INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Perception of Non-Contrastive Variations in American English by Japanese Learners: Flaps are Less Favored Than Stops

Kiyoko Yoneyama[1], *Mafuyu Kitahara*[2] *and Keiichi Tajima*[3]

[1]Daito Bunka University, Japan
[2]Sophia University, Japan
[3]Hosei University, Japan

yoneyama@ic.daito.ac.jp, mafuyu@sophia,ac.jp, tajima@hosei.ac.jp

Abstract

Alveolar flaps are non-contrastive allophonic variants of alveolar stops in American English. A lexical decision experiment was conducted with Japanese learners of English (JE) to investigate whether second-language (L2) learners are sensitive to such allophonic variations when recognizing words in L2. The stimuli consisted of 36 isolated bisyllabic English words containing word-medial /t/, half of which were flap-favored words, e.g. *city*, and the other half were [t]-favored words, e.g. *faster*. All stimuli were recorded with two surface forms: /t/ as a flap, e.g. *city* with a flap, or as [t], e.g. *city* with [t]. The stimuli were counterbalanced so that participants only heard one of the two surface forms of each word. The accuracy data indicated that flap-favored words pronounced with a flap, e.g. *city* with a flap, were recognized significantly less accurately than flap-favored words with [t], e.g. *city* with [t], and [t]-favored words with [t], e.g. *faster* with [t]. These results suggest that JE learners prefer canonical forms over frequent forms produced with context-dependent allophonic variations. These results are inconsistent with previous studies that found native speakers' preference for frequent forms, and highlight differences in the effect of allophonic variations on the perception of native-language and L2 speech.

Index Terms: alveolar flaps, canonical forms, allophonic variants, second-language learners

1. Introduction

Studies on second-language (L2) speech learning have often focused on L2 learners' production and perception of sounds that signal lexical distinctions in L2, e.g. [1]. Much less attention has been given to sounds that do not signal lexical contrasts. However, learning to produce and perceive non-contrastive sounds could be important for L2 learners, particularly if they want to achieve native-level performance.

In American English (AE), intervocalic alveolar stops are realized as alveolar flaps, e.g. *better*, *rider*, *get up*, *need it*. Alveolar flaps are allophonic variants of /t/ and /d/ and do not signal lexical contrasts with other sounds in English.

The present line of research has focused on the production and perception of AE alveolar flaps by Japanese learners of English (JE learners). The Japanese language has a single liquid consonant, which is often described as an alveolar flap, e.g. [2][3][4][5][6]. This provides an interesting scenario for studying the effect of learners' first language (L1) on L2 learning. That is, Japanese speakers produce and perceive alveolar flaps in the context of their L1, i.e. as a phonetic realization of the Japanese liquid consonant. If the ability to produce and perceive alveolar flaps positively transfers to L2, then JE learners might be expected to produce and perceive alveolar flaps appropriately in English. However, since alveolar flaps are allophonic variants of /t/ and /d/ in AE, not a phonetic realization of /r/ as in Japanese, JE learners may have difficulty appropriately associating alveolar flaps with the correct sounds in English.

As a first step in addressing this issue, we looked at the English Read by Japanese Corpus, compiled by Minematsu and his colleagues [7]. This corpus contains English utterances by 202 Japanese students from 20 Japanese universities across Japan. They were native speakers of Japanese whose English exposure was very limited. They read 120 English sentences, among other materials. We then searched for potentially flappable segments. We checked whether intervocalic /t/ and /d/ which would typically be produced as alveolar flaps by native AE speakers were produced as such by the Japanese speakers, in a total of 12000 tokens in 477 sentences. Surprisingly, the result showed that flaps were produced in only 8 tokens, in 4 kinds of sentences. Flap rate, or the rate of producing alveolar flaps, in this corpus was virtually 0%. This means that Japanese university students typically do not produce alveolar flaps in English, despite the fact that they produced alveolar flaps in L1.

Next, we conducted a production experiment [8], to examine how often more advanced JE learners might produce alveolar flaps in English. The participants were 40 Japanese learners of English with some experience living in North America, ranging from 2 weeks to 10.5 years. They read 65 words and phrases embedded in a carrier sentence, e.g. *Say 'party' now*, *Say 'get it' now*. We then conducted auditory transcription and acoustic analysis of the target segment. The results showed that 37.6% of the target segment were transcribed as an alveolar flap. The percentage varied depending on factors such as item type, age of arrival, duration of stay, and TOEFL score. These results suggest that JE learners with some experience living in North America do frequently produce alveolar flaps in English.

