How L2 learners perceive English prosody

Shinobu Mizuguchi[1], Tim Mahrt[2], and Koichi Tateishi[3]

[1]Kobe University, Japan, [2]WOVN Technologies, Inc., [3]Kobe College, Japan
[1]*mizuguti@kobe-u.ac.jp*,[2]*tmahrt@gmail.com*, [3]*tateishi@mail.kobe-c.ac.jp*

Abstract

It is important for listeners to perceive boundaries and prominences of an utterance correctly for good communication. L2 learners have more difficulties in speech perception than L1 listeners. This paper investigates how Japanese learners of English as a foreign language (JEFLs) perceive English prosody. We have conducted Rapid Prosody Transcription (RPT) experiments on spontaneous speech and read texts with JEFLs and L1 listeners and compared how they perceive boundaries and prominences. We also conducted a perception experiment on narrow focus in English to consider how JEFLs process acoustic cues.

Our findings are that JEFLs can perceive English narrow focus with almost perfect accuracy. JEFLs are also able to perceive prosodic boundaries and prominences in read text well, or even better than L1 listeners. Both L1 listeners and JEFLs have similar perceptual strategies toward prosody processing; they use pause for a boundary cue and duration for a prominence cue. A difference is found in spontaneous speech; L1 listeners use acoustic, syntactic (e.g. S and SBAR) and non-syntactic cues (e.g. Discourse Markers and Disfluencies), but JEFLs cannot handle non-syntactic cues. With these non-syntactic 'distractors', JEFLs are likely to be too confused to perceive boundaries correctly.

Keywords: L2 learners, English prosody, perception of boundary and prominence, Rapid Prosody Transcription (RPT)

1. Introduction

It is well-known that prosody plays an important role in communicating the syntactic, semantic, and pragmatic information of spoken language (cf. Bolinger [1], Ladd [2], among many others). Previous studies in Japan (cf. Saito and Ueda [3], among others) have found that Japanese learners of English as a Foreign Language (JEFLs) misplace prominence, and produce English with a narrower F0 range and less intensity than L1 speakers. Prosody perception by JEFLs is, on the other hand, understudied. This paper investigates how JEFLs perceive English prosody in natural speech. For this study, we need to analyze prosodically annotated speech drawn from a variety of speech materials. Cole developed Rapid Prosody Transcription (RPT) with her colleagues. It is a system for collecting prosodic annotations from untrained listeners (cf. Cole et al. [4], among others); listeners mark prominences with a circle and boundaries with a vertical line on transcribed speech as in (1) while listening to recorded speech.

(1) i really don't know i think in today's world what they call the nineties that uh it's just like everything is changed like when i grew up …

Cole et al. [4] conducted a series of RPT experiments on English spontaneous speech; they used 72 excerpts of 11-22 seconds long in duration drawn from the Buckeye corpus, a dataset of conversation-style speech collected from the interviews of 97 speakers from Columbus, OH. Under RPT, each word is assigned a p-score and a b-score, which is the sum of individuals who marked the word for bearing a prominence or boundary respectively, divided by the number of participants. Cole and her colleagues recruited 36 native speakers of English (NS/L1) and considered inter- and intra-speaker variations in p-scores and b-scores. They found that (i) b-scores (0.63 on Fleiss' kappa) are higher than p-scores (0.40 on Fleiss' kappa) in inter-speaker agreement and (ii) the best predictors of boundary perception of English prosody are the syntactic categories S & SBAR (0.541 on Kendall's tau), and the second-best predictor is vowel duration (0.369 on Kendall's tau) (cf. Figure 1[1] and Figure 2).

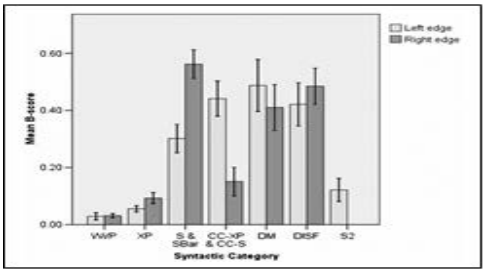


Figure 1: Mean b-scores at syntactic and non-syntactic categories at left-edge and right-edge (adapted from [4])

