PERCEPTION OF JAPANESE VOWEL LENGTH BY AUSTRALIAN ENGLISH LISTENERS

James Whang[1], Kakeru Yazawa[2], Paola Escudero[3]

[1]Saarland University, [2]Waseda University, [3]Western Sydney University
research@jameswhang.net, k-yazawa@aoni.waseda.jp, paola.escudero@westernsydney.edu.au

ABSTRACT

This study presents an experiment that investigates the perception of Japanese vowel length by Australian English (AusE) listeners. Previous studies found that AusE listeners utilize duration, spectra, and vowel inherent spectral change (VISC) to categorize native vowels. However, while Japanese vowels have phonemic vowel length distinctions, they typically do not exhibit VISC. In the absence of VISC, AusE listeners were expected to rely primarily on durational cues to categorize long/short Japanese vowels as long/short AusE monophthongs that match in height and backness. Participants generally showed the predicted categorization patterns, but unexpected results were also found in how Japanese /e/ and /u, uu/ were categorized. The results are discussed in terms of perceptual cue weighting in AusE (duration, spectra, VISC) and how this differs from the better-studied perception of Japanese vowels by American English (AmE) listeners.

Keywords: Australian English, Japanese, cross-linguistic perception, vowels, cue weighting

1. INTRODUCTION

Second language perception models such as the Speech Learning Model (SLM) [10], the Perceptual Assimilation Model (PAM) [2], and the Second Language Linguistic Perception (L2LP) model [7] posit that nonnative perception patterns are predictable from phonetic and phonological similarities between the first (L1) and second (L2) language sound categories. More importantly for the purposes of the current study, the acoustic cues used to establish sound contrasts are often weighted differently across languages and even dialects of the same language. For example, despite tense vowels being systematically longer than their lax counterparts in American English (AmE) (e.g., /iː/ vs. /ɪ/), native AmE listeners rely primarily on spectral cues when distinguishing the vowels [11]. However, in Australian English (AusE), the tense-lax vowel pairs have significant spectral overlap [4], making spectral cues less reliable for tense/lax vowel distinction. AusE listeners instead seem to assign higher weights to non-spectral cues compared to AmE listeners, such as duration and the presence vs. absence of large formant movements within the vowels [20].

The importance of spectral cues for AmE listeners is also evident during their L2 perception. AmE listeners show a strong tendency to categorize both long and short Japanese vowels as tense [16], presumably because Japanese vowels are generally peripheral (i.e., spectrally more tense vowel-like) regardless of phonemic length and because duration is an underutilized cue in AmE. However, unlike the case of AmE, the L2 perception patterns of native listeners of other dialects of English are understudied. We investigate, therefore, the perception of Japanese vowels by AusE listeners, who have been shown to be sensitive to durational cues at least in their L1.

We first compare the vowel systems of the languages in focus. First, the AusE vowel system consists of 13 monophthongs, and as summarized in Table 1, they differ from their AmE counterparts in that the tense/lax pairs are typically regarded as having long/short contrasts instead. Long vowels are approximately 1.5 times the length of their corresponding short vowels [5]. It should also be noted that although the lax high front vowel symbol /ɪ/ is used for the vowel in 'hid,' the vowel category has significant spectral overlap with /iː/ [4]. The 'feared' vowel /ɪə/ is included in the list of monophthongs because although vowel-rhotic sequences in AusE can be realized as centralizing diphthongs similar to British varieties of English (e.g., [bɪə] 'beer'), younger AusE speakers often produce them as long monophthongs (e.g., [bɪː] 'beer'), especially in closed syllables [4, 5]. Furthermore, some long vowels that are phonologically treated as monophthongs tend to exhibit large formant movements (e.g., /iː/ → [ᵊiː]), referred to as vowel inherent spectral change (VISC) [5, 6]. VISC in short vowels and monophthongized 'rhotic' vowels (e.g., /ɪə/) are negligible or significantly smaller than in long vowels, making VISC an important perceptual cue along with duration for distinguishing vowels that are spectrally similar (e.g., /iː, ɪə, ɪ/ →

[ᵊiː, ɪː, ɪ]) [8, 9, 20]. AusE listeners, therefore, are sensitive to duration and VISC when categorizing monophthongs, which contrasts with AmE listeners who rely primarily on spectral differences [16].

