Perception of Prosodic Prominence and Boundaries by L1 and L2 Speakers of English

Gábor Pintér1, Shinobu Mizuguchi2, Koichi Tateishi3
1 School of Languages and Communication, Kobe University, Japan  
2 Graduate School of Inter-Cultural Studies, Kobe University, Japan  
3 Department of English, Kobe College, Japan

Abstract
This study provides a direct comparison of boundary and prominence perception strategies between Japanese EFL learners and native speakers of English using the Rapid Prosody Transcription (RPT) method. Although RPT experiments are available for both native English speakers [1], [2], [3] and Japanese EFL learners [4], a direct comparison of the available data is problematic as the stimuli sets used in the experiments are not identical. The present research addresses this issue by using identical stimuli sets across L1 and L2 listeners. The data for native English speakers was taken from RPT experiments carried out by Jennifer Cole with Yoonsook Mo and colleagues [1-3, 5-8]. The non-native data was collected by re-running a subset of RPT tasks reported in the work by Cole and Mo with 108 Japanese undergraduate students. Although native speakers perceived more boundaries and prominent words than L2 speakers, the results outlined surprisingly similar perceptual strategies. A strong correlation was found between the responses of native speakers and Japanese learners of English in both boundary and prominence perception tasks. In boundary perception both groups relied heavily on silent pauses and vocal fillers. In prominence detection the responses correlated with vowel duration, maximum amplitude, and maximum pitch in this specific order for both language groups.

Index Terms: speech prosody, perception of L2 speech, Rapid Prosody Transcription, Japanese EFL, boundaries, prominence

1. Introduction
Recent developments in the research of L2 prosody demonstrate that judgments of degree of foreign accent and intelligibility of learners’ speech relies heavily on prosodic cues. Although segmental mistakes, such as confusion between L and R segments, can be a source of misunderstanding, unexpected use of prosodic cues—such as timing, intensity and pitch—can render speech unintelligible [9], [10], [11]. These findings suggest that substantial improvements in terms of intelligibility can be achieved in educational settings by raising awareness of prosodic cues of the target language and providing guidelines for corrections. In order to do so, however, it is necessary to know the differences between L1 and L2 perceptual strategies. Although descriptions discussing prosodic cues of individual languages are not uncommon, comparative studies, especially those that allow for direct comparison of perceptual responses, are rare. This study aims to fill this gap by comparing results of prosody perception tasks carried out in parallel with native speakers and Japanese EFL learners.

The current study fits into a paradigm that relies on the use of the Rapid Prosody Transcription (RPT) method. The RPT task provides a simple way to register listeners’ perceptions of prosodic boundaries and prominence. A series of RPT experiments have been carried out with native speakers of English, investigating how prosodic structure is perceived in spoken English [1-3], [5], and [6]. For example, [6] found that besides silent pauses, phrase final vowel duration, and to a lesser extent vowel intensity, also play an important role in boundary perception. [3] showed that syntactic clause boundaries are the strongest predictors for boundary perception in American English. Various acoustic features, such as normalized vowel energy and maximum F0 of the stressed vowel were reported to be cueing prosodic prominence in spontaneous English [7], [8].

There are also some cross-linguistic studies using the RPT paradigm. For example, [12] investigated how North-American speakers perceive prosodic information in the English of Latin American Spanish speakers. An indirect comparison of Japanese learners’ and native English speakers’ perception of prominence and boundaries is available in [4]. The current research also compares Japanese and native listeners of English; however, it has an improvement over the latter paper, as the comparisons are based on responses drawn from the same stimuli set. Namely, our experiments are re-runs of a native RPT experiment [1], using Japanese EFL learners as participants.

Based on previous research, the following hypotheses can be formulated for non-native speakers of English. First, we expect that Japanese EFL speakers achieve lower levels of inter-speaker agreement than native speakers, because of the diversity of learners’ perceptual strategies. Second, since pause is a universally available acoustic cue, we expect that Japanese listeners rely on pauses during boundary detection to an extent that is comparable to native speakers. Third, as Japanese is a pitch-accent language, Japanese EFL speakers are expected to be more sensitive to pitch than to other acoustic cues in prominence perception.