10.21437/ISAPh.2018-17

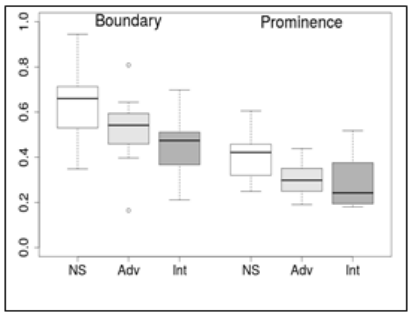


Figure 2: Inter-speaker agreement
(Unit: Fleiss kappa, adapted from [5])

We [5] replicated Cole et al.'s RPT experiment on two groups of JEFLs[2] (108 Intermediate learners (Int, TOEFL PBT mean score: 493.7) and 15 Advanced learners (Adv, TOEFL PBT mean score: 595)), with 11 excerpts from the Buckeye corpus. Figure 2 illustrates the inter-annotator variation of the boundary and prominence perception on Fleiss' kappa; boundary scores are 0.63 (L1) > 0.521 (Adv) > 0.458 (Int), and prominence scores are 0.401 (L1) > 0.305 (Adv) > 0.284 (Int). We can see that we have a task effect: the inter-annotator agreement is higher for making boundaries than for prominences. We also have a group effect; L1 listeners perceive prominences and boundaries more than JEFLs.

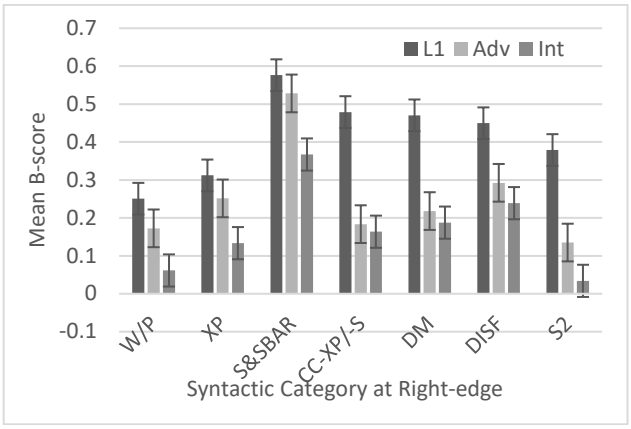
To investigate the causes of the group effect further, we compared how syntactic and non-syntactic categories function as boundary cues. Figure 3 shows that the syntactic categories of S & SBAR and CC-XP/-S at the right-edge are boundary cues for L1, but not for JEFLS.

We have conducted a one-way ANOVA on the b-scores of S & SBAR between L1 and JEFLs, and the difference is significant between L1 and Int ($F(1,22) = 4.664$, $p=0.042$), but not between L1 and Adv ($F(1,22) = 0.049$, $p=0.4408$). As for CC-XP/-S, the difference is significant between L1 and JEFLs ($F(1,22) = 4.860$, $p=0.0382$ between L1 and Int, $F(1,22) = 4.419$, $p=0.0472$ between L1 and Adv).

As for non-syntactic categories, Discourse Markers (DMs) and Disfluencies (DISFs) are boundary cues for L1, but not for JEFLs ($F(1,22) = 5.518$, $p=0.0282$ between L1 and Adv, and $F(1,22)=5.433$, $p=0.0293$ between L1 and Int, on one-way ANOVA).

We have a different result from Cole et al. [4] as for acoustic cues; post-word pause is a stronger cue for boundary perception than vowel duration with our materials[3] (cf. Table 1).

Cole et al. [4] claim that the best predictors of boundary perception of English prosody are the syntactic categories S & SBAR, and the second-best predictor is vowel duration for L1 listeners. Our study



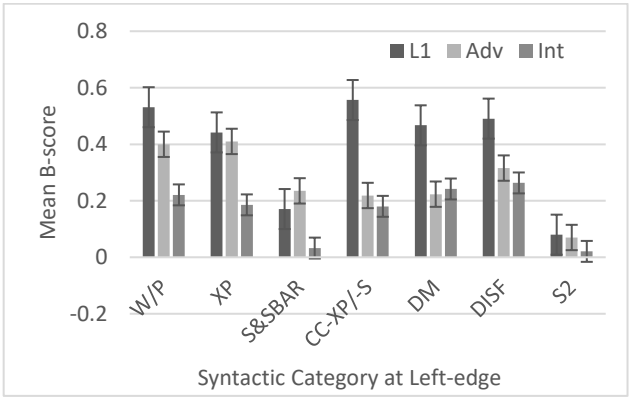


Figure 3: Mean b-scores per category at right-edge (above) and at left-edge (below)