Table 1: Australian English monophthongs.

Long		Short	
Vowel	Word	Vowel	Word
/iː/	'heed'	/ɪ/	'hid'
/eː/	'haired'	/e/	'head'
/ɐː/	'hard'	/ɐ/	'hud'
/oː/	'hoard'	/ɔ/	'hod'
/ʉː/	'food'	/ʊ/	'hood'
/ɜː/	'heard'	/æ/	'had'
/ɪə/	'feared'		

Japanese is a five vowel system (/i, e, a, o, u/) with contrastive length (e.g. /toge/ 'thorn' vs. /tooge/ 'mountain pass'), making a total of ten monophthongal vowels. Although /u/ has long been described as high back unrounded [ɯ], younger speakers consistently articulate the vowel as rounded [17] and central [14, 17] (i.e., [ʉ]). Roundedness, therefore, is correlated with backness: back vowels /o, u/ are rounded while /i, e, a/ are not. On one hand, Japanese vowels are similar to AusE vowels in that duration is an important cue for vowel categorization and that spectral differences between long and short vowels are negligible. On the other hand, Japanese vowels differ from AusE vowels in that the magnitude of durational difference is larger in Japanese, approximately two to three times longer than short vowels [12, 13] compared to 1.5 times longer, and that Japanese vowels lack VISC. Given the two vowel systems, we hypothesize the following: (i) AusE listeners should categorize long and short Japanese vowels as long and short AusE vowels and (ii) in the absence of VISC, AusE listeners should show more reliance on spectral cues.

2. METHODS

2.1. Participants

Ten female Australian English listeners were recruited for the experiment at Western Sydney University, Sydney, Australia. Nine of the participants were undergraduate or graduate students between the ages 20 and 27, born and raised in the greater Sydney area. Six of the nine were monolinguals, and three were heritage speakers of Turkish, Punjabi, or Arabic. The tenth participant (age 33) had resided in Australia since the age of 5 and was an English-Spanish bilingual. All participants were compensated for their time in the form of class credit or monetary compensation.

2.2. Stimuli

The stimuli were 10 Japanese vowels—five short /i, e, a, o, u/ and five long /ii, ee, aa, oo, uu/—embedded in 3 phonetic contexts (/bVp, dVt, gVk/) spoken by 10 native Japanese speakers (5 male, 5 female) for a total of 300 tokens. The speakers were students or graduates of universities in Japan who had spent the majority of their lives in Tokyo and surrounding areas (aged 21 - 27, mean = 23.9). The recording took place in an anechoic chamber at Waseda University, using a SONY F-780 microphone with a sampling frequency of 44,100 Hz and 16-bit quantization. The speakers read aloud the sentence "*CVCe - CVCo - CVCe to CVCo ni wa V ga aru*" ('CVCe - CVCo - In CVCe and CVCo there is V'), and then the /e/ in the underlined /CVCe/ were clipped to create a /CVC/ stimuli (Japanese does not generally allow a syllable coda). Lexical pitch accent was placed on the first syllable. The stimuli were then manipulated to have a peak intensity of 70 dB SPL in Praat [3].

2.3. Procedure

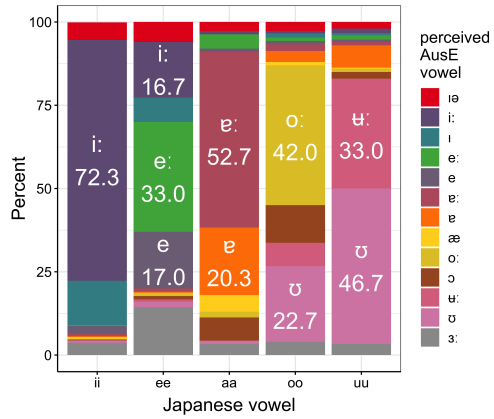
The experiment was a forced-choice task, where the participants had to categorize the vowel in the aforementioned /CVC/ stimuli presented in isolation. The participants had to choose from a list the word that contains the vowel that best matches the vowel they heard in the stimuli. The words in the list all had the shape /hVd/ with the exception of two words, which had a /fVd/ shape (as in Table 1). Participants were asked to choose as quickly as possible. The stimuli were presented in random order through noise-isolating headphones, and participants selected their answer choices by clicking the word choice with a mouse. A break was programmed to occur after 150 tokens (midpoint of experiment), which ended when participants clicked the mouse. The experiment was conducted using PsychoPy2 (*v1.90.2*) [18], which recorded participant responses and reaction times.

