Focus perception in Japanese: Effects of focus location and accent condition

Albert Lee, Faith Chiu, and Yi Xu

Citation: Proc. Mtgs. Acoust. **29**, 060007 (2016); doi: 10.1121/2.0000441
View online: https://doi.org/10.1121/2.0000441
View Table of Contents: https://asa.scitation.org/toc/pma/29/1
Published by the Acoustical Society of America

Proceedings of Meetings on Acoustics

Volume 29 http://acousticalsociety.org/

172nd Meeting of the Acoustical Society of America

Honolulu, Hawaii
28 November – 2 December 2016

Speech Communication: Paper 5aSC37

Focus perception in Japanese: Effects of focus location and accent condition

Albert Lee
Department of Linguistics and Modern Languages, Chinese University of Hong Kong; Department of Linguistics, University of Hong Kong, Hong Kong; albertlee@hku.hk

Faith Chiu and Yi Xu
Department of Speech, Hearing and Phonetic Sciences, University College London, London, UK; faith.chiu.11@ucl.ac.uk, yi.xu@ucl.ac.uk

This study explores the contexts in which native Japanese listeners have difficulty identifying prosodic focus. Theories of intonational phonology, syntax, and phonetics make different predictions as to which focus location would be the most challenging to the native listener. Lexical pitch accent further complicates this picture. In a sentence with mixed pitch accent conditions (e.g. Unaccented-Accented-Unaccented), the lexical accent would naturally stand out as more prominent than the unaccented words in terms of modifications to the F0 contour, thus potentially resembling focus. A focus identification task was conducted with 16 native listeners from the Greater Tokyo area. Natural and synthetic stimuli were played to the listeners who then chose which word of the sentence was under focus. Neutral (or broad) focus was also an option. Stimuli contrasted in accent condition and focus location. Results showed a highly complex interplay between these two factors. For example, accented narrow foci were always more correctly identified (51%) than unaccented ones (28%), whereas the identification rate for final focus was the lowest (31%) among all focus locations. These results are discussed with reference to the research literature on focus production and formal representation of intonation.

Published by the Acoustical Society of America

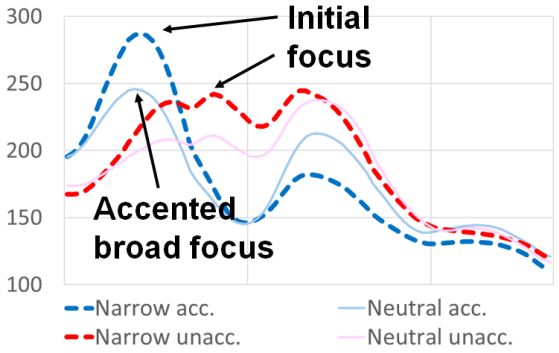
 [DOI: 10.1121/2.0000441]

1. INTRODUCTION

By now a great deal is known about the production of Japanese focus prosody. Specifically, narrow focus in Japanese is characterized by on-focus fundamental frequency (f_o) range expansion and post-focus f_o range compression (Beckman & Pierrehumbert, 1986; Ishihara, 2011; Lee & Xu, 2012; Pierrehumbert & Beckman, 1988) alongside the modification of non-f_o cues such as duration and formant frequency (Maekawa, 1997). However, since narrow focus can also be marked by syntactic (i.e. fronting the syntactic constituent under narrow focus, aka 'scrambling') and morphological (i.e. using the focus particles *dake* 'only' or *mo* 'too') means, it remains an open question how much weight prosody alone carries in signaling narrow focus in the discourse. Whereas focus markers can be easily elicited in a *production* task where speakers are told to express narrow focus without using morpho-syntactic means, *identifying* focus could be a daunting task when there are only phonetic cues in stimuli lacking explicit morpho-syntactic markers. Focus perception could be even harder if only f_o cues are available, because f_o is delegated a wide range of communicative functions (e.g. focus, emotion, sentence type); a given prosodic marker (e.g. a raised f_o peak) could be associated with numerous different communicative meanings.

Some phonological, syntactic, and phonetic factors further complicate the issue. As summarized in Venditti et al (2008), initial, penultimate, and final foci in Japanese are all potentially confusable with neutral (i.e. broad) focus for different reasons, respectively described below.