Table 1: Acoustic cues (Unit: Kendall's tau)

		L1	ADV	INT
boundary	Silent pause	0.615	0.550	0.522
	Final vowel duration	0.384	0.3	0.318
prominence	Stressed vowel duration	0.280	0.264	0.312
	Max pitch	0.112	0.049	0.058

[5] shows that JEFLs are different from L1 listeners in the use of syntactic, non-syntactic and acoustic cues. The purpose of this paper is to investigate how JEFLs perceive English prosody in different kinds of speech. Since we [5] did a RPT experiment on English natural speech which is endowed with syntactic, non-syntactic and acoustic cues, our next step is to conduct perception experiments on English natural speech other than spontaneous speech. Thus, we next conducted an RPT experiment with read news stories which contain syntactic cues and acoustic cues, but do not contain Discourse Markers or disfluencies. We also conducted a perception experiment with contrastive narrow foci that did not contain meaningful syntactic cues or disfluencies. We will compare the results with the ones in Cole et al. [4] and Mizuguchi et al.[5], and see whether or not there is a difference between L1 listeners and JEFLs

in their perceptual strategies depending on the materials.

2. Perception experiment on read text

2.1. Methods and participants

We conducted a perception experiment on read texts in English, i.e. natural speech without non-syntactic boundary cues of DMs and DISFs. We used RPT, as in Cole et al. [4] and Mizuguchi et al. [5]. The materials are 11 excerpts of 11 to 22 seconds long in duration from the news stories of *the Voice of America*, as in (2).

(2) each year nearly a thousand young people come through community centers run by philadelphia based non profit jevs …

We recruited 30 L1 listeners, 17 Adv JEFLs (F9, M8, mean age: 22.6, TOEFL PBT mean score: 580.6) and 30 Int JEFLs (F9, M21, mean age: 18.9, TOEFL PBT mean score: 474.1). L1 listeners sat for a WEB-based experiment, developed by one of our authors (cf. Mahrt [7]). JEFLs were recruited at a Japanese university and they did a paper-based experiment. All of our participants reported no auditory difficulties, and the same instructions were given to all the three groups. Our read texts contain no DMs or DISFs. We therefore predicted a smaller difference in perception of read texts between L1 listeners and JEFLs.

2.2. Results

Table 2 shows the results. To our surprise, JEFLs show a stronger inter-listener agreement than L1 listeners in b-score. Advanced JEFLs show stronger inter-listener agreement than L1 listeners in p-score, too. Also all the three groups use pause as a boundary cue and duration as a primary prominence cue, followed by pause.

Figure 4 illustrates mean b-scores per syntactic category, and we can see that there is no significant difference among the three groups of listeners in how syntactic categories are used as boundary cues.

Table 2: Inter- and intra-speaker agreement

agreement			LI	ADV	INT
Inter-speaker (Unit: Fleiss' kappa)		b-score	0.6	0.81	0.66
		p-score	0.39	0.42	0.31
Intra-speaker (Unit: Kendall's tau)	boundary	pause	0.72	0.72	0.61
		duration	0.34	0.34	0.35
	prominence	pause	0.29	0.29	0.27
		duration	0.46	0.48	0.49