3. RESULTS

The prediction going into the experiment was that AusE listeners should primarily exhibit duration-based perception patterns, categorizing long Japanese vowels as long AusE vowels and short Japanese vowels as short AusE vowels that match in height and backness. Although this was generally the case, there were some exceptions. Figure 1 first shows the cate-

gorization of long Japanese vowels.

Figure 1: Response percentages for long Japanese vowels. Only responses > 15% labeled.



First, /uu/ was more frequently categorized as /ʊ/ rather than the expected /ʉː/. Second, while most Japanese vowels were matched with AusE vowels in terms of height and backness, /oo/ was an exception, which was categorized as either mid back /oː/ or high back /ʊ/. This suggests uncertainty with respect to the height of the vowel. A linear mixed effects regression model was fit to test how long the participants took in categorizing each of the Japanese long vowels using the *lme4* [1] package for *R* [19] and the *lmerTest* [15] package to obtain *p*-values for the model. The model included random intercepts for each participant and stimulus item. The results are shown in Table 2.

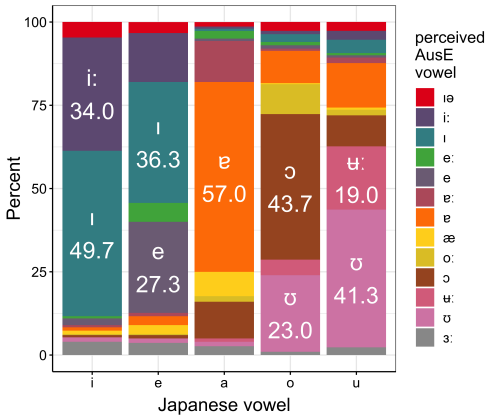
Table 2: Reaction times in long vowel categorization ($^{***}p < 0.001$). Baseline = /oo/.

	Estimate	S.E.	
(Intercept)	4.35	0.41	***
/ii/	−1.33	0.22	***
/ee/	−0.83	0.22	***
/aa/	−0.81	0.22	***
/uu/	−0.93	0.22	***

At 4.35 seconds, the mean reaction time when categorizing /oo/ was significantly longer than every other vowel. This suggests that /oo/ was relatively more difficult to categorize than other long vowels, perhaps because the participants found the vowel to be ambiguous in height. Pairwise comparisons were also conducted using the *difflsmeans* function in *R*, which revealed additionally that reaction times for /ii/ were significantly shorter than for /ee/ (-0.50 seconds; $p = 0.04$) and for /aa/ (-0.52 seconds; $p = 0.03$). No other comparison was significant.

Figure 2 shows that all short Japanese vowels were categorized as short AusE vowels, but again with some unexpected categorization patterns.

Figure 2: Response percentages for short Japanese vowels. Only responses > 15% labeled.



First, /e/ was most often categorized as the short high front vowel /ɪ/ rather than the expected mid front vowel /e/. Second, although /o/ was most often categorized as the expected short mid back vowel /ɔ/, it was also often categorized as the short high back vowel /ʊ/, which was also the case with /oo/.

A mixed effects model was fit to the short vowel reaction times as well with the same fixed and random effects structure as with the long vowels, and the results showed that reaction times were not significantly different among /e, o, u/ (Table 3). Pairwise comparisons were also conducted, which additionally revealed that reaction times between /i, a/ were not significantly different from each other and that reaction times for /e, o, u/ were all significantly longer than for /i, a/ ($p < 0.01$), suggesting that participants found it relatively easier to categorize /i, a/ than the other three vowels.

Table 3: Reaction times in short vowel categorization ($^{***}p < 0.001$). Baseline = /o/.

