Native English listeners’ perceptions of prosody in L1 and L2 reading

Caroline L. Smith, Paul Edmunds
Department of Linguistics, University of New Mexico, Albuquerque, NM, USA
caroline@unm.edu, pedmunds@unm.edu

Abstract
Rapid Prosody Transcription (RPT) was used to investigate listeners’ perceptions of prosody in reading by native and non-native English speakers. RPT offers a language-independent tool to access listeners’ holistic understanding of prosody. Listeners hear an audio recording of speech while following along on an orthographic, unpunctuated transcript of the recording. They indicate their perception of phrasal boundaries or prominent words by marking them on the transcript in real time. Our listeners showed higher agreement for boundary-marking in the native speakers’ reading than the non-natives’. Listeners marked more boundaries in the non-natives’ reading, likely because the non-natives paused more often, although listeners partially compensated by not marking boundaries as often when non-natives made short pauses. For prominence, rates of agreement were higher for the non-natives. This may be due to listeners’ marking fewer prominences in the non-natives’ reading, meaning that they agreed on the absence of prominent words. Compared to acoustic analysis, studying listener reactions provides more insight into what aspects of non-native prosody are most salient. This may be useful in guiding learners to the most effective ways to improve their prosody.

Index Terms: prosody, perception of L2 speech, phrasal boundaries, prominence

1. Introduction
Of all the dimensions involved in speaking, prosody is one where accurate production may be most critical for speakers using their second language (L2), as so much of the suprasegmental information in an utterance must be used by the listener in order to interpret the intended message. If a speaker’s realization of prosodic qualities does not match a listener’s expectations, the message may be misinterpreted or confused. The present study investigates this problem by testing how native English listeners perceive two aspects of prosody, phrasal boundaries and prominence, in the speech of native and L2 speakers of English.

Rapid Prosody Transcription [1], [2] was used to compare listeners’ perceptions of the phrasing produced by native and non-native speakers of English reading aloud. In Rapid Prosody Transcription (RPT), listeners hear an audio recording of speech while following along on an orthographic, unpunctuated transcript of the recording. They indicate their perception of phrasal boundaries or prominent words by marking them on the transcript in real time. Good rates of agreement have been achieved in this task, making it a useful tool to access naïve listeners’ perceptions of some aspects of prosodic structure. Here we adapt this technique to examine native English listeners’ perceptions of prosodic phrasing and prominence as produced by native and non-native English speakers. Given the importance of prosody in effective communication, it could be useful to evaluate non-native speakers’ success at producing phrasing and prominence that is readily perceptible. Comparison of speakers with different language backgrounds can help to distinguish aspects of the distribution and production of prosody that are language-specific from those that characterize many or all languages.

RPT may be more suited to evaluating non-native speakers than an expert transcription scheme that incorporates language-specific assumptions, since learners may transfer some aspects of the prosody of their L1 to their L2 production [3], so that a language-specific system is a poor representation of their production. With the goal of evaluating the impression that speech makes on listeners, RPT may be more useful than acoustic analysis in that it reflects listeners’ adaptation to different speakers. Listeners respond to the signal as a whole, while measurement of individual acoustic parameters may fail to capture the relevant dimensions.

We examined listeners’ perceptions of two dimensions of prosody: phrasal boundaries and prominent words. In English, a variety of acoustic modifications cue both of these, including changes to F0, duration, intensity and segmental quality. Our expectation is that non-native speakers do not fully control the cues used by native English speakers to indicate prosodic patterns, so our first hypothesis is that listeners will show a lower rate of agreement when responding to the speech of non-natives.

2. Recordings used in study
To enable a controlled comparison of the speakers’ productions, they were recorded reading a standard text, the Rainbow Passage [4]. Recordings were made of 12 native speakers of American English and 12 non-native speakers, with six men and six women in each group. The non-native speakers were all advanced learners of American English whose native language is Latin American Spanish.

All speakers were recorded reading the Rainbow Passage three times as part of a larger set of recordings. For each speaker, the recording with the fewest errors was selected to be played to listeners. For all but two speakers (one female native and one female non-native), there were reading errors in all recordings. These were transcribed accurately in the transcripts that were presented to the listeners, so the number of words in each recording varied from a minimum of 329 to a maximum of 382. The non-native speakers averaged 351 words in 149 s, the natives 345 words in 116 s.