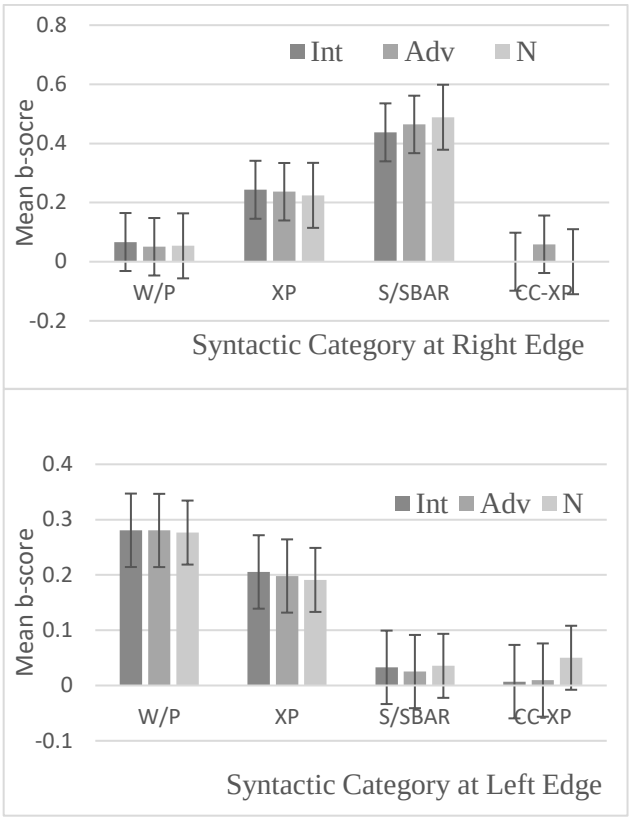


Figure 4. Mean b-scores per syntactic category at right edge (above) and left edge (below), where N is L1 listeners

2.3. Discussion

The RPT experiments on English spontaneous speech show a group effect and a task effect (cf. Figure 2), but Table 2 shows that there is no group effect observed in b-scores and p-scores. What is interesting is the fact that L1 listeners have no difference in inter-speaker agreement on b-scores and p-scores between spontaneous speech and read texts (cf. Figure 2 and Table 2). JEFLs, on the other hand, show a higher inter-speaker agreement in b-scores and p-scores on read texts than on spontaneous speech. This experiment has shown that for materials without DMs and DISFs, JEFLs are more easily able to perceive boundaries and prominences than for materials with DMs and DISFs.

Figure 4 shows that S & SBAR at the right edge are boundary perception cues for all the groups; even Int JEFLs use S & SBAR as a syntactic boundary cue. No significant difference is observed among the three groups in the use of the major categories as well as minor categories like XPs and W/P at the right edge. When learners are deprived from the non-syntactic boundary cues of DMs and DISFs, JEFLs can perceive boundaries as well as L1 listeners. We can claim that our prediction of a more narrow difference in perception between JEFLs and L1 listeners is borne out.

As for acoustic cues, Cole et al. [4] considered vowel duration and pause as boundary cues, but they did not consider prominence cues. We have found that all the three groups use pause as a boundary cue and duration as a primary prominence cue, followed by pause (cf. Table 2).

3. Perception experiment on narrow focus

Our second experiment is a perception experiment on narrow focus in English. This experiment keeps the syntactic context constant, and thus listeners can only rely on acoustic cues to perform the task, in contrast to our earlier experiments. Focus is the information in the sentence that is assumed by the speaker not to be shared by the speaker and the hearer (cf. Jackendoff [8]). The literature (cf. Halliday [9] and many others henceforth) classifies focus into broad focus and narrow focus. The former, for instance, can signal the foci *the shed, painted the shed,* or the whole sentence in (3), depending on the context. The latter narrows the possible range of foci to a particular constituent such as only *the shed*.

(3) John painted *the shed* yesterday. (Halliday [9])

Languages vary in how narrow focus is acoustically marked, and Lee et al. [10] conducted a cross-linguistic perception experiment, using ten digit numbers of the form XXX-XXX-XXXX, where one number in the series was produced with focus under a Question-Answer sequence as in (4). Given only the response portion, participants were asked to identify which number was focused.

(4) A: Is Mary's number 215-418-5623?
B: No, the number is 215-417-5623.

They found that for languages with stronger acoustic marking of narrow focus, such as with longer duration, higher intensity and higher F0, listeners were able to identify the focused material with almost perfect accuracy. The perception identification rate of American English is 97.8% by L1 listeners.