	Estimate	S.E.	
(Intercept)	3.80	0.35	***
/i/	−1.04	0.19	***
/e/	−0.09	0.19	
/a/	−0.82	0.19	***
/u/	−0.06	0.19	

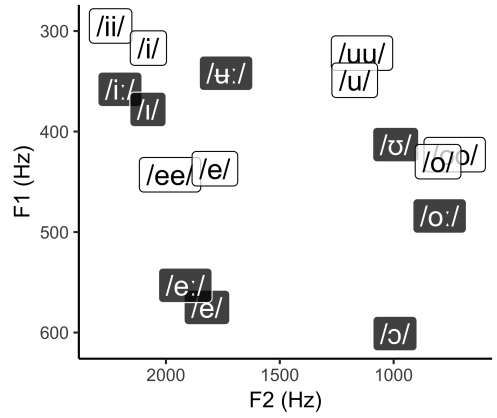
4. DISCUSSION

Overall, the results suggest that AusE listeners do make use of durational cues to categorize Japanese vowels. Long Japanese vowels were consistently categorized as long AusE counterparts and likewise for short vowels, supporting our hypothesis. This contrasts with AmE listeners, who have been shown to

be generally insensitive to durational cues, relying primarily on spectral cues [11, 16]. The difference is due to the fact that AmE vowels have tense/lax distinctions where lax vowels are more central than their tense counterparts. The same tense/lax categories, however, exhibit significantly more spectral overlap in AusE than in AmE, motivating AusE listeners to rely on other cues such as duration. The dialectal difference in perception patterns is in line with the predictions of models such as SLM, PAM, and L2LP, where nonnative perception is characterized by the properties of L1 categories, including how various acoustic cues such as spectral quality, duration, and VISC are weighted.

There were some exceptions to the duration-based perception patterns. Although /uu/ was categorized as the expected long vowel /ʉː/ more often than /u/, listeners more frequently chose the short /ʊ/ category for both vowels, suggesting that duration was being ignored in this case. There are two possible interrelated factors driving the observed pattern, the first of which is simply that Japanese /uu/ is spectrally ambiguous to AusE listeners. As can be seen in Figure 3 (adapted from [6, 16]), /ʉː/ is much more fronted than /ʊ/, articulated near /ɪ/ in AusE, suggesting that it is perhaps more accurately a high front rounded vowel (e.g., [ʏ]). Japanese /uu, u/, however, are intermediate between AusE /ʉː/ and /ʊ/ with the height of the former but backness of the latter. To an AusE listener, therefore, Japanese /uu/ is either an exceptionally high and long /ʊ/ or an exceptionally backed and VISC-less /ʉː/. Faced with these choices, AusE listeners seem willing to prioritize spectral quality over VISC and duration.

Figure 3: Mean central formants of AusE vowels (black box) [6] and Japanese vowels (white box) [16].



Spectral ambiguity also explains the categorization patterns of Japanese /oo, o/ as AusE /oː, ʊ/ and Japanese /e/ as AusE /ɪ/. /oo, o/ have the height of /ʊ/ but backness of /oː/. Japanese /e/ is similarly ambiguous in that it is higher than AusE /e/ but also lower and further back than AusE /ɪ/.

The second possible factor is that /ʉː/ often exhibits VISC in the F3 dimension (i.e., gradual lip rounding) [4], which both /uu/ and /ʊ/ lack. It could be the case that AusE listeners might be prioritizing the lack of F3 VISC itself as a cue for the short vowel /ʊ/ over the long duration as a cue for the long vowel /ʉː/. However, based on previous findings that VISC plays a complementary role to duration in AusE vowel categorization [20], it seems unlikely that (the lack of) VISC alone could override duration. In addition, the /uu/ categorization pattern contrasts with /ii/, which despite the similar lack of VISC was consistently categorized as the long AusE vowel /iː/. The main difference between /uu/ and /ii/ is perhaps that whereas /uu/ matches only in duration with /ʉː/, /ii/ matches with /iː/ in terms of spectral quality in addition to duration. The diverging patterns between /uu/ and /ii/ categorization in particular suggest again that although AusE listeners use durational cues, it is only when either spectral or (lack of) VISC cues also match. When there is a spectral and VISC mismatch, AusE listeners seem willing to ignore duration.