We then turned to perception, and conducted a word identification experiment [9], in order to examine how accurately JE learners perceive English words containing alveolar flaps. The participants were 39 JE learners with no experience living in English-speaking communities. The stimuli were English minimal pairs that contrasted between /t/ or /d/ vs. /r/ or /l/, e.g. *genetic* vs. *generic*, *fading* vs. *failing*. The /t/ or /d/ was produced as either an alveolar stop or a flap.

http://dx.doi.org/10.21437/Interspeech.2017-207

The task was a minimal pair 2-alternative forced-choice identification task. The results showed that accuracy was significantly lower for tokens produced with a flap (80.1%) than for those produced with a stop (97.9%, $p < .001$). Thus, JE learners have difficulty correctly identifying words pronounced with alveolar flaps. This suggests that JE learners are sensitive to allophonic variations in L2.

In light of these results, the present study addressed the following research question: If JE learners are sensitive to allophonic variations in L2, can they utilize this variation when recognizing words in spoken English? Previous studies have shown that native AE listeners utilize allophonic variation when recognizing words in English, e.g. [10][11][12]. We wanted to see whether L2 learners behave similarly.

There are two possible hypotheses. First, the *frequency-of-exposure hypothesis* says that frequent forms are recognized more accurately and quickly, e.g. [10][11][12][13][14][15]. So the word *city* would be recognized more accurately if it was produced with an alveolar flap than an alveolar stop, since *city* is frequently produced with a flap in AE. However, the word *faster* would be recognized more accurately if it was NOT produced with a flap, since *faster* is frequently produced with a stop in AE. On the other hand, the *canonical-form-dominance hypothesis* says that canonical forms or dictionary pronunciations are recognized more accurately and quickly, e.g. [16][17][18][19][20][21]. According to this hypothesis, both the words *city* and *faster* would be recognized more accurately if it was produced with a stop than if it was produced with a flap, since canonical forms for both words would contain a stop. Pitt and his colleagues [12] found that native AE speakers were more accurate with frequent forms rather than canonical forms, supporting the frequency-of-exposure hypothesis. We wanted to see whether JE learners would show a similar pattern or not.

2. Lexical Decision Experiment

2.1. Participants

Participants were 53 Japanese learners of English. They were undergraduate students who were born and raised in Japan. They have been learning English as a second language for at least 6 years from junior high school. They had relatively limited English experience, although they are considered to have better English skills compared to most Japanese undergraduate students. They had either pre-1st EIKEN or TOEIC scores of 700 or higher at the time of participation.

2.2. Materials

The target stimuli consisted of 36 isolated bisyllabic English words containing word-medial /t/, half of which were flap-favored words, e.g. *city*, *better*, and the other half were [t]-favored words, e.g., *faster*, *custom*. All stimuli were recorded with two surface forms: /t/ produced as a flap, e.g. *city* with a flap, or /t/ produced as [t], e.g. *city* with [t]. Each participant heard 9 target words each in four word conditions as shown in Table 1.

The filler stimuli consisted of 72 isolated words containing no word-medial /t/, half of which were real English words, e.g. *young*, *museum*, and the other half were nonwords, e.g. *gool* [gul], *sprink* [sprɪŋk].

All the stimuli were recorded by a male AE speaker from the Midwest who was an undergraduate linguistics major with some phonetics training.

Table 1: *Materials (Target Words)*.

Stop-favored words		**Word Condition**	**Description**
faster	filter	t_t	stop-favored words pronounced with a stop
custom	instance		
lifted	distant		
master	practice		
safety	actor	t_flap	stop-favored words pronounced with a flap
sister	active		
system	after		
western	plastic		
doctor	justice		
Flap-favored words		**Word Condition**	**Description**
party	better	flap_flap	flap-favored words pronounced with a flap
pretty	city		
water	daughter		
writer	getting		
motor	later	flap_t	flap-favored words pronounced with a stop
native	letter		
notice	little		
voted	matter		
item	meeting		

2.3. Procedures

Participants were tested individually in a quiet environment. The stimuli were presented through headphones at a comfortable listening level. The experiment was controlled by E-Prime software [22]. The participants performed a lexical decision task where they listened to the stimulus word carefully and decided if it was a real word or not, as quickly and accurately as possible. The stimuli were counterbalanced in the lists using a Latin Square design, so that the participants only heard one of the two surface forms for each word. The test session consisted of 108 trials, in which the stimuli were presented in a random order. A practice session consisting of 4 trials was given before the test session.