Figure 1. Averaged f_o contours of initial narrow vs. neutral × accented vs. unaccented focus (data from Lee & Xu, 2015). Each contour is averaged across 10 speakers (5 repetitions each).



A. Initial focus (phonological factor)

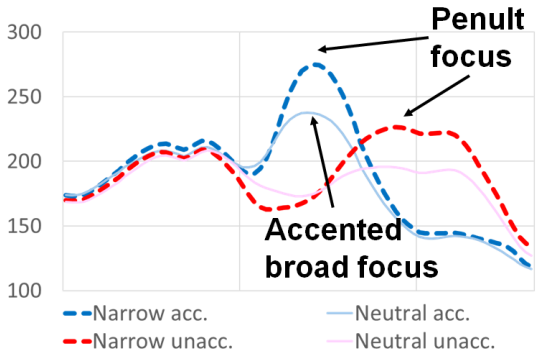
Venditti et al (2008:504) noted that initial focus and broad focus may be ambiguous because 'there has to be at least one IP-initial rise at the beginning of every well-formed utterance (in Japanese), and when there is no narrower focus prompting an IP break and reset later on, the rise from the utterance initial [%L] makes the immediately next [H] target (whether a phrasal [H–] or the [H] of a [H*+L]) the highest (most prominent) peak in the utterance'. In other words, where f_o is not 'reset' utterance-medially by a later narrow focus, the highest peak will be on the first word of the utterance. Meanwhile, when narrow focus is utterance-initial, on-focus expansion will raise the first peak, but will not change the fact that it is the highest in the first place. There

thus exists ambiguity as the listener cannot determine if the initial peak has been raised by focus or is intrinsically high (see Figure 1, blue contours).

B. Penultimate focus (syntactic factor)

Penultimate focus would be indistinguishable from neutral (more precisely, broad focus on the entire object-verb phrase) due to the 'focus projection' principle. Focus projection predicts that placing prosodic focus on the object NP leads to two possible interpretations: narrow focus on the NP and broader focus on the VP. It follows that for an SVO language, like English, final focus and broad focus on the VP would be ambiguous (Gussenhoven, 1983), whereas for a SOV language like Japanese broad focus on the VP would be indistinguishable from narrow focus on the object NP, i.e. penultimate focus (Ishihara, 2011; Ito, 2002; Lee & Xu, 2012). The same has also been observed in Turkish (Ipek, 2011), another SOV language. While in production (laboratory speech), the distinction between the two focus conditions is marked by on-focus raising (see Figure 2) post-focus compression appears to be absent (overlapping blue contours towards the end). Listeners thus have one cue less to rely on compared to initial focus, possibly making focus perception more difficult.

Figure 2. Averaged f_0 contours of penultimate narrow vs. neutral × accented vs. unaccented focus



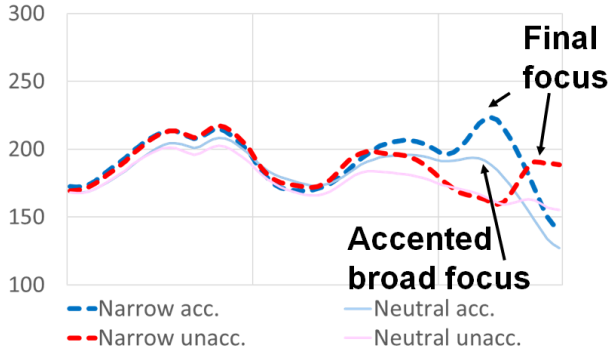
C. Final focus (phonetic factor)

Across languages it has been shown that final focus is prosodically expressed much less effectively than an earlier focus (Botinis, Bannert, & Tatham, 2000; Botinis, Fourakis, & Gawronska, 1999; Cooper, Eady, & Mueller, 1985; Rump & Collier, 1996; Xu, 1999). In general, an utterance-final word bearing narrow focus 'is produced with less relative emphasis' (Cooper et al., 1985, p. 2147 investigating focus in English). For SVO languages, part of the reason would be focus projection as discussed in §1.2. Meanwhile, Liu & Xu (2005) suggested that this could be the result of the conflicting needs to encode both sentence type (questions vs. statements) and focus in the sentence-final word. As Japanese also marks questions with an utterance-final boundary tone (Beckman & Pierrehumbert, 1986; Pierrehumbert & Beckman, 1988), an overladen utterance-final word would have less room for f_0 modification for focus. If acoustic cues in production are ambiguous in the first place, listeners would be easily confused in perception too.