Table 3: Median z-score values of focused digits (adapted from Lee et al. [10])

	South Kyungsang Korean	Seoul Korean	Tokyo Japanese	Suzhou Wu	Standard French	Mandarin Chinese	American English
Production	median z-score values of focused digits						
duration	0.64	0.13	0.10	0.48	1.73	1.19	0.95
intensity	-0.26	0.24	-0.24	0.53	0.97	0.36	1.28
pitch	1.00	0.62	0.60	0.61	1.17	3.13	2.96
Perception identification rate	55.6%	44.6%	-	-	-	94.9%	97.3%

Weak — Prosodic Focus Marking — Strong

3.1. Procedures and participants

We replicated Lee et al.'s experiment on English with JEFLs. We recruited 18 intermediate and advanced JEFLS (F5, M13, mean age 20.5) at two Japanese universities. The materials are 30 randomized phone-numbers embedded in a carrier sentence as in (4), recorded by a male speaker of Midwest American English.

3.2. Results and discussion

The perception identification rate is 98.6%. English narrow focus is coded by the prosodic cues of F0, duration and intensity, and L1 listeners as well as JEFLs perceive narrow focus correctly. This fact leads us to claim that although JEFLs are poorer at perceiving boundaries and prominences (cf. Figure 2), it is not due to an inability not to perceive acoustic cues.

4. General discussion and conclusion

We have compared JEFLs with L1 listeners in how they perceive English prosody in perception experiments. The materials we used were spontaneous speech, read news stories, and contrastive narrow focus. Table 4 summarizes the results of RPT and narrow focus perception experiments considered in this paper. We divided the perception cues in English prosody among three categories: syntactic, non-syntactic and acoustic. Spontaneous speech has all the three types of cues, read news stories has two types, and narrow focus is endowed with acoustic cues only.

Table 4: Perception cues and agreement rate

	Perception Cues			Boundary: Inter-speaker Agreement (Unit: Fleiss' kappa)			Prominence: Inter-speaker Agreement (Unit: Fleiss' kappa)		
	syn-tactic	non-syntactic	acous-tic	LI	Adv	Int	LI	Adv	Int
spontaneous speech	✔	✔	✔	0.63	0.52	0.49	0.40	0.31	0.28
read news story	✔	*	✔	0.6	0.81	0.66	0.39	0.42	0.31
							Perception Identification Rate (%)		
narrow focus	*	*	✔	-			97.3%	98.6%	

The purpose of this paper is to investigate how JEFLs perceive English prosody and to compare their strategies with the ones used by L1 listeners. To our surprise, JEFLs show similar perception strategies as L1 listeners. JEFLs can perceive narrow focus with as high accuracy as L1 listeners, which means they can process acoustic cues correctly. Also JEFLs are able to perceive prosodic boundaries and prominences in read text well, or even better than L1 listeners. Table

2 leads us to claim that JEFLs use pause as a boundary cue and duration as a prominence cue almost to the same extent as L1 listeners. Figure 4 shows that JEFLs have a good command of syntactic categories as boundary cues. A difference between L1 listeners and JEFLs is found in the perception of spontaneous speech (cf. Figure 2). What is interesting is that L1 listeners show no difference in boundary and prominence perception between the materials with and without non-syntactic cues. JEFLs, however, have big differences in boundary and prominence perception in the same environments. This implies that the non-syntactic cues of DMs and DISFs disturb JEFLs in perceiving English prosody, but not L1 listeners. With these non-syntactic 'distractors', JEFLs are likely to be too confused to perceive boundaries correctly.

Also Intermediate students have, in fact, difficulties to use S & S BAR at the right edge as a boundary cue. As for processing S & S BAR, there is no difference between L1 and Adv JEFLs, but there is between L1 and Int JEFLs, as observed in Section I. This implies that poorer syntactic abilities lead to poorer perception of boundaries and hence poorer perception of prominence. Further study on the correlation between syntactic and acoustic cues is expected.

What remains unexplained is a task effect; why do even L1 listeners show a low inter-speaker agreement in perception of prominence in spontaneous speech (cf. Figure 2)? Both L1 and L2 listeners clearly understand some prosodic phenomenon (like contrastive narrow focus) but more subtle ones may be difficult. We need to explain why. One thing we can think of is that narrow focus is coded by prosodic cues (cf. Table 2), while broad focus is less saliently coded prosodically (cf. Table 1). This is one of the reasons why prominence in spontaneous speech is harder to perceive. Bishop [11] argues that narrowly focused objects were heard more prominent than when the same objects were part of broader VP or sentence foci. Though much needs to be clarified, this may be a possible account. Another possible account is that spontaneous speech processing is more complex, since it requires boundary information as well as prominence information to be processed. We already know that we need to consider syntactic and non-syntactic cues for processing boundaries. Broad focus perception, therefore, demands more information in addition to acoustic information, which makes processing complex.