2.4. Analyses

2.4.1. Word Accuracy

A logistic mixed effects model was used for the analysis of word accuracy. The random factors were crossed intercepts for subject and word which was nested under the favored environment, i.e., [t]-favored and flap-favored. The fixed factor was word condition, i.e., t_t, flap_t, t_flap, and flap_flap (see Table 1).

Figure 1 displays boxplots of the word recognition accuracy across the word conditions. The logistic mixed effects model demonstrated that the word condition was significant, $\chi^2(3) = 17.21$, $p < 0.001$, suggesting that the word recognition accuracy was significantly different across the conditions. Post-hoc analyses for each pair-wise comparison demonstrated that the accuracy in the flap_flap condition was significantly lower than in the t_t condition, $\chi^2(1) = 8.56$, $p = 0.021$, and the flap_t condition, $\chi^2(1) = 19.08$, $p < 0.01$. Other contrasts were not significantly different, $p > 0.05$. For the post-hoc analyses, p-values were adjusted with Bonferroni correction.

2.4.2. Reaction Times

Figure 2 displays the reaction time across the word conditions. A linear mixed effects model was used for the reaction time

analysis. Both the random and fixed factors were the same as in the word accuracy analysis. The linear mixed effects model demonstrated that the word condition did not significantly affect the reaction time, $\chi^2(3) = 6.33$, $p > 0.05$. Reaction times by JE learners in the present study were much longer than those by AE listeners reported in Pitt *et al*.'s study [12].

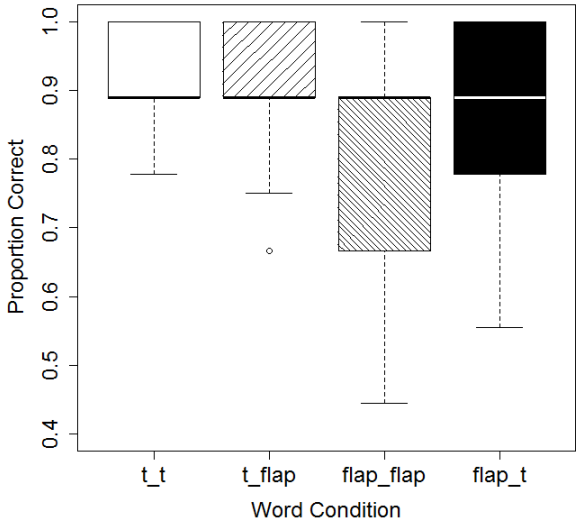

Figure 1: *Word identification accuracy across the word condition described in Table 1.*

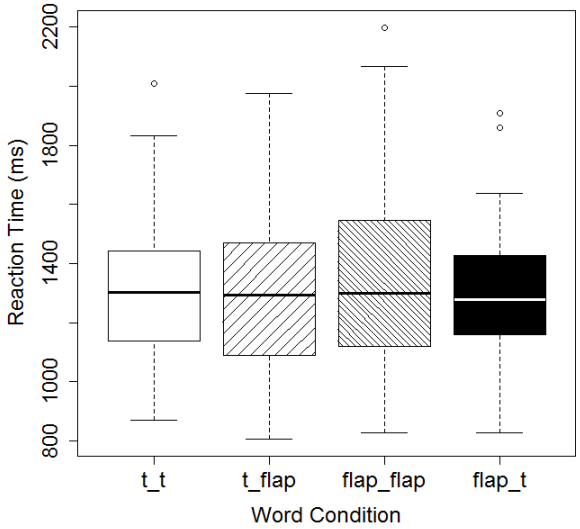

Figure 2: *Reaction time across the word condition described in Table 1.*

3. Discussion

3.1. Canonical form vs. frequency of exposure

The present study investigated possible effects of allophonic variations in the recognition of spoken English words by Japanese learners of English. If JE learners are sensitive to the frequency of occurrence of allophonic variants, then they should recognize *city* pronounced with a flap more accurately and quickly than *city* pronounced with [t] (frequency-of-exposure hypothesis). However, if JE learners instead prefer canonical forms or dictionary pronunciations, then they should recognize *city* pronounced with [t] more accurately and quickly than *city* pronounced with a flap (canonical-form-dominance hypothesis).

Results from the lexical decision task in the present study indicated that JE learners identified flap_t words, i.e. *city* with [t], more accurately than flap_flap words, i.e. *city* with a flap; the former is a canonical form while the latter is a frequent form in AE. JE learners were also more accurate at identifying t_t words, i.e. *faster* with [t], more accurately than flap_flap words, i.e. *city* with a flap; both of these are frequent forms in AE but only the former is a canonical form. These results together are in line with the canonical-form-dominance hypothesis. That is, JE learners are more accurate at recognizing AE words when they do not contain context-sensitive allophonic variations, but are rather produced in a canonical form.