The recorded readings were segmented in Praat [5] into “phrases” demarcated by any pause of more than 150 ms duration. The word “phrase” is used here without theoretical implications to refer to a group of words demarcated by interruptions to the flow of speech (silence or filled pauses). On average, non-native speakers had longer durations of both speaking time (excluding pauses) and pause time. The total number of words produced varied over a similar range for both native and non-native speakers, so the non-native speakers’ longer speaking time implies a slower speech rate.
3. Experiment 1: Phrasing

Impressionistically, the non-native speakers seemed to have more frequent breaks in the flow of their speech than did the native speakers. Many of these breaks seem to be hesitations rather than intentional pauses, but listeners might still interpret them as boundaries. We thus expected our listeners to mark phrasal boundaries more frequently in the non-natives’ speech.

3.1. Method

The listeners were native speakers of American English recruited from introductory linguistics classes who were randomly assigned to one of two groups. Eleven listeners responded to recordings of the native speakers, and a different eleven listeners heard the recordings of the non-natives. They all reported having American English as their native and dominant language.

The instructions were adapted from those used by Cole et al. [2], and read in part: “A feature of normal speech that we are interested in is the way speakers break up an utterance into chunks. These chunks group words in a way that helps the listener interpret the utterance, and are especially important when the speaker produces long stretches of continuous speech. … A chunk may be as small as a single word, or it may contain many words, and speakers can vary quite a bit in the size of the chunks they produce.” Because the listeners were untrained, no linguistic description was given for the type of unit they are to listen for. We use the word “phrase” here simply to refer to the groups of words between two boundaries perceived by listeners.

Listeners participated individually in a quiet office. Presentation of the recordings was controlled by a Praat script; listeners clicked on a button to initiate playing of each recording, which they heard over headphones. They responded to one practice recording with a different text and voice before beginning the experiment, and each listener heard the recordings in a different random order. The practice recording was of a native American English speaker for the listeners assigned to the native-speaker group; listeners assigned to the non-native-speaker group heard a practice recording of a non-native speaker. They marked phrase boundaries by writing a slash / on the paper transcript as they listened. These responses were then tabulated in Excel spreadsheets. One listener failed to mark any boundaries in one recording of a non-native speaker.

Most statistical analyses reported here were calculated in Excel. A boundary score (b-score) was calculated for each word in each recording. The b-score is equal to the proportion of listeners who marked a boundary following that word. B-scores range from 0 (no listeners marked a boundary after the word) to 1 (all listeners marked a boundary).

The Online Kappa Calculator [6] was used to calculate the kappa statistic with free marginals [7], which assesses rates of agreement taking chance agreement into account.

3.2. Results

3.2.1. Agreement among listeners

Our first hypothesis, that listener agreement would be lower for non-natives’ speech, was supported. The mean kappa for listeners who heard native speakers was 0.91; the mean with non-native speakers was 0.87. The difference between the kappa values for listener responses to the two groups of speakers (n=12 for each group) was significant in a t-test (two-tailed, equal variances, t=3.5, p=.002). In addition, agreement was significantly lower for male non-native speakers compared to female non-natives, but there was no difference between the genders for the native speakers.

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>0.91</td>
<td></td>
</tr>
<tr>
<td>Non-Native</td>
<td>0.90</td>
<td>0.84</td>
</tr>
</tbody>
</table>

3.2.2. Frequency of marking boundaries

Our second hypothesis was also supported: listeners marked significantly more boundaries in recordings of non-native speakers (mean 33.6) than for native speakers (mean 27.6). This difference was also significant in a t-test (t=3.28, p=.003). Their rate of boundary marking (which takes into account the varying number of words in the different recordings) was also more frequent with the non-native speakers: the mean number of words per marked boundary was 10.7 for the non-natives compared to 12.7 for the native speakers (t=3.73, p=.001).

Next we compared how frequently different listeners marked boundaries. Looking again at the number of boundaries marked per recording, but taking the mean for each listener rather than for each speaker, the range of values was greater for the different listeners than for the different speakers. Because of the greater variability when the data were averaged this way, there was no significant difference in the rate of marking among the two groups of listeners (t=1.28, p>.1). Since these two groups of listeners were composed of different individuals, the lack of difference between the groups validates the claim that the significant differences between speaker groups are due to differences in the speakers, not the listeners.

3.2.3. Relation to occurrence of pauses

Using Praat, all non-speech intervals of 150 ms or longer were identified. These are referred to as pauses although the shorter ones may not be perceived as such.

