .01*
Residuals	936	111.1	0.12		

As can be seen in Figure 1, the choice score increases across gates for the *late boundary condition*, while it decreases across gates for the *early boundary condition*. Post hoc comparisons of the choice scores revealed that (a) for the *late boundary condition*, the score at "the first gate: and "the pre-early gate" were significantly lower than "the early DP gate" which in turn had a lower score than "the late DP gate", "the post-late gate" and "the last gate"; (ii) for the *early boundary condition*, the choice scores of "the first gate" and "the pre-early gate" were significantly lower than "the early DP gate", "the late DP gate", "the post-late gate", and "the last gate". The gates with statistically the same scores are enclosed in the same ellipses in Figure 1.

The critical gate is "the early DP gate". This gate is when listeners detected a boundary in the *early boundary condition*, and when listeners encountered the absence of a boundary in the *late boundary condition*. The listeners did not give different scores until they reached this gate, and they did not show any difference in scores even at "the pre-early gate". This suggests that not enough boundary cues were provided before this disambiguation point.

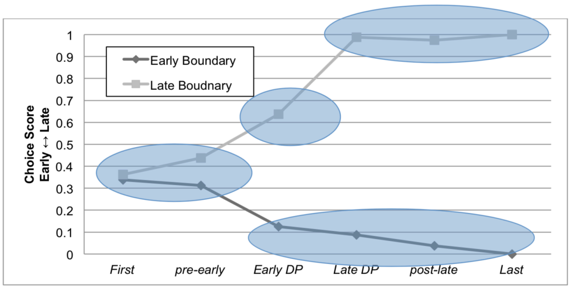

Figure 1: *Average choice score across Gates and Boundary Conditions. Ellipses indicate no significant difference within a condition.*

Table 2 presents the average choice scores as well as the standard deviations. In the *early boundary condition,* the standard deviation gradually declined as the listeners progressed through the sentence, whereas in the *late boundary condition*, the variation in score showed a sudden reduction after "the early DP gate". It appears that for the *late boundary condition*, listeners were not so sure about their choice (i.e., more variation in score) as they detected a possible absence of a boundary at "the early DP gate"; however, the variation got reduced when they detected a real boundary at "the late DP gate".

Table 2: Average choice scores and standard deviations across Gates and Boundary Conditions.

	First	Pre-early	Early DP	Late DP	Post-late	Last
	mean (sd)	mean(sd)	mean (sd)	mean (sd)	mean (sd)	mean (sd)
Late	0.36 (0.48)	0.44 (0.50)	0.64 (0.48)	0.99 (0.11)	0.98 (0.16)	1.00 (0)
Early	0.34 (0.48)	0.31 (0.47)	0.13 (0.33)	0.09 (0.28)	0.04 (0.19)	0 .00(0)

The choice score results show that listeners made accurate judgments only after they reached "the early DP gate", whether the sentence was from the *early boundary condition* or the *late boundary condition*.

3.2. Confidence Rating

The confidence ratings were entered for repeated measures ANOVA in R with the same factors, "gate" and "boundary condition". The statistical results are shown in Table 3. A significant main effect was found only in "gate".

Table 3. Repeated measures ANOVA for Confidence Rating

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
BoundaryCondition	1	3	3	0.004	0.947
Gate	5	118796	23759	42.183	< .01*
BoundaryCondition × Gate	5	936	187	0.332	0.893
Residuals	936	527190	563		

In a typical gating experiment, "recognition point", the point when listeners make a correct choice without further changes and their confidence rating about the choice is 80% or higher, is usually examined. With the threshold (= 80%) in mind, we find that listeners' confidence rating went beyond 80% after they reached "the late DP gate" in both conditions, as shown with the dashed line at 80% in Figure 2.

Therefore, in terms of confidence, the critical gate was "the late DP gate". It seems that whether the listeners detected a boundary or not determined their confidence about their choice.

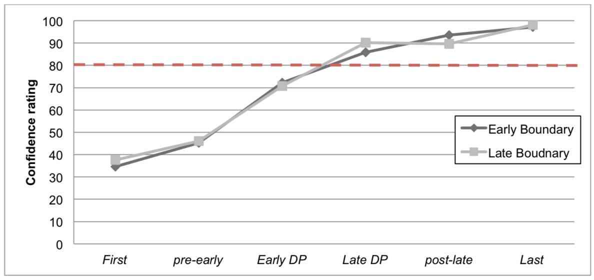

Figure 2: Average confidence ratings across Gates and Boundary Conditions.