Figure 3 shows that although there is clear evidence of on-focus raising that tells apart narrow vs. broad foci, the pre-focus portions of the f_0 contours are largely overlapping. How

sensitive listeners are to the f_0 difference in the final word alone would determine their ability in identifying narrow final focus.

Figure 3. Averaged f_0 contours of final narrow vs. neutral × accented vs. unaccented focus



D. The effect of accent and role of non-f_0 cues

A further perplexing fact is that Japanese has lexical pitch accent. Unlike lexical tones, of which members are deemed equal in prominence within a language (except, for example, the Neutral Tone in Mandarin which is the 'weaker' tone), an accented mora in Japanese naturally stands out among unaccented ones; in turn an accented word would stand out among unaccented words (see for example Lee, Prom-on, & Xu, in press; Pierrehumbert & Beckman, 1988 for the acoustical differences between accented and unaccented words). Thus in a broad focus utterance where an accented word is surrounded by unaccented words (i.e. Unaccented-Accented-Unaccented, henceforth UAU), the accented word stands out and could be misperceived as bearing narrow focus.

Finally, beside f_o, focus has been reported to affect duration (Maekawa, 1997; Xu, 1999), voice quality (Sluijter, van Heuven, & Pacilly, 1997) and formant frequency (Maekawa, 1997), all of which could serve as cues to focus perception. It is possible that without these cues, f_0 patterns associated with focus would not be very effective.

Given the intricate nature of the above-listed issues, the first goal of this paper is to compare if there is a focus location that is most indistinguishable from broad focus in declarative utterances in Japanese, and if so, whether it is initial, medial or final focus. Secondly, we want to find out how pitch accent may interact with focus location in affecting listeners' perception of focus. Thirdly, it will be interesting to know how well listeners can identify focus when f_0 is the only cue. A series of perception experiments were conducted to answer these questions, using both natural and synthesized stimuli, as described in the following sections.

2. PRE-TEST: NATURALNESS JUDGMENT[1]

Our goal is to test the effects of focus location and accent condition on focus identification using resynthesized stimuli, which are better controlled and free from cross-repetition variation. To do so, it is necessary to first establish that resynthesized and natural stimuli are not different

[1] A version of this Section appeared in Lee & Xu (2015).

for our purpose. In this experiment, we compare how natural the two types of stimuli sound to the native listeners.

All the stimuli used here and in §3 were adopted from Lee & Xu (2015). There are 128 utterances in total, made of 2 sentence lengths (8 or 11 morae) × 8 accented conditions (2 accent ^ 3 words) × 4 focus conditions (Initial / Medial / Final / Neutral) × 2 sources (natural / synthesized). The natural stimuli were resynthesized using PENTAtrainer2 (Prom-on & Xu, 2012; Xu & Prom-on, 2014). The training corpus consisted of 6,400 natural utterances, based on which articulatory parameters were extracted in terms of pitch height, slope and articulatory strength. These articulatory parameters were then used to generate the synthetic stimuli. See Lee & Xu (2015) for details of the procedures and accuracy of the f_o resynthesis. PENTAtrainer2 was chosen because of its ability to synthesize f_o with high accuracy, with Pearson's $r > .97$ (Lee et al (2014); Liu et al (2013), which is well suited for the purpose of finding out the difficulty in focus perception in natural settings.

