Japanese Vowel Devoicing Modulates Perceptual Epenthesis

Alexander J. Kilpatrick[1], Shigeto Kawahara[2], Rikke L. Bundgaard-Nielsen[3], Brett J. Baker[1], Janet Fletcher[1]

[1]University of Melbourne, [2]Keio University [3]MARCS Institute for Brain, Behaviour and Development, Western Sydney University

alex.kilpatrick@unimelb.edu.au

Abstract

This study investigates a relationship between perceptual epenthesis and vowel devoicing in Japanese. Across two experiments, epenthetic vowels are compared in environments where devoicing and deletion occur. In Experiment 1, participants assign illicit /VCCV/ and /VCVC/ tokens to /VCuCV/ and /VCVCu/ categories and judge how well tokens fit to the allocated category. In Experiment 2, participants discriminate between phonotactically illicit and licit tokens in AXB tests. The results show that illicit tokens are a better match to—and more difficult to discriminate from—their perceptually nearest legal counterpart when the target vowels are deleted than when they are merely devoiced.

Index Terms: Phonology, Phonetics, Japanese, Perceptual Epenthesis, Vowel Devoicing, Perceptual Assimilation.

1. Introduction

Standard Japanese (hereafter: Japanese) phonotactics do not allow non-homorganic consonant clusters and word-final, non-nasal consonants. As a result, Japanese listeners sometimes perceive an illusory, epenthetic /u/, which serves to perceptually repair the input to adhere to Japanese phonotactics [1], when they are exposed to such violations. It has been proposed that /u/ is epenthesised in these contexts because it is the shortest of all Japanese vowels [2], making it the phonetically minimal element of the language [1]. In line with a novel extension of the Perceptual Assimilation model [3], which predicts and accounts for the influence of L1 transitional probability on L2 perception [4], we propose that perceptual epenthesis is a process whereby illicit non-homorganic consonant clusters and word-final, non-nasal consonants are assimilated to their perceptually nearest and most predictable match. This process of assimilation reduces—or even eliminates—the perceptual distance between tokens that contain ether illicit consonant clusters (VCCV) or word-final, non-nasal consonants (VCVC) and tokens that adhere to Japanese phonotactics (VCVCV), resulting in the illusory vowel effect.

Vowel devoicing in Japanese can occur with all phonemically short vowels (hereafter: vowels); however, it is only systematic with high vowels (/u/ and /i/) [5]. Japanese vowel devoicing typically occurs either between voiceless consonants (/C̥VC̥/) or after a voiceless consonant at the end of a word (/C̥V#/) [6]. A comprehensive corpus analysis identifies several other contributing factors that influence vowel devoicing including the manner of articulation of the preceding (C_1) and following (C_2) consonant, as well as variation among individual phones [7]. For example, when /u/ occurs between two voiceless consonants, it undergoes devoicing 84% of the time, this increases to 98% if the C_1 is a voiceless fricative and the C_2 is a voiceless plosive, and further still to 99% if the C_1 is an /s/ and the C_2 is a /p/ (See Table 1). Others have proposed that the predictability of the preceding consonant in CV sequences [8] or the frequency of the words that carry the target vowel [9] can affect deletion/devoicing.

In Japanese, devoiced vowels that follow stops are typically phonetically realized differently from those that follow fricatives and affricates. Some argue that devoiced vowels that follow fricatives and affricates are deleted (See discussion in [6]) and that these instances of deletion may result in consonantal syllables [10]. Electropalatography [10] and electromagnetic articulography [11] studies have also suggested that these vowels are undergoing deletion, with speakers not exhibiting lingual configurations typical of the Japanese /u/ in some devoiced tokens.

In the following, we refer to these post-fricative devoiced allophones as "deleted" vowels in order to distinguish them from devoiced vowels that follow voiceless stops. Phonological vowel deletion in these contexts would mean that voiceless non-homorganic consonant clusters were phonotactically permissible in Japanese and therefore unlikely to elicit perceptual epenthesis; numerous experiments—including those featured in the present paper—have shown that this is not the case. Instead, we propose that some surface level representation of deleted vowels remain but that these near-zero allophones act as a better match to vowel-less sequences. These near-zero allophones maintain fewer or less salient acoustic cues than devoiced or voiced vowels, making them perceptually minimal.