In summary, the more gates the listeners heard, the more confident they were about their choice scores. To conclude from the choice score and the confidence rating results, it was only at "the early DP gate" and "the late DP gate" that the listeners clearly discriminated between the two boundary conditions. In addition, listeners showed greater confidence in their choices after they reached beyond "the early DP gate". This suggests that listeners needed to perceive enough prosodic cues at this point to produce the perception results.

3.3. Acoustic measures

Given that listeners started to give accurate judgments at "the early DP gate" and "the late DP gate", we are interested in the acoustic cues that they received *before* and at these disambiguation points in each condition. The acoustic measures were obtained from the voiced portion of each syllable before and including "the late DP gate". In the following figures, the syllables are right-aligned: '1' is "the late DP gate", '0' is "the early DP gate", '-1' is "the pre-early gate", and so on.

Among all the acoustic measures listed in the previous section, only the ones related to duration and f0 showed significant results.

3.3.1. Final Lengthening

With "boundary condition" as the independent variable, paired *t*-tests at each gate revealed that durations were only significantly different at "the early DP gate" ($t(7)=10.71$, $p<.05$) and "the late DP gate" ($t(7)=5.68$, $p<.05$). The results are displayed in Figure 3. The increase in duration at both gates suggests that there was a duration contrast caused by final lengthening in each boundary condition. The final lengthening was only realized at the last syllable in each condition rather than increasing gradually through the sentences. Thus, it could have been a cue to make the two boundary conditions distinct from each other.

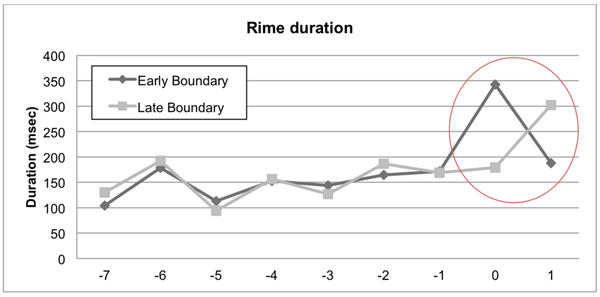

Figure 3: Mean duration (ms) of the two boundary conditions at the nine positions.

3.3.2. Final pitch declination

Paired *t*-tests at each position revealed that f0 mean and f-median in the two boundary conditions differ significantly at “the late DP gate” (f0 median: $t(7) - 6.77$, $p < .05$; f0 mean $t(7)$-5.61, p<.05). This is because “the late DP gate” (= point “1”) in the *early boundary condition* began a new tone sandhi group and so involved f0 reset. On the other hand, f0 range differs at “the early DP” (= point “0”) ($t(7)$=2.93, p<.05), but not at “the late DP gate”. Consistent with previous studies, f0 range at a prosodic boundary is found to be wider (i.e., “the early DP gate” in the *early boundary condition*).

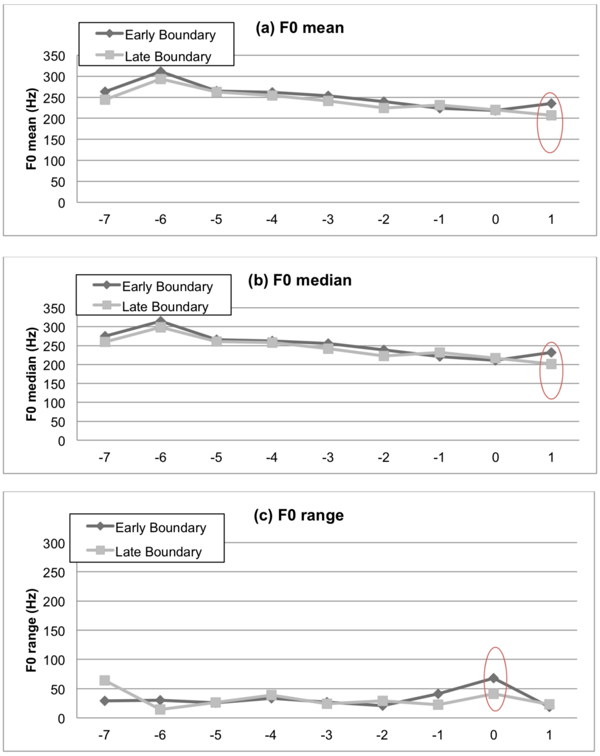

Figure 4: f0 mean, f0 median, and f0 range across the two boundary conditions at the nine positions.