Table 1. Target stimuli used in the present paper ('A' stands for accented words, 'U' for unaccented; the accented mora is underlined and boldfaced)

		Word I		Word II		Word III
Short	A	***me****i-ga* May が May-NOM	×	***mo****mo* 腿 thigh	×	*-o* ***mi****ta* を見た -ACC saw
	U	*mei-ga* 姪が Niece-NOM		*momo* 桃 peach		*-ni nita* に似た -DAT resembled
Long	A	***mu****umin-ga* ムーミンが Moomin-NOM	×	***bu****dou* 武道 martial arts	×	*-o* ***mi****ta* を見た -ACC saw
	U	*noumin-ga* 農民が Farmer-NOM		*budou* 葡萄 grapes		*-ni nita* に似た -DAT resembled

16 native listeners (3 male) of Japanese were recruited for a naturalness judgment task. They were all born and raised in the Greater Tokyo area (Tokyo, Saitama, Kanagawa, and Chiba), and aged between 23 and 37 years old (mean age = 27.9). Most subjects had arrived in the UK for less than a year, except one who had arrived for 12 months, and another who had spent two years in the USA. In subsequent analyses these two listeners were not found to behave differently from the other listeners on the whole in any discernable way. None reported any history of speech or hearing impairment. Written informed consent was obtained from all participants in this pre-test and in Experiments 1 and 2 below. This study was approved by the UCL Research Ethics Committee (Project number: SHaPSetXU002).

The experiment took place in a quiet room in University College London. Participants were randomly assigned to one of two groups. One group judged the longer utterances (N = 64) while the other heard the shorter ones. Subjects were seated in front of a laptop computer, which displayed the Praat (Boersma & Weenink, 2016, version 5.4) ExperimentMFC interface, and wearing circumaural headphones. They listened to each stimulus and rated the naturalness on a 1~5 scale, with 5 being the most natural. Each stimulus could be replayed up to three times.

Results of the naturalness judgment test can be found in

Table 2. We are interested in whether the Type of Stimuli (original vs. synthesized) affects how a listener rates the naturalness of stimuli. A one-way repeated measures ANOVA shows that Type of stimuli has no main effect on naturalness judgment rating. This suggests that the two types of stimuli sounded equally natural to the native listeners. The grand mean rating of natural stimuli is 3.83, which is close to that of synthesized stimuli (3.74, out of a 1~5 scale).

Table 2. Naturalness ratings of synthetic vs. natural stimuli (and standard deviation).

	Natural	Resynthesis
Initial	3.71 (±1.33)	3.61 (±1.3)
Medial	3.79 (±1.34)	3.72 (±1.33)
Final	3.77 (±1.14)	3.79 (±1.25)
Neutral	4.04 (±1.15)	3.85 (±1.33)
Average	**3.83 (±1.25)**	**3.74 (±1.3)**

3. EXPERIMENT 1: PILOT STUDY

Having shown that resynthesized and natural stimuli sounded equally natural, the next step is to determine if they would yield the same results in a focus identification task. We recruited 7 native listeners (4 male) to take part in a pilot study. They were students who had moved to Hong Kong or England for less than 6 months at the time of the experiment. One of them had also lived in the USA for 4 years. No one reported any history of speech or hearing impairment. The stimuli were identical to those in the naturalness judgment task (§2) except that in this task the participants were to identify the narrow focus in a 4AFC (Word 1 / 2 / 3 / No emphasis) task. On this occasion all the participants heard the longer sentences (8 accented conditions × 4 focus conditions × 2 sources × 3 repetitions = 192 trials).

Table 3 shows that the accuracy of focus identification is highly similar between natural and resynthesized stimuli. For Final and Neutral foci, resynthesized stimuli even yielded better accuracy than natural stimuli. A paired samples t-test revealed that identification accuracy rates did not differ between natural and resynthesized stimuli, $t(27) = 1.238$, $p = .227$. It is thus safe to proceed to Experiment 2, where only resynthesized stimuli are used.

Table 3. Mean identification accuracy by focus condition (and standard deviation).

	Natural	Resynthesis
Initial	83% (±18%)	72% (±21%)
Medial	67% (±18%)	46% (±20%)
Final	44% (±20%)	45% (±24%)
Neutral	66% (±18%)	80% (±12%)
Average	**65% (±22%)**	**61% (±25%)**

4. EXPERIMENT 2: TRIAL WITH RESYNTHESIZED STIMULI

16 native listeners of Japanese were recruited to participate in a 4AFC task. The participants also took part in the naturalness rating task in §2. The stimuli were the same as those in §2 (the

synthetic subset only). Each participant was assigned to one of two groups, hearing either the longer (11 morae) or the shorter (8 morae) sentences. For a given listener, there were 8 accent conditions (2 accent ^ 3 words) × 4 focus conditions (Initial / Medial / Final / Neutral) × 3 repetitions = 96 trials in total.