One explanation for why listeners marked boundaries more frequently with the non-native speakers is that these speakers paused more often: regression $r^2 = 0.58$ for number of pauses as a predictor for the number of words between boundaries. But listeners may also have compensated somewhat: Figure 1 shows that fewer listeners marked boundaries where there were pauses under one second in the speech of the non-natives (ESL) than in the speech of the native speakers (NS). This suggests that they understood that some of these short pauses in the non-natives’ speech were not marking boundaries in the structure, but were due to some kind of disfluency. Most of the native speakers also experienced some disfluency in their reading, so it is possible that the same kind of compensation could have occurred to a lesser extent in listeners’ responses to native speakers’ reading.
Figure 1: Boundary scores associated with different pause durations produced by native speakers (NS) and non-natives (ESL).

4. Experiment 2: Prominence

When the experimenters listen to the recordings of the non-natives, we perceive that they place approximately equal emphasis on all content words as they read. Rasier and Hiligsmann [3] claim that L2 learners in general “seem to have much difficulty distinguishing between old and new information and therefore tend to emphasize nearly each word in the utterance.” A study of reading in English by non-native speakers [8] noted much the same behavior in the Spanish-speaking participants, who are described as producing all prominences with maximal amplitude. And Hualde has suggested that in L1 Spanish, speakers tend to place pitch accents on almost all content words [9]; in contrast, English speakers produce a pitch accent on just a subset of content words. We therefore expected that when listening to the non-natives, our listeners might have difficulty deciding that any one word was more prominent than another.

If the non-natives transfer their L1 habits to English, it might lead to listeners’ perceiving almost all words as pitch accented. Alternatively, the listeners might bring their English-based expectations to bear, and assume that only a minority of words will receive prominence. That strategy, coupled with the non-natives’ lack of differentiation among words, might result in listeners’ marking a very small number of words as prominent. We therefore expected that the listeners would mark either far more or far fewer prominences in the speech of the non-natives.

4.1. Method

The listeners in this experiment were recruited from the Psychology department’s subject pool. They all reported having American English as their native and dominant language, and had no training in linguistics. None of them had participated in Experiment 1. Twelve listeners responded to recordings of the native speakers, and a different twelve heard the recordings of the non-natives.

The listeners in Experiment 2 practiced on the same practice recording as for Experiment 1, and presentation of the recordings was controlled in the same way. Listeners were instructed to underline words that they perceived as prominent. The instructions, adapted from [1], read in part: “The prominent words are in a sense highlighted for the listener, and stand out from other non-prominent words. Prominent words are often distinguished by being louder or longer than you would expect them to be.” One listener failed to mark any prominences in two recordings of non-native speakers.

A prominence score (p-score) was calculated for each word in each recording. The p-score is equal to the proportion of listeners who marked the word as prominent. Like the b-scores, p-scores range from 0 (no listeners marked the word as prominent) to 1 (all listeners marked it). Other analyses were done as in Experiment 1.

4.2. Results

4.2.1. Agreement among listeners

Surprisingly, the results of Experiment 2 contradicted the hypothesis that listener agreement would be lower for non-natives’ speech. The mean kappa for listeners who heard native speakers was 0.63; the mean with non-natives was 0.78. The difference between the kappa values for the two groups of speakers (t=12 for each group) was significant in a t-test (two-tailed, equal variances, t=11.3, p<.001).

Kappa values between 0.6 and 0.8 are said to indicate “substantial” agreement [10]; even the kappas for the native speaker group were still in this range. But the kappa values for listeners responding to the native speakers were all lower than those for any non-native speaker. While lower rates of agreement for prominence than boundaries have been consistently reported in other studies using RPT, e.g., [1], [11], the rates obtained in this study for the native speakers are unusually low. Also, unlike what was found for boundary-marking, there was no difference in agreement between male and female speakers, within speaker type.

Table 2. Free marginal kappa statistics, listeners’ prominence marking for different speaker groups.

<table>
<thead>
<tr>
<th></th>
<th>Female</th>
<th>Male</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native</td>
<td>0.64</td>
<td>0.63</td>
</tr>
<tr>
<td>Non-Native</td>
<td>0.78</td>
<td>0.79</td>
</tr>
</tbody>
</table>

4.2.2. Frequency of marking prominent words

Listeners marked significantly more prominences in recordings of native speakers (mean 43.9) than for non-native speakers (mean 22.8). This difference was significant in a t-test (t=9.8, p<.001). (This is the opposite of the pattern observed in Experiment 1, where listeners marked more boundaries for the non-natives.) For prominences with the native speakers, the range of means averaged across all listeners was 33.8 to 51.4 prominences marked per recording; for the non-native speakers, the range was 17.7 to 32.7. On average, listeners marked a prominence every 8.0 words for the native speakers compared to every 15.8 words for the non-natives (t=9.20, p<.001). Even for the native speakers, this is slightly less frequent than the listeners in [1], who heard spontaneous speech from native speakers and marked prominences every 4 – 8 words.