In summary, the results for f0 mean and f0 median confirmed that pitch declines as a sentence continues and they also provide evidence for pitch reset phrase initially. The results for f0 range suggests that a wider f0 range tends to appear phrase-finally.

For further analysis, the correlations between listener’s responses and acoustic measures at each position were examined, and the multiple linear regressions were carried out in R. In Table 4, the ticks indicate the acoustic measures that contributed significantly to the repression equation. The proportion of explained variance (R^2) are also given.

Table 4. Multiple regression analyses for choice scores and confidence ratings vs. acoustic measures.

	Choice Score	Confidence Rating
Duration	√	√
F0 range	√	√
F0 slope	√	√
F0 median	√	√
F0 mean		√
H1*-H2*	√	
HNR05	√	√
HNR15		
HNR25	√	
HNR35	√	√
CPP	√	√
R^2	0.05	0.19
F	5.69	27.97
p	< .01	< .01

Previous studies have established that duration and f0 determine which interpretation listeners would assign to a syntactically ambiguous sentence. The acoustic analyses show clear differences in duration and f0 for the comparison of the two boundary conditions and thus support this claim. The results also suggest that sentences were disambiguated only at the disambiguation points, but not earlier. However, the multiple regression results revealed that listeners’ responses (both choice scores and confidence ratings) correlate with not only duration and f0 measures but also voice and noise cues.

4. Conclusion

Previous studies have established that duration and f0 determine which interpretation listeners would assign to a syntactically ambiguous sentence. The acoustic analyses show clear differences in duration and f0 for the comparison of the two boundary conditions and thus support this claim. The results also suggest that sentences were disambiguated only at the disambiguation points, but not earlier. However, the multiple regression results revealed that listeners’ responses (both choice scores and confidence ratings) correlate with not only duration and f0 measures but also voice and noise cues.

5. References

[1] I. Lehiste, “Phonetic disambiguation of syntactic ambiguity,” *Glossa*, 7, pp. 107–122, 1973.

[2] I. Lehiste, J. Olive, and L. Streeter, “Role of duration in disambiguating syntactically ambiguous sentences,” *Journal of the Acoustical Society of America*, 60, pp. 1199–1202, 1976.

[3] L. Streeter, “Acoustic determinants of phrase boundary location,” *Journal of the Acoustical Society of America*, 64, pp. 1582-1592, 1978.

[4] D. Scott, “Duration as a cue to the perception of a phrase boundary,” *Journal of the Acoustical Society of America*, 71, pp. 996-1007, 1982.

[5] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fond, “The use of prosody in syntactic disambiguation,” *Journal of the Acoustical Society of America*, 90, pp. 2956-2970, 1991.
[6] C. Beach, “The interpretation of prosodic patterns of points of syntactic structure ambiguity: evidence for cue trading relations,” *Journal of Memory and Language*, 30, pp. 644-663, 1991.
[7] A. Geers, “Intonation contour and syntactic structure as predictors of apparent segmentation,” *Journal of the Acoustical Society of America*, 4, pp. 411-458.
[8] S. Speer, M. Kjelgaard, and K. Dobroth, “The influence of prosodic structure on the resolution of temporary syntactic closure ambiguities,” *Journal of Psycholinguistic Research*, 25, pp. 247-268.
[9] V. Aubergé, T. Grépillat, and A. Rilliard, “Can we perceive attitudes before the end of setnences? The gating paradigm for prosodic contours,” *Proceedings of the European Conference on Speech Communication and Technology*, 1997, vol. 2, pp. 871-874.
[10] V. J. van Heuven, J. Haan, E. Janse, and E. J. van der Torre, “Perceptual identification of sentence type and the time-distribution of prosodic interrogativity marker in Dutch,” *ETRW Workshop on Prosody*, 1997, pp. 317-320.
[11] F. Grosjean, C. Hirst, “Using prosody to predict the end of sentences in English and French, normal and brain-damaged subjects,” *Language and Cognitive Processes*, 11, pp. 107-134.
[12] F. Grosjean, “How long is the sentence? Prediction and prosody in the on-line processing of language,” *Linguistics*, 21, pp. 501-529.
[13] G. Kuo, “Perceived prosodic boundaries in Taiwanese and their acoustic correlates,” *Interspeech 2012 Proceedings.*
[14] Y. Shue, P. Keating, C. Vicenik, K. Yu, “VoiceSauce: a program for voice analysis,” *Proceedings of the ICPhS XVII*, pp. 1846-1849.