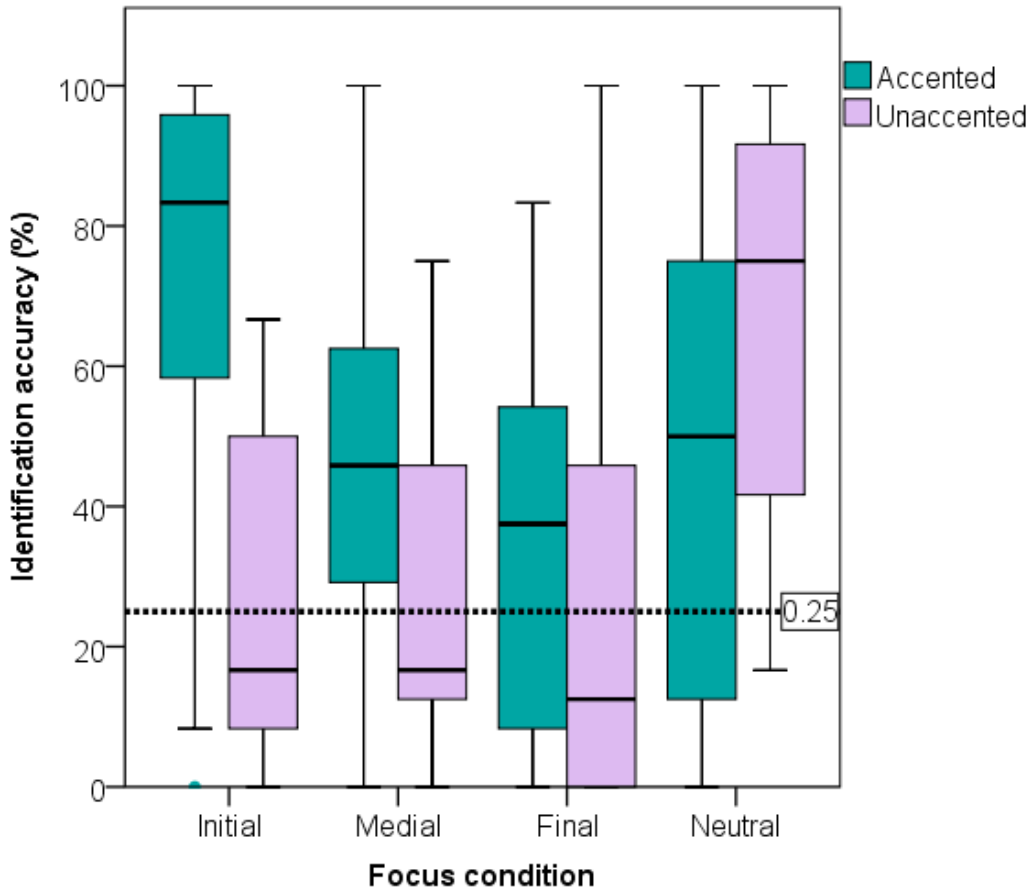
Figure 4 shows that, in general, focus is identified more accurately when the word bearing narrow focus is accented. Here the accent condition for Neutral focus is that of Word 1. In Neutral Focus statements, the focus condition is more accurately identified if Word 1 is unaccented. For all narrow focus conditions, an accented focus (turquoise box) yields higher identification accuracy than an unaccented focus (lilac box). On the other hand, for a broad focus statement, lexical pitch accent on the first word appears to make focus identification more difficult (47% vs. 67 %, chance = 25%, i.e. dotted line in Figure 4). On the whole, the most easily identified focus condition in statements is Neutral (57%), followed by Initial (49%), Medial (39%), and Final (31%). Two-way Repeated Measures ANOVA shows that the main effects of focus condition ($F(3,45) = 3.213$, $p = .032$) and accent condition of the focused word ($F(1,15) = 12.967$, $p = .003$) are significant, as is their interaction ($F(3,45) = 14.355$, $p < .001$). Post-hoc Tukey HSD tests indicated that the difference in focus identification accuracy between accented and unaccented foci was significant ($p = .003$); for focus condition, only the difference between final and neutral foci reached statistical significance ($p = .028$).