2. Methodology
Our understanding of prosodic structure is based on temporally aligned labels for boundaries and prominence (cf. ToBI [13]). In a series of experiments Jennifer Cole and her colleagues (e.g., [1], [3], [5]) demonstrated that not only trained transcribers, but also native listeners can provide consistent and valuable information about how prosodic boundaries and prominence are perceived. Since the task utilized in their RPT experiments does not require any prior training, it was selected as an appropriate device for our comparative study.
The native English data for the comparisons were generously offered by Cole and her collaborators, together with a set of audio files that were used as stimuli in their experiments. The current study used this set of stimuli in a series of RPT tests to collect responses from Japanese EFL learners.

2.1. Participants

The participants of the experiment were recruited from undergraduate students of Kobe University and Kobe College. In total, 108 students participated in the experiment. The actual number of response sets used was smaller, as those individuals who failed to provide a response for any one of the audio recordings were entirely removed as participant from the analysis. The English language proficiency of the participants was intermediate, ranging between 404 and 608 points on the TOEFL PBT scale ($M = 493.7$, $SD = 37.33$; note: for Kobe College students the scores were transformed from TOEIC scores).

2.2. Data and Procedure

In order to ensure direct comparison, the experiments with Japanese EFL learners used audio recordings that had been used previously in the monolingual RPT experiment of [1]. The audio stimuli of the original experiment consisted of 36 pieces of 10–24 second long recordings taken from the Buckeye corpus of spontaneous speech [14]. A relatively short and easily comprehensible set of 11 sound files were selected from the original set to be presented to non-native speakers. A printed transcript for each sentence was also created, leaving out punctuation and capitalization.

The procedure of the experiments followed that of its monolingual counterpart. The experiments consisted of two blocks. In the prominence block, the participants were asked to circle those words in their transcripts that they perceived as prominent while listening to the audio stimuli. In the boundary block, they were instructed to use vertical lines to demarcate words that form a perceptual chunk. The audio files were played twice through classroom loudspeakers, during which time participants could mark and revise their responses.

3. Results

The printed data was digitalized and tested for inter-speaker agreement, and analyzed for correlation with various acoustic cues.

3.1. Inter-speaker agreement

In order to calculate inter-speaker agreement, those participants who failed to provide response for any one of the eleven audio stimuli were removed from the data set. This clean-up was carried out for boundary and prominence conditions separately, leaving 103 participants for boundary and 100 participants for prominence responses. Based on this data, inter-speaker agreement among Japanese EFL participants was found to be higher in the boundary task ($Fleiss' \kappa=0.475$, $n=103$) than in the prominence task ($Fleiss' \kappa=0.272$, $n=100$), which is in accordance with previous studies with native speakers [1], [2], [3].

It was not possible to calculate Fleiss' kappa scores over the whole data set for native speakers because the identity of speakers were not consistent across the stimuli. The number of raters ranged between 16 to 24 as a function of audio stimuli. In order to provide a direct comparison with Japanese EFL speakers, Fleiss' Kappa coefficients were calculated for each audio stimuli separately. The results confirmed the indirect findings of [4] in that inter-speaker agreements were higher in native speakers in both boundary and prominence conditions.

![Figure 1: L1 versus L2 Fleiss' Kappa scores](image1)

As another measure for inter-speaker agreement, Cohen's Kappa coefficients were calculated for all possible Japanese EFL rater pairs. Similar to Fleiss' kappa scores, agreement about boundaries were on average higher ($max = 0.827$, mean = 0.459, $SD = 0.149$, $n = 5253$) than for prominence marks ($max = 0.574$, mean = 0.275, $SD = 0.099$, $n = 4950$). The distribution of the agreement scores were unimodal, close to normal.