In line with the aforementioned extension of the Perceptual Assimilation model, we also propose that the perceptual minimality of the expected allophone has an influence on the perceptual distance between illicit sequences and their perceptually nearest, predictable match. Indeed, we argue that this is due to sequences that predictably stimulate perceptually minimal allophones eliciting less discriminable illusory vowels because the assimilation distance between the illicit sequence and its nearest match is narrowed. We test this hypothesis in two experiments that examine epenthetic vowels which occur after voiceless fricatives, voiceless plosives and voiced consonants. In Experiment 1, participants assign VCCV and VCVC tokens to VCVCV categories and assign goodness of fit (GoF) ratings to how well they adhere to a given category. In Experiment 2, participants discriminate VCCV and VCVC tokens from VCVCV tokens in a series of AXB discrimination experiments.

.

Table 1. *List of experimental tokens and rates at which /u/ undergoes devoicing in Japanese discourse. Here, Env .% = devoicing rates in the specific environment tested, Manner = rates based on the manner of articulation of the C_1 and C_2 in /C̥VC̥/ sequences and Voicing = devoicing rates based purely on the voice/voicelessness of the C_1 and C_2 [7].*

Location	Allophone	Licit Token	Illicit Token	Environment	Env. %	Manner	Voicing
Medial	Deleted	/esupo/	/espo/	/s_p/	99%	98%	84%
Medial	Devoiced	/ekupo/	/ekpo/	/k_p/	88%	80%	84%
Medial	Devoiced	/epuso/	/epso/	/p_s/	60%	74%	84%
Medial	Voiced	/egupo/	/egpo/	/g_p/	N/A	N/A	2%
Medial	Voiced	/ezubo/	/ezbo/	/z_b/	N/A	N/A	1%
Medial	Voiced	/ebuzo/	/ebzo/	/b_z/	N/A	N/A	1%
Final	Deleted	/epusu/	/epus/	/s_#/	N/A	N/A	N/A
Final	Devoiced	/esupu/	/esup/	/p_#/	N/A	N/A	N/A

2. Method

2.1. Stimuli

A full list of all 16 tokens appears in Table 1. The stimuli were produced by three phonetically trained female Australian English (AustE) speakers. These were recorded in a recording studio located at the University of Melbourne and had a bit depth of 64kb/sec and a sample rate of 48kHz. Each speaker produced five consecutive repetitions of each of the 16 tokens. The first and fifth repetitions were not used in Experiment 1 to avoid any effects of list initial unfamiliarity and list final intonation patterns. The remaining three tokens were excised with a 20 ms ramp-in and a 10 ms ramp-out. On average, /u/ duration in medial licit tokens was 85 ms (range 67-106 ms, *SD* =13 ms); average target /u/ duration in word-final licit tokens was 152 ms (range 95-279 ms, *SD* = 40 ms). Contrasting licit/illicit token pairs were designed so that the production of licit stimuli would predictably produce varying allophones in the target /u/; deleted (e.g., /es**u**po/), devoiced (e.g., ep**u**so) and voiced (e.g., /ez**u**bo/) (see Table 2).

2.2. Participants

34 undergraduate students from the Mita campus of Keio University were recruited as participants for Experiments 1 and 2. Participants were all L1 Japanese speakers, born to L1 Japanese speaking parents. Participants were recruited by word of mouth. Participant ages ranged from 18 to 26 (*M* = 20, *SD* = 1.6) and were selected on the basis of limited exposure to languages other than Japanese although all participants had previously studied English due to it being a compulsory subject in the Japanese education system.

2.3. Procedure: Experiment 1

Experiment 1 took place in a quiet room located at the Mita campus of Keio University. The experiment consisted of a single block of trials which contained all 16 tokens. Participants were asked to categorise tokens into 8 categories. These categories were presented to the participants as on-screen buttons with Hiragana labels (categories, tokens and Hiragana labels presented in Table 3). Tokens were drawn at random from a library of 144 stimuli (16 tokens x 3 speakers x 3 repetitions each). Upon assigning each token to a category, participants were asked to assign a GoF rating to indicate how well the token fit to the assigned category. This was presented to participants as a Likert scale ranging from 1 to 7. To explain that a low score was supposed to indicate a poor fit, the 1 on the Likert scale was labelled 違う (different) and the 7 was labelled 同じ (identical). We predict that listeners will assign higher GoF ratings to illicit tokens with C_1 voiceless fricatives due to deleted vowels being a better match to vowel-less sequences.

2.4. Procedure: Experiment 2

Experiment 2 was conducted directly after Experiment 1 in the same location. In Experiment 2, participants were required to respond to 192 AXB discrimination trials, 24 triads for each of the 8 licit/illicit contrasts (Table 2). To avoid speaker or phone sequence bias, tokens were organized into six speaker sequences (123, 132, 213, 231, 312, 321) and each of the speaker sequences was organised into four token sequences (AAB, ABB, BAA, BBA). All contrasts were presented to participants in a single block from which each AXB triads were drawn at random with a replace paradigm so that any trial that timed out was replayed later during the experiment. Both discrimination accuracy and response times were recorded. Tokens were spaced with a 1000 ms inter-stimulus interval. Here we predict that listeners will have greater difficulty discriminating between contrasts with voiceless fricatives in the C_1 position due to the smaller perceptual distance between the deleted vowel and the vowel-less sequence.