The combination of accent conditions also affects identification accuracy. Table 4 shows that, as predicted in §1, identification accuracy was the lowest in UAU, whereas focus in all-accented (AAA) utterances was the most correctly identified. One-way repeated measures ANOVA shows that the effect of accent combination on identification accuracy is significant, $F(7,105) = 5.658$, $p < .001$. Post-hoc Tukey HSD tests confirmed that identification accuracy was significantly lower in UAU than in AAA ($p < .001$), AUA ($p = .009$), and UAA ($p = .019$).

Table 4. Mean identification accuracy and RT by accent combination.

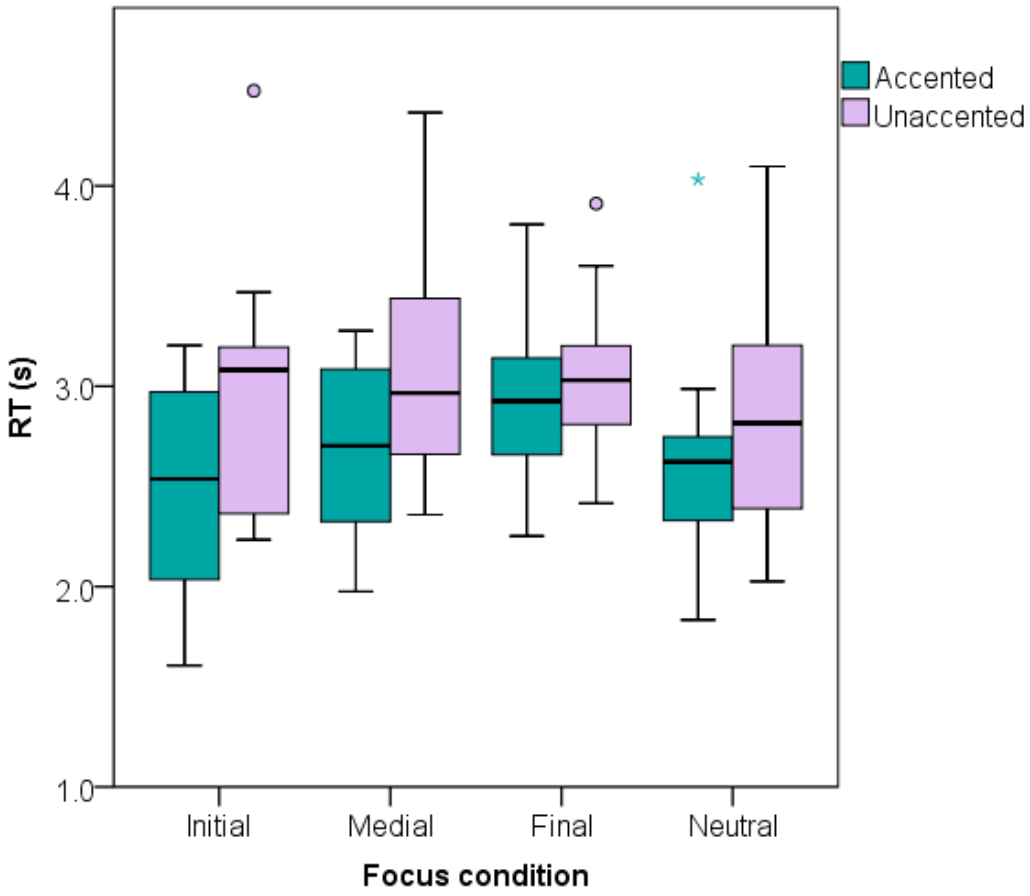
Accent comb.	Accuracy	S. D.	RT (s)	S. D.
AAA	0.56	0.22	2.88	1.27
AUA	0.48	0.17	2.89	1.26
UAA	0.47	0.16	2.83	1.09
UUU	0.45	0.20	3.09	1.33
AAU	0.45	0.21	2.84	1.13
AUU	0.41	0.17	2.92	1.31
UUA	0.36	0.11	3.00	1.63
UAU	0.32	0.11	2.89	1.32

Figure 4. Mean identification accuracy by focus condition and accent condition of the focused word (or of Word 1 in case of neutral focus).