5. CONCLUSION

The current study presented an experiment investigating how native AusE listeners perceive nonnative Japanese vowels. The results showed that AusE listeners rely both on spectral and durational cues when categorizing Japanese long/short vowels, which differ from the results of previous studies with AmE listeners showing a general underutilization of durational cues. This is presumably due to the relative (un)importance of durational cues in the two English dialects, which is in line with the notion that L1-specific cue weighting drives nonnative and L2 speech perception. The results further showed that the role of (the lack of) VISC seemed to depend on the vowels being categorized. A follow-up study on how AusE listeners' perceive Japanese diphthongs (or vowel sequences) might shed more light on the complex interplay of VISC with spectral and duration cues in AusE listeners' perception of both L1 and L2 vowels.

6. ACKNOWLEDGMENTS

This research was supported by the Japan Society for the Promotion of Science [18J11517], the Australian Research Council (ARC) [FT160100514], and the ARC Centre of Excellence for the Dynamics of Language [CE140100041]. Thanks to Nicole Traynor for help running the experiments.

7. REFERENCES

[1] Bates, D., Mächler, M., Bolker, B., Walker, S. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1), 1–48.

[2] Best, C. T. 1995. A direct realist view of cross-language speech perception. In: Strange, W., (ed), *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York Press 171–204.

[3] Boersma, P., Weenink, D. 2018. Praat: doing phonetics by computer (Version 6.0.43) [Computer program].

[4] Cox, F. 2006. The acoustic characteristics of /hVd/ vowels in the speech of some Australian teenagers. *Australian Journal of Linguistics* 26(2), 147–179.

[5] Cox, F., Palethorpe, S. 2007. Australian English. *Journal of the International Phonetic Association* 37(3), 341–350.

[6] Elvin, J., Williams, D., Escudero, P. 2016. Dynamic acoustic properties of monophthongs and diphthongs in Western Sydney Australian English. *The Journal of the Acoustical Society of America* 140(1), 576–581.

[7] Escudero, P. 2005. *Linguistic perception and second language acquisition: Explaining the attainment of optimal phonological categorization*. PhD thesis Utrecht University.

[8] Escudero, P., Best, C. T., Kitamura, C., Mulak, K. E. 2014. Magnitude of phonetic distinction predicts success at early word learning in native and non-native accents. 5, 1059.

[9] Escudero, P., Mulak, K. E., Elvin, J., Traynor, N. M. 2017. "Mummy, keep it steady": Phonetic variation shapes word learning at 15 and 17 months. *Developmental Science* e12640.

[10] Flege, J. E. 1995. Second language speech learning: Theory, findings, and problems. In: Strange, W., (ed), *Speech perception and linguistic experience: Issues in cross-language research*. Timonium, MD: York Press 233–277.

[11] Hillenbrand, J. M., Clark, M. J., Houde, R. A. 2000. Some effects of duration on vowel recognition. *The Journal of the Acoustical Society of America* 108(6), 3013–3022.

[12] Hirata, Y. 2004. Effects of speaking rate on the vowel length distinction in Japanese. *Journal of Phonetics* 32(4), 565–589.

[13] Hirata, Y., Tsukada, K. 2009. Effects of speaking rate and vowel length on formant frequency displacement in Japanese. *Phonetica* 66, 129–149.

[14] Hisagi, M., Nishi, K., Strange, W. 2008. Acoustic properties of japanese and english vowels: Effects of phonetic and prosodic context. *Japanese/Korean Linguistics* 13, 223–224.

[15] Kuznetsova, A., Brockhoff, P. B., Christensen, R. H. B. 2017. lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software* 82(13), 1–26.

[16] Nishi, K., Strange, W., Akahane-Yamada, R., Kubo, R., Trent-Brown, S. A. 2008. Acoustic and perceptual similarity of Japanese and American English vowels. *The Journal of the Acoustical Society of America* 124(1), 576–588.

[17] Nogita, A., Yamane, N., Bird, S. 2013. The Japanese unrounded back vowel /ɯ/ is in fact unrounded central/front [ʉ - ʏ]. *Ultrafest VI Program and Abstract Booklet* 39–42.

[18] Peirce, J. W. 2007. Psychopy—psychophysics software in Python. *Journal of Neuroscience Methods* 162(1-2), 8–13.

[19] R Development Core Team, 2008. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing Vienna, Austria.

[20] Williams, D., Escudero, P., Gafos, A. I. 2018. Spectral change and duration as cues in Australian English listeners' front vowel categorization. *The Journal of the Acoustical Society of America* 144(3), EL215–221.