3. Results: Experiment 1

3.1. Categorisation Rate

Participants categorised most tokens to their perceptually nearest phonotactically licit category where the phonotactic violation is repaired by a /u/. All illicit tokens but one were categorised according to this prediction at a rate of 90% or greater. The one illicit token that did not adhere to a 90% categorisation rate was /egpo/, which was categorised as /ekupo/ 25% of the time. Licit tokens were also assigned to their predicted category at a rate of 90% or greater except in the case of the /epusu/ token which was categorised as /epuso/ 30% of the time and was only assigned to its predicted /epusu/ category 67% of the time; this is less than its illicit counterpart, /epus/, which was categorised as /epusu/ 95% of the time.

Table 2. *AXB contrasts organized by word position and the most likely target allophone in licit tokens.*

.	Deleted	Devoiced	Voiced
Word Medial	/esupo/-/espo/		/ezubo/-/ezbo/
		/epuso/-/epso/	/ebuzo/-/ebzo/
		/ekupo/-/ekpo/	/egupo/-/egpo/
Word Final	/epusu/-/epus/	/esupu/-/esup/	

Table 3. *Categorisation rates and Goodness of Fit ratings for licit and illicit tokens. Goodness of Fit ratings are presented in parenthesis. Categorisation rates less than 1% are not featured.*

	Medial Contrasts						Final Contrasts	
	えすぽ	えくぽ	えぷそ	えぐぽ	えずぼ	えぶぞ	えすぷ	えぷす
	esupo	ekupo	epuso	egupo	ezubo	ebuzo	esupu	epusu
/esupo/	98% (5.67)				1% (2.5)			1% (2.75)
/espo/	94% (5.81)				1% (2)			5% (5)
/ekupo/		90% (5.44)		10% (4.43)				
/ekpo/		91% (5.18)		9% (3.59)				
/epuso/			91% (5.5)			5% (2.93)	4% (3.83)	
/epso/			95% (5.18)			1% (2.75)	4% (3.17)	
/egupo/				100% (6.03)				
/egpo/		25% (4.13)		75% (4.79)				
/ezubo/					100% (5.43)			
/ezbo/	3% (4.5)				97% (4.62)			
/ebuzo/						100% (5.72)		
/ebzo/						100% (5.21)		
/epusu/			30% (4.47)			3% (2.56)	67% (4.62)	
/epus/			4% (3.00)			1% (2.75)	95% (4.78)	
/esupu/	9% (4.86)							90% (4.97)
/esup/	9% (1.93)							91% (3.81)

3.2. Goodness of Fit Rating

Overall, licit tokens achieved a higher average GoF rating (5.37) than illicit tokens (4.91); $t(271) = 6.24$, $p = < 0.001$. The only tokens that did not adhere to this pattern were those with a voiceless fricative preceding the epenthetic context. In these deleted contexts, the illicit tokens achieved higher GoF ratings than the licit tokens, /espo/ (5.81) was rated significantly higher than /esupo/ (5.67); $t(33) = -2.95$, $p = 0.006$, and /epus/ (4.78) was rated significantly higher than /epusu/ (4.62); $t(33) = -6.56$, $p = < 0.001$, despite being assigned to the /esupo/ and /epusu/ categories respectively.

Table 4. *Average scores for licit and illicit tokens, difference between scores and results from paired sample t-tests.*

Licit		Illicit				
Token	**GoF**	**Token**	**GoF**	**Diff.**	***t***	***p***
esupo	5.67	espo	5.81	-0.14	-6.87	< 0.001
ekupo	5.44	ekpo	5.18	0.26	10.92	< 0.001
epuso	5.5	epso	5.18	0.32	10.11	< 0.001
egupo	6.03	egpo	4.79	1.24	24.54	< 0.001
ezubo	5.43	ezbo	4.62	0.81	25.54	< 0.001
ebuzo	5.72	ebzo	5.21	0.51	17.29	< 0.001
epusu	4.62	epus	4.78	-0.16	-5.32	< 0.001
esupu	4.97	esup	3.81	1.16	31.11	< 0.001
Average	5.42		4.92	0.5		