It is important for listeners to perceive the boundaries and prominences of an utterance correctly for good communication. Studies on the segmental production by JEFLs are being developed, but prosody perception is still understudied. L2 learners have more difficulties in speech perception than L1 listeners, but we have shown that L1 listeners and L2 learners have similar perceptual strategies toward prosody processing. We hope our study will be extended to a deeper understanding of human cognition.

5. Acknowledgements

We would like to express our gratitude to the experiment participants at Kobe University and Kobe College. We also thank Prof. Jennifer Cole for letting us share her RPT materials and data. Thanks also go to the Voice of America for letting us use their materials for our experiments. This study is supported by the Japan Society for Promotion of Science (JSPS) grants #245042 and #15K02480, given to Shinobu Mizuguchi.

6. References

[1] Bolinger, D. 1986. *Intonation and Its Use: Melody and Spoken English*, Stanford: Stanford University Press.

[2] Ladd, D.R. 1996. *Intonational Phonology*. Cambridge: Cambridge University Press.

[3] Saito, H., Ueda,I. 2011. Misplacement of nuclear stress by Japanese learners of English. *Onsei Kenkyu*: *Journal of the Phonetic Society of Japan* 15(1), 87-95. (in Japanese)

[4] Cole, J., Mo, Y., Hasegawa-Johnson, M. 2010. Signal-based and expectation-based factors in the perception of prosodic prominence. *Language and Cognitive Processes* 1, 425-452. DOI: https://doi.org/10.1515/labphon.2010.022.

[5] Mizuguchi, S., Pintér, G., Tateishi, K. 2016. Natural speech perception cues by Japanese learners of English. *Proceedings of PacSLRF 2016*, 151-156.

[6] Roy, J., Cole. J., Mahrt, T. 2017. Individual differences and patterns of convergence in prosody perception. *Laboratory Phonology: Journal of the Assoiation for Labolatory Phonology* 8(1), 22. DOI: http://doi.org/10.5334/labphon.108.

[7] Mahrt, T., 2015. Language markup and experimental design software (LMEDS), Retrieved from http://prosody.beckman.illinois.edu/lmeds.html.

[8] Jackendoff, J. 1972. *Semantic Interpretation in Generative Grammar*. Cambridge, Mass.: The MIT Press.

[9] Halliday, M.A.K.,1966-1968. Notes on transitivity and theme in English Part 1 – Part 3. *Journal of Linguistics* 2 and 3.

[10] Lee, B. Y-Cheol, Wang, S., Chen,M., Adda-Decker, A., Nambu, S., Liberman, M. 2015. A crosslinguistic study of prosodic focus. *Proceedings of ICASSP 2015 and 2015 IEEE International Conference*, 4754-4758.

[11] Bishop, J. B. 2011. English listeners' knowledge of the broad versus narrow focus contrast. *International Society of Phonetic Sciences (ICPhS)* XVII, 312-315.

[1] In addition to the coding of b-scores and p-scores, each word is annotated for the highest level syntactic boundary that coincides with its right and left edges, separately, based on a manual syntactic parse using the Penn Treebank annotation guidelines (Marcus, Marcinkiewicz and Santorini 1993). A left-edge marks a boundary preceding the word, and a right-edge marks a boundary marking after the word. (i), for example, shows how S is marked at both edges.

(i) Left edge: uh |$_S$ it's just you …
Right edge: still live in the city of Columbus $_S$| uh

The syntactic phrasings are S, SBAR, S2, Coordinate Conjunction (CC-S/-XP), Phrase (XP), Within Phrase (W/P), Disfluency (DISF), and Discourse Marker (DM). See Cole et al. 2010 [4] for details of the other syntactic phrasings used.

[2] The number of transcribers vary among the three groups. Based on an analysis of inter-annotator agreement proposed by Roy et al. 2017 [6], we consider RPT annotations from a group of 13 annotators to be reliable in the sense that they are reproducible with an expected difference of less than 5% of the estimated SD of the true population.

[3] Though we used the same excerpts with the ones used in Cole et al. 2010, our materials are of smaller portion than theirs, which might lead to a different result.