Contrary to the accuracy data, reaction time data did not show significant differences across the word conditions. Reaction times were generally much longer in the present study than in Pitt *et al*.'s study [12] with native AE speakers. These results suggest that lexical decision takes longer in L2 than in L1, and does not systematically vary by whether a word was produced in a canonical form or in a frequent form.

Results from the present study with JE learners are in sharp contrast to results reported in Pitt *et al*.'s study with AE speakers, which reported that AE speakers are more accurate and quicker at recognizing words that are produced with appropriate allophonic variations, supporting the frequency-of-exposure hypothesis. Thus, the frequency-of-exposure hypothesis, which applied to native speakers, is not directly applicable to L2 learners. A possible backdrop for this discrepancy is that the input to average Japanese learners of English is highly supported by the word spellings because of the shortage of oral input. The English teaching guidelines in Japan do not clearly state that American English is the only target variety to be taught. However, all the government authorized textbooks are based on American English, so this variety of English is the most likely variety to be acquired by learners. Traditionally, the English education system has focused on reading and writing abilities more than listening and speaking abilities. Practice on listening and speaking abilities in class has been extremely limited, which has been a criticism for the English teaching system in Japan. Although the most recent updates on the English teaching guidelines [23] place greater emphasis on oral communicative skills in classroom teaching than ever, reading comprehension is still a dominant element of English teaching and much less time is spent on oral input, compared to the English education systems in other countries. Furthermore, even if more time is spent on oral skills in English classes, flap forms are hardly attested in the L2 English produced by English teachers in Japanese schools. JE learners therefore are not sufficiently exposed to either the frequent form, e.g. *city* with a flap, or to the canonical form, e.g. *city* with [t]. As a result, average Japanese leaners of English tend to produce the canonical form which is also supported by the orthographic form. Another possible interpretation might be like this: the input to learners may represent the frequency distribution of the target language to some extent, but the sound in question (alveolar flap) may not be recognized by learners as an allophone of /t/. In other words, the phonological structure

of L2 may not be fully acquired yet, and learners' lexical representation of L2 words may be built upon L1 phonology.

3.2. Production vs. perception

Let us now compare between JE learners' production and perception of AE alveolar flaps. The production experiment with JE learners [8] indicates that JE learners with some experience living in North America, even those with just two weeks of study-abroad in the US, *can* produce alveolar flaps. However, the perceptual identification experiment in Kitahara *et al*.'s study [9] and the lexical decision experiment in the present study suggest that JE learners do not accurately recognize words produced with alveolar flaps. These results together suggest that production skills may precede perception skills, at least when JE learners try to learn alveolar flaps in AE. The finding that production precedes perception in L2 speech learners is consistent with some studies such as those by Sheldon and Strange [24], but is inconsistent with other studies such as those by Bradlow *et al*. [1] who found an opposite trend.

The reason that production precedes perception in the acquisition of AE alveolar flaps by JE learners may be the following. Because Japanese speakers produce alveolar flaps as a phonetic realization of the liquid consonant in Japanese, JE learners may be able to utilize the same set of articulatory movements to produce alveolar flaps in English as well. However, at the phonological level, alveolar flaps and [t] belong to distinct phonemes in Japanese, while they belong to the same phoneme in English. Thus, JE learners need to restructure the sound-phoneme associations when listening to English. That is, for accurate perception, JE learners need to learn the phonological structure of English at a more abstract level. It is possible that perception may take more time than production because perception requires phonological restructuring while production does not necessarily.

Of course, the above explanation might not go beyond speculation since the perception data and production data in question are not directly comparable. The participants for this study as well as [9] only have limited experience in AE. We plan to conduct the same perception experiment with advanced JE leaners who have more extensive exposure to AE through residence in US for a longer term, in order to test if our explanation is valid.

4. Conclusions

In this paper, we have examined the perception of alveolar flaps and stops by non-advanced Japanese learners of English. The word accuracy data in a lexical decision experiment suggest that JE learners prefer stops (i.e. canonical forms) than flaps (i.e. allophonic variants). However, we have reported elsewhere that learners with some experience living in North America *can* produce alveolar flaps. There seems to be an apparent inconsistency between perception and production. How do learners develop their ability to produce flaps even though they might have a hard time perceiving them? Our next step would be to tap further into JE learners' perception to examine the extent to which advanced learners have acquired the phonology of L2. To test this, phoneme-monitoring and or phonological priming tasks might be necessary. Placing this issue in a bigger picture, our question might be rephrased as this: can learners alter their *phonology* after the critical period? Further research is needed to address this and other related questions.