4. Results: Experiment 2

4.1. Medial Contrasts

The discrimination accuracy results support our hypothesis that deletion contexts (e.g., /es**u**po/) are harder to discriminate from vowel-less tokens than devoiced or voiced contexts (e.g., /ep**u**so/ or /ez**u**bo/). Of the medial AXB tests, participants were least accurate at discriminating between deleted contrasts (M = 68%, SD = 14%), followed by devoiced contrasts (M = 75%, SD = 16%) and finally voiced contrasts (76%, SD = 14%). A one-way ANOVA between voicing conditions was conducted to compare the effect of voicing of the predictable allophone in the target position on test accuracy. The ANOVA revealed a significant main effect; $F(2, 201) = 3.1$, $p < 0.05$. Post hoc comparisons using the Bonferroni correction revealed a significant difference between deleted and voiced contrasts ($p < 0.05$) but not between deleted and devoiced ($p = 0.185$) or devoiced and voiced contrasts ($p = 1$).

In medial contrasts, response time results largely mirror the results in terms of accuracy whereby participants required more time to respond to contrasts that were difficult to discriminate. Participants took longest to respond to deleted contexts (M = 1279 ms, SD = 136 ms), followed by devoiced contexts (M = 1261 ms, SD = 135 ms), and finally voiced contexts (M = 1241 ms, SD = 139 ms). A one-way ANOVA of voicing conditions on response time also revealed a significant effect; $F(2, 4895) = 6.5$, $p < 0.01$. As with accuracy, a post hoc comparison with Bonferroni correction revealed a significant difference between deleted and voiced conditions ($p < 0.01$) but not between deleted and devoiced ($p = 0.087$) or devoiced and voiced ($p = 0.338$).

4.2. Word-Final Contrasts

As with medial contrasts, participants were less accurate at discriminating between the word-final deleted contrast (/epusu/-/epus/ M = 81%, SD = 8%) compared to the word-final devoiced contrast (/esupu/-/esup/ M = 86%, SD = 12%). A paired-samples t-test was conducted to compare the accuracy results of /epusu/-/epus/ and /esupu/-/esup/ contrasts. This revealed a significant difference between the two word-final contrasts ($t(33) = 2.5$, $p < 0.05$). Participants also took longer to respond to the /epusu/-/epus/ contrast (M = 1325 ms, SD = 131 ms) compared to the /esupu/-/esup/ contrast (M = 1299 ms, SD = 156 ms). A paired samples t-test calculated on this difference revealed a significant difference ($t(813) = 2.19$, $p < 0.05$).

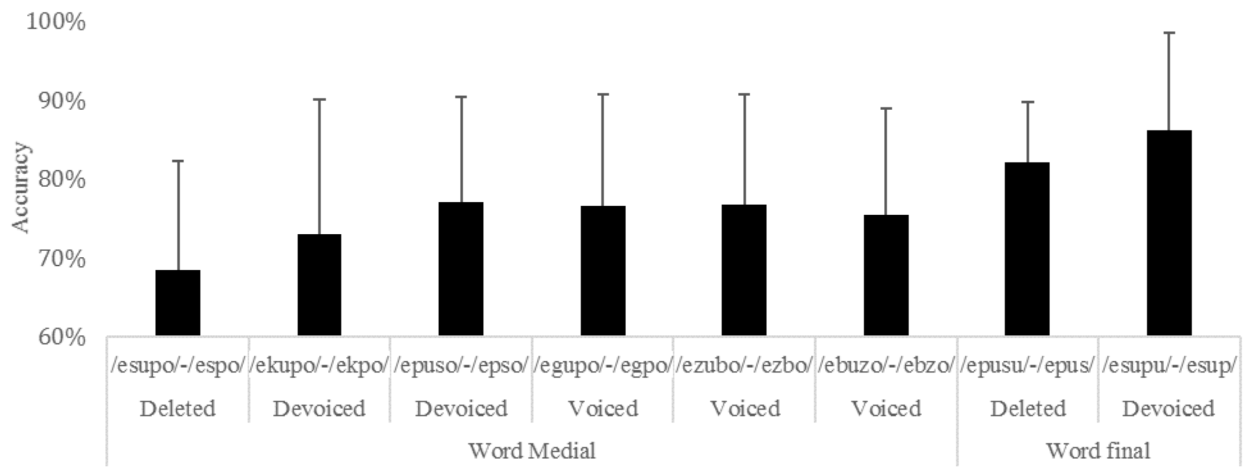

Figure 1. *AXB test accuracy. Error bars represent standard deviation.*

5. Discussion

5.1. Experiment 1: Categorisation and Goodness of Fit

In Experiment 1, we predicted that participants would assign higher GoF ratings to tokens with a voiceless fricative in the C_1 position. This hypothesis is supported by the results of this experiment whereby /espo/ and /epus/ tokens achieved higher GoF ratings than even their licit counterparts. We propose that this is likely due to the perceptual minimality of deleted vowels. Sequences that predictably elicit perceptually minimal vowels are a better match to vowel-less sequences than the voiced vowels produced by our AustE speaking volunteers.