In terms of reaction time (RT) (from the onset of the playback), participants generally responded faster when the narrow focus was accented (Figure 5). However, for broad focus statements, the better identification accuracy with an unaccented Word 1 was not accompanied by a shorter RT. Although all target stimuli were < 1.7 s in duration and mean RT was > 2.5 s, it appears that listeners responded more quickly when narrow focus occurs earlier. Two-way repeated measures ANOVA reveals that both focus condition ($F(3,45) = 4.413$, $p = .008$)) and the accent condition of the focused word ($F(1,15) = 11.380$, $p = .004$)) have a significant main effect on RT, whereas their interaction does not. Post-hoc Tukey HSD tests confirmed that listeners responded 312 ms faster to accented foci than to unaccented foci ($p = .004$). RT was also 249 ms faster for initial focus compared to final focus ($p = .019$), and 253 ms faster for neutral focus compared to final focus ($p = .017$). The contrasts between other focus conditions did not reach statistical significance. There is a general negative correlation between identification accuracy and RT ($r = -.230$, $p = .005$).

Figure 5. Mean RT by focus condition and accent condition of the focused word (or of Word 1 in case of neutral focus).



5. DISCUSSION

We set out to investigate which of the theoretically ambiguous focus conditions would be the most confusable with broad focus. Our results indicated that broad focus was generally more accurately identified; then within narrow focus, the earlier the focus the higher the identification accuracy. This seems to suggest that, as far as Japanese statements are concerned, 'phonetic' influence is stronger than syntactic and phonological ones. This is not to say that the prediction by intonational phonology (i.e. initial and broad focus being ambiguous) was not borne out. As it stands, identification accuracy is low across the board, with the highest being 57% (i.e. Neutral), showing that the aforementioned influences were all in play, and that when f_o is the only cue available it is generally hard to correctly identify narrow focus (or the lack thereof) even for native listeners.

The observed effect of accent condition of the focused item on identification accuracy can be interpreted as unaccented words being difficult for prosodic focus to realize. As seen in Figure 4, where the narrow focus is unaccented, identification is near chance. This characteristic of Japanese word prosody thus makes its focus different from that in languages like Mandarin, where focus identification is not known to be much lower for a particular lexical tone. Conceivable reasons for this discrepancy include that unaccented words do not have sharp f_o turning points and that post-focus compression is absent after an unaccented focus (Ishihara,

2011; Lee & Xu, 2012), hence leaving less room for the speaker to manipulate f_o and in turn for focus to stand out.

Since lexical pitch accent aids focus identification, it is understandable the AAA condition yields the highest accuracy (56%). Likewise the UUU condition was challenging, too, with mean accuracy at 45%, but not as challenging as the UAU condition (32%). While unaccented words make it difficult for prosodic focus to realize, it is an accented word standing out among unaccented words that is the most confusing to native listeners.

That RT is inversely correlated with identification accuracy is not surprising. In our data, an earlier (accented) narrow focus sees both shorter RT and higher accuracy. What is interesting is perhaps neutral focus, where the more accurately identified unaccented word condition does not see a corresponding shorter RT. Possibly, some of the Neutral Focus judgments were a result of listeners' failure to identify any narrow focus, thus resorting to the default choice (i.e. Neutral). The exact reason behind requires further investigation.

It is somewhat surprising to see that f_o cues alone, as conveyed by the synthetic f_o contours, was just as effective in conveying focus as the naturally focused utterances, which would have contained all the segmental, voice quality and duration cues (Maekawa, 1997; Sluijter et al., 1997; Xu, 1999). Equally intriguing is the finding that f_o contours generated by PENTAtrainer2 using parameters automatically extracted from natural speech was just as effective in conveying focus as the original f_o patterns in natural speech. This, plus the further finding that the synthetic f_o contours sounded just as natural as the original ones, suggest that such model-based synthetic prosody is worth further exploration in future research.

ACKNOWLEDGMENTS

This work is partially supported by the UCL Overseas Research Scholarship awarded to AL and FC.

REFERENCES

Beckman, M. E., & Pierrehumbert, J. B. (1986). Intonational structure in Japanese and English. *Phonology Yearbook, 3*, 255–309.