5. Acknowledgements

We thank all the participants of our experiments. We thank Dr. Yasuaki Shinohara (Waseda Univ.) for his dedicated support with statistical analyses. A part of this paper was presented at the 5th joint meeting of ASA and ASJ in Hawaii, 2016. We appreciate valuable comments and suggestions from the participants. This work was supported by JSPS-Kakenhi grant # 16K02646, 15K02492, and 26370508.

6. References

[1] A. R. Bradlow, D. B. Pisoni, R. Akahane-Yamada and Y. Tohkura, "Training Japanese listeners to Training Japanese listeners to identify English/r/and/l: IV. Some effects of perceptual learning on speech production identify English/r/and/l: IV. Some effects of perceptual learning on speech production," *The Journal of the Acoustical Society of America*, vol. 101, no. 4, pp. 2299-2310, 1997.

[2] S. Kawakami, *Nihongo Onsei Gaisetsu*, Tokyo: Oufusha, 1977 (in Japanese).

[3] Y. Amanuma, K. Otsubo, and O. Mizutani, *Nihongo Onseigaku*, Tokyo: Kuroshio Publisher, 1982 (in Japanese).

[4] T. Arai, "A case study of spontaneous speech in Japanese," in *ICPhS 14 – 14th International Congress of Phonetic Sciences, August 1-7, San Francisco, CA, USA, Proceedings*, 1999, pp. 615-618.

[5] T. Arai, N. Warner, and S. Greenberg, "Analysis of spontaneous Japanese in a multi-language telephone-speech corpus," *Acoustical Science and technology*, vol. 28, no. 1, pp. 46-48, 2007.

[6] T. Arai, "Acoustic characteristics of Japanese /r/ sounds and errors in children's speech," *Annual Spring Meeting of the Acoustical Society of Japan, March 13-15, Tokyo University of Technology, Tokyo, Japan, Proceedings*, 2013, pp. 349-352.

[7] N. Minematsu, K. Tomiyama, K. Yoshimoto, K. Shimizu, S. Nakagawa, M. Dantsuji, and S. Makino, "English Speech Database Read by Japanese Learners for CALL System Development," *LREC 2002 – 3rd International Conference on Language Resources and Evaluation, May 29-30, Las Palmas, Canary Islands, Spain, Proceedings*, 2002, pp. 896-903.

[8] K. Tajima, M. Kitahara, and K. Yoneyama, "Production of an Allophonic Variant in a Second Language: The Case of Intervocalic Alveolar Flapping," *JELS*, vol. *32*, 2015, pp.139-145.

[9] M. Kitahara, K. Tajima, and K. Yoneyama, "Perception of flaps, stops, and liquids by Japanese learners of English," *29th meeting of Phonetic Society of Japan, October 3-4, Kobe University, Kobe, Japan, Proceedings,* 2015, pp. 74-79.

[10] C. M. Connine, "It's not what you hear but how often you hear it: On the neglected role of phonological variant frequency in auditory word recognition," *Psychonomic Bulletin and Review*, vol. 11, no. 6, pp. 1084–1089, 2004.

[11] C. M. Connine, L. J. Ranbom, and D. J. Patterson, "Processing variant forms in spoken word recognition: The role of variant frequency," *Perception & Psychophysics,* vol. 70, no. 3, pp. 403–411, 2008.

[12] M. A. Pitt, L. Dilley, and M. Tat, "Exploring the role of exposure frequency in recognizing pronunciation variants," *Journal of Phonetics*, vol. 39, pp. 304-311, 2011.

[13] J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," *IEEE ICASSP-92 – IEEE International Conference on Acoustics, Speech, & Signal Processing, March 23-26, San Francisco, CA, USA, Proceedings*; 1992, pp. 517–520.

[14] H. Mitterer and M. Ernestus, "Listeners recover /t/s that speakers reduce: Evidence from /t/-lenition in Dutch," *Journal of Phonetics*, vol. 34, pp. 73-103, 2006.

[15] H. Mitterer and J. M. McQueen, "Foreign subtitles help but native-language subtitles harm foreign speech perception," *PLoS One*, vol. 4, A146-A150, 2009.

[16] M. Ernestus and R. H. Baayen, “The comprehension of acoustically reduced morphologically complex words: The roles of deletion, duration and frequency of occurrence,” *ICPhS 2007 – 16th International Congress of Phonetic Sciences, August 6-