Next we compared how frequently different listeners marked prominences. Looking again at the number of prominences marked per recording, but taking the mean for each listener rather than for each speaker, the range of values was greater for the different listeners than for the different speakers. (This is consistent with what was found in Experiment 1.) For the listeners who heard the native speakers, the range of mean prominences marked per recording was 23.3
to 96.0. For the listeners who heard the non-natives, the range was 11.1 to 60.8 prominences marked per recording by different listeners. Although these ranges overlapped, the difference between the two groups was significant (t=2.87, p<.01). This is unlike what was found in Experiment 1, and suggests that the differences between groups could possibly be due to differences between the two groups of listeners, rather than between the two groups of speakers.

The higher agreement for non-natives can be at least partly explained by the lower frequency with which listeners marked prominences in the non-natives’ reading. That is, the agreement was primarily over the absence of prominences, rather than their presence. Looking just at those words that received a p-score of 0.66 or higher (meaning at least nine out of twelve listeners marked them), only five words met this criterion in all of the non-natives’ recordings, out of a total of 4219 words. Two of these occurred in the speech of one female non-native speaker, the other three in a second female non-native. In contrast, for the native speakers, a total of 54 words (out of 4161) had p-scores above 0.66, and all but one native speaker had at least one word that was identified with this level of consensus.

The low frequency of marking prominence in the non-natives’ recordings suggests that the second explanation hypothesized in section 4 was supported: the listeners expected prominence to occur on a subset of content words, and so they marked only a few as prominent. The task of identifying which words to mark may have been particularly difficult in the non-natives’ speech, and the listeners’ solution was to mark fewer words.

5. Boundaries and Prominences

Comparing listeners’ patterns of marking boundaries and prominences, one striking result is that for all of the non-native speakers, listeners marked more boundaries than prominences, whereas the reverse was true for the native speakers. (Recall that different listeners participated in each task.)

If the boundaries demarcate salient phrasal breaks, then in the speech of the native speakers, listeners perceived multiple prominences within each “phrase.” This is consistent with English allowing both nuclear and pre-nuclear pitch accents within an intonational phrase. But in the speech of the non-natives, listeners perceived, on average, less than one prominence per phrase. Two explanations are possible: listeners marked some boundaries that were not really boundaries, possibly coinciding with breaks due to disfluency. Another possibility is that listeners failed to mark (or speakers failed to indicate) even a single, nuclear prominence in some phrases. These factors could be combining to result in a lower ratio of perceived prominences to perceived boundaries.

6. Discussion and Conclusions

Native English listeners were less consistent in their perceptions of phrasal boundaries in reading aloud by non-native speakers of English than in reading by other native English speakers. They tended to hear more boundaries in the non-natives’ readings, but at many locations just a few listeners marked a boundary. Listeners were able to compensate to some extent for the large number of pauses produced by the non-natives, because they marked fewer boundaries when non-natives made a short pause than when native speakers made a pause of comparable duration.

The listeners responded rather differently to the task of marking prominent words. They marked fewer words as prominent in the non-natives’ readings than in the natives’, and only five words in the non-natives’ readings were identified as prominent by a consensus of listeners. Thus the non-native speakers failed almost completely to produce prominences that were discernible to the listeners. The listeners had low rates of agreement for marking prominence in the native speakers’ reading, but a consensus of them did agree that 54 words were prominent (across eleven of the twelve speakers). Given the overall higher frequency of marking prominence in the native speakers, it seems that the low agreement rates are due to different listeners perceiving (or choosing to mark) different prominences. The native speakers do seem to have produced prominences that the listeners could hear, but the listeners responded inconsistently.

These results provide some suggestions as to what native English listeners notice in reading by non-native speakers. They seemed to be able to compensate to a fair extent for the pauses made by the non-natives that were a consequence of hesitations, not boundary indicators. But the listeners had much more trouble perceiving prominences in the non-natives’ reading. Further analysis of the acoustics of the natives and non-natives’ productions will be necessary in order to identify the source of this trouble.

7. Acknowledgements

We would like to thank the Psychology Department (Jane Smith, Chair) for facilitating our use of their subject recruitment facilities.

8. References