Boersma, P. P. G., & Weenink, D. J. M. (2016). Praat: Doing phonetics by computer. Retrieved from http://www.praat.org/

Botinis, A., Bannert, R., & Tatham, M. (2000). Contrastive tonal analysis of focus perception in Greek and Swedish. In A. Botinis (Ed.), *Intonation: Analysis, modelling and technology* (pp. 97–118). Dordrecht: Springer.

Botinis, A., Fourakis, M. S., & Gawronska, B. (1999). Focus identification in English, Greek and Swedish. In *Proceedings of the 14th International Congress of Phonetic Sciences (ICPhS 1999)* (pp. 1557–1560). San Francisco, CA.

Cooper, W. E., Eady, S. J., & Mueller, P. R. (1985). Acoustical aspects of contrastive stress in question-answer contexts. *Journal of the Acoustical Society of America, 77*(6), 2142–2156.

Gussenhoven, C. H. M. (1983). Testing the reality of focus domains. *Language and Speech, 26*(1), 61–80.

Ipek, C. (2011). Phonetic realization of focus with no on-focus pitch range expansion in Turkish. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)* (pp. 140–143). Hong Kong.

Ishihara, S. (2011). Focus prosody in Tokyo Japanese wh-questions with lexically unaccented wh-phrases. In *Proceedings of the 17th International Congress of Phonetic Sciences (ICPhS 2011)* (pp. 946–949). Hong Kong.

Ito, K. (2002). Ambiguity in broad focus and narrow focus interpretation in Japanese. In *Proceedings of the 1st International Conference on Speech Prosody (SP2002)* (pp. 411–414). Aix-en-Provence, France.

Lee, A., Prom-on, S., & Xu, Y. (in press). Pre-low raising in Japanese pitch accent. *Phonetica*.

Lee, A., & Xu, Y. (2012). Revisiting focus prosody in Japanese. In *Proceedings of the 6th International Conference on Speech Prosody (SP2012)* (pp. 274–277). Shanghai.

Lee, A., & Xu, Y. (2015). Modelling Japanese intonation using PENTAtrainer2. In *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*. Glasgow, Scotland.

Lee, A., Xu, Y., & Prom-on, S. (2014). Modeling Japanese f_o contours using the PENTAtrainers and AMtrainer. In *Proceedings of the 4th International Symposium on Tonal Aspects of Languages (TAL 2014)* (pp. 164–167). Nijmegen.

Liu, F., Xu, Y., Prom-on, S., & Yu, A. C. L. (2013). Morpheme-like prosodic functions: Evidence from acoustic analysis and computational modeling. *Journal of Speech Sciences, 3*(1), 85–140.

Maekawa, K. (1997). Effects of focus on duration and vowel formant frequency in Japanese. In Y. Sagisaka, W. N. Campbell, & N. Higuchi (Eds.), *Computing prosody: Computational models for processing spontaneous speech* (pp. 129–153). New York, NY: Springer.

Pierrehumbert, J. B., & Beckman, M. E. (1988). *Japanese Tone Structure*. Cambridge, MA: Massachusetts Institute of Technology.

Prom-on, S., & Xu, Y. (2012). PENTATrainer2: A hypothesis-driven prosody modeling tool. In *Proceedings of the 5th IESL Conference on Experimental Linguistics* (pp. 93–100). Athens, Greece.

Rump, H. H., & Collier, R. (1996). Focus conditions and the prominence of pitch-accented syllables. *Language and Speech, 39*(1), 1–17.

Sluijter, A. M. C., van Heuven, V. J. J. P., & Pacilly, J. J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America, 101*(1), 503–513.

Venditti, J. J., Maekawa, K., & Beckman, M. E. (2008). Prominence marking in the Japanese intonation system. In S. Miyagawa & M. Saito (Eds.), *The Oxford handbook of Japanese linguistics* (pp. 456–512). New York, NY: Oxford University Press.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of f_o contours. *Journal of Phonetics, 27*(1), 55–105.

Xu, Y., & Prom-on, S. (2014). Toward invariant functional representations of variable surface fundamental frequency contours: Synthesizing speech melody via model-based stochastic learning. *Speech Communication, 57*, 181–208.