![Figure 2: Distribution of Cohen's Kappa scores for L2 listeners](image2)

3.2. P-scores and B-scores

Probabilities of receiving a prominence or boundary mark is considered to be the basis for further statistic analysis and comparison with previous research. These $p$(rominence)- and b(oundary)-scores are calculated by dividing the number of prominence and boundary marks by the number of valid participants. A score of 1.0 means that all participant marked the answer sheet; a score of 0.0 means that none did.

A pairwise comparison of boundary scores reveals that there is a strong correlation between native and non-native marking rates ($r = 0.876$, $p < 0.01$). Also, as it is clear from Figure 3, native speakers tended to achieve higher b-scores in general than non-native speakers. Nevertheless, there are some cases that registered with only Japanese but not native English listeners.

![Figure 3: Distribution of PScores for L2 Participants](image3)
In the prominence scores there is also a strong correlation between native and Japanese EFL speakers, albeit to a lesser extent ($r = 0.749$, $p < 0.01$).

Although the correlations were strong, Wilcoxon signed tests showed that neither boundary ($Z = 333, p < 0.001$) nor prominence scores ($Z = 66993, p < 0.001$) were identical between the native and non-native speakers.

These findings imply that both groups of listeners behave by-and-large similarly, but there are still differences in their perceptions of boundary and prominence.

### 3.3. Acoustic cues

The audio recordings were transcribed at word and segmental levels using Praat [15]. Various acoustic features were extracted from the audio data in order to investigate their correlations with boundary and prominence scores.

#### 3.3.1. Acoustic Cues for Boundaries Perception

The most obvious, cross-linguistic acoustic correlate of boundaries are silent pauses (e.g., [2], [16]) and fillers (e.g., “er”, “um”). As expected, $b$-scores were found to be correlated with non-word intervals at a statistically significant level both with native and with Japanese EFL speakers. Interestingly, however, the extent of correlation was greater in native speakers ($Kendall's \tau = 0.615, p < 0.001$) than in Japanese speakers of English ($\tau = 0.522, p < 0.001$). This asymmetry was present even when non-word intervals were categorized into silent pauses (NS $\tau = 0.525$, JP EFL $\tau = 0.484$) and vocalic fillers (NS $\tau = 0.401$, JP EFL $\tau = 0.335$).

Another acoustic predictor that has been reported to show correlation with perceived boundaries is final vowel duration [2]. This finding was confirmed in our data, as significant correlation was found between $b$-scores and the duration of final vowel in words. As with pauses, the degree of correlation was higher for native speakers than for Japanese EFL learners (NS $\tau = 0.324$, JP EFL $\tau = 0.274$). This result, however, has to be considered carefully, because final vowel duration naturally correlates with the duration of a following pause (in our data: $\tau = 0.297, p < 0.001$). Since $b$-scores showed higher correlation with pauses than with vowel duration, we can conclude that pausing takes precedence over vowel duration as a perceptual cue for prosodic boundaries.

#### 3.3.2. Acoustic cues for prominence perception

Amplitude, pitch and segment duration are the most widely acknowledged acoustic correlates of prominence. The following table summarizes the correlation of some of the most relevant vowel-related acoustic cues with prominence scores. All acoustic measurements were normalized at stimuli level.

<table>
<thead>
<tr>
<th>feature</th>
<th>NS</th>
<th>JP EFL</th>
</tr>
</thead>
<tbody>
<tr>
<td>vowel duration</td>
<td>0.318</td>
<td>0.346</td>
</tr>
<tr>
<td>max pitch</td>
<td>0.126</td>
<td>0.065</td>
</tr>
<tr>
<td>max intensity</td>
<td>0.123</td>
<td>0.147</td>
</tr>
<tr>
<td>mid-point intensity</td>
<td>0.102</td>
<td>0.126</td>
</tr>
<tr>
<td>mid-point pitch</td>
<td>0.096</td>
<td>0.024</td>
</tr>
</tbody>
</table>

As previous RPT studies of [7], [8] and [4] pointed out, unlike with boundaries, there is no single acoustic predictor for prominence. Surprisingly, the correlation coefficients reveal rather similar patterns between native and Japanese EFL listeners. Among the observed features, vowel duration proved to be the strongest cue (NS: $\tau = 0.318$, JP EFL $\tau = 0.346$). Contrary to the expectations, pitch-related cues did not dominate the responses of Japanese EFL listeners.