5.2. Experiment 2: AXB Discrimination

In Experiment 2, we predicted that listeners would have more difficulty discriminating between contrasts where the epenthetic context would predictably undergo vowel deletion. This hypothesis is reflected in both medial and word-final discrimination accuracy results which show that contrasts were less discriminable when the C_1 was a voiceless fricative compared to other consonants. Deleted contrasts were significantly more difficult to discriminate than devoiced or voiced contrasts. This suggests that contrasts are more discriminable when the epenthetic vowel would predictably elicit voicing if the token were spoken by a Japanese speaker.

6. Conclusion

The present report demonstrates that Japanese listeners are more likely to perceive an epenthetic /u/ when the C_1 is a voiceless fricative when compared with voiceless stops or voiced consonants. In line with the aforementioned extension of PAM [4], phonotactically unattested sequences are assimilated to a predictable match. When the epenthetic context is preceded by a voiceless fricative, the assimilation distance is shortened due to the perceptual minimality of "deleted" vowels. One possible explanation for the difference between C_1 voiceless fricative and C_1 voiceless plosive contexts is that the turbulent aperiodic energy of the fricative masks the acoustic cues of the target vowel more substantially than the release of the stop. This masking makes these near-zero allophones a better match to vowel-less sequences, encouraging assimilation to the target phoneme. This assimilation reduces or eliminates the perceptual distance between illicit sequences and their nearest, most predictable match, making illicit tokens (e.g., /espo/) more acceptable and making contrast pairs (e.g., /espo/-/esupo/) less divergent.

7. Acknowledgements

We would like to thank our research participants and Keio University. We also thank the W.T. Mollison scholarship committee at the University of Melbourne, as well as Cathleen Benevanto, Rosey Billington, Katie Jepson and Eleanor Lewis.

8. References

[1] Dupoux, E., Parlato, E., Frota, S., Hirose, Y., & Peperkamp, S. "Where do illusory vowels come from?", Journal of Memory and Language, 64(3), 199-210, 2011.

[2] Arai, T., Warner, N., & Greenberg, S. "OGI tagengo denwa onsei koopasu-ni okeru nihongo shizen hatsuwa onsei no bunseki" [analysis of spontaneous Japanese in OGO multi-langauge telephone speech corpus], The Spring Meeting of the Acoustical Society of Japan, 1, 361-362, 2001.

[3] Best, C. T. "The emergence of native-language phonological influences in infants: A perceptual assimilation model." The development of speech perception: The transition from speech sounds to spoken words, 167(224), 233-277, 1994.

[4] Kilpatrick, A. J., Bundgaard-Nielsen, R. L., & Baker, B. J. "Japanese Co-occurrence Restrictions Influence Second Language Perception", Applied Psycholinguistics, In Press.

[5] Maekawa, K. "Hatsuwa sokudo ni yoru yūsei kukan no hendō" IEICE Technical Report SP89-148. 47–53, 1990.

[6] Fujimoto, M. "Vowel Devoicing", in H. Kubozono [Ed], Handbook of Japanese Phonetics and Phonology, Walter de Gruyter, 167-214, 2015.

[7] Maekawa, K., & Kikuchi, H. "Corpus-based analysis of vowel devoicing in spontaneous Japanese: an interim report.", in J. van de Weijer, K. Nanjo & T. Nishihara [Eds], Voicing in Japanese, 84, 205-228, 2005.

[8] Whang, J. "Recoverability-driven coarticulation: Acoustic evidence from Japanese high vowel devoicing." The Journal of the Acoustical Society of America, 143, 1159, 2018.

[9] Kilpatrick, A. J., Bundgaard-Nielsen, R. L., & Baker, B. J. "Japanese Vowel Deletion Occurs in Words in Citation Form", Proceedings of the 16th Australasian International Conference on Speech Science and Technology, 325-328, 2016.

[10] Matsui, M. "On the input information of the C/D model for vowel devoicing in Japanese." Journal of the Phonetic Society of Japan 21:1, 127-140, 2017.

[11] Shaw, J., & Kawahara, S. "The lingual gesture of devoiced /u/ in Tokyo Japanese." Journal of Phonetics, 66, 100-118, 2018.