### 4. Discussion

#### 4.1. Overview

Results on measurements of inter-speaker agreement repeated results of previous studies and fulfilled our expectations. First, the level of agreement on prosodic boundaries was higher than that on prominent words for both native and Japanese EFL learning groups. Second, non-native agreement scores for both boundary and prominence were lower than for their native counterparts. This lower agreement could be due to the diversity of perceptual strategies language learners adopt. This argument is supported by the fact that Japanese learners of English normally do not receive any unified training concerning pronunciation skills.

The inspection of boundary and prominence scores revealed surprising similarities between native and Japanese speakers of English. The correlation was higher with boundary scores, which implies, in accordance with [4], that prosodic boundaries are possibly encoded by language independent cues, such as silent pause or hesitation. Indeed, these non-word intervals were found to be similarly important for both language groups in prosodic boundary perception. This
verifying our second initial hypothesis. Interestingly, native speakers showed greater reliance, in absolute values, to these universal acoustic cues.

As pointed out in previous research, prosodic prominence is not encoded clearly by a single acoustic feature (see [6], [7], [8], and [4]). The best available acoustic cue for prominence in the current research was found to be vowel duration, but its correlation with p-scores was significantly lower ($\tau_u = 0.346$) than the best acoustic cue's correlation with b-scores ($\tau_u = 0.522$). This result refutes the initial hypothesis that Japanese EFL speakers are sensitive to pitch in their prominence perception.

The lower levels of correlation with acoustic cues is related to rather low levels of inter-speaker agreement in prominence perception, especially in non-native speakers (Fleiss' kappa $= 0.272$). The fact that listeners do not agree about the position of prominence implies that they employ different perceptual strategies, which rules out the possibility of prominence having strong correlation with a single acoustic cue. The hypothesis that assumes that native speakers utilize various prominence encoding strategies can also explain why non-native speakers are less successful in achieving native-like performance in prominence perception.

### 4.2. Future directions

The next logical step for this research would be to include syntactic cues in the analysis. The research of [3] indicated that syntactic information greatly contributed to boundary perception in native speakers. For Japanese EFL learners, however, this effect would be expected to be smaller, because their L2 syntax processing is less automatized.

Another logical step would be to investigate the effect of English proficiency level on perceptual strategies. This step would require recruiting English learners with a greater diversity in proficiency level, as the participants in this study were rather homogeneous in level.

A less obvious extension of the research would be to cluster participants by their Cohen's kappa agreement scores, and find differences over perceptual strategies of the clusters. This bottom-up approach would be expected to highlight differences in perceptual strategies both within and across language groups. This bottom-up clustering seems to be a more promising approach than the standard proficiency-based approach since proficiency measured by such tests as TOEIC or TOEFL does not necessarily imply spoken language fluency, or native-like perception of prosodic structures.

### 5. Conclusions

This study was a replication of the RPT experiments of [1], this time with Japanese EFL learners as participants. The use of identical stimuli sets made direct comparison of native and Japanese EFL speaker responses possible. The direct comparison of L1 and L2 perceptual data found unexpected similarities in the prosody perception strategies of the two language groups. Although the experiment results for native and non-native listeners were found to be statistically different, boundary and prominence scores showed higher correlation than expected. The reliance on acoustic cues in both cases was also similar. Boundary perception was largely encoded by inter-word pauses. Acoustic cues for prominence were less consistent, but vowel duration was found to be a relatively good predictor of prominence for both native and Japanese speakers of English.

### 6. Acknowledgements

We would like to express our deepest gratitude to professor Jennifer Cole for her unending help and expert advice throughout the project. This work was supported by The Ministry of Education, Culture, Sports, Science and Technology KAKENHI (24520542) assigned to the second author.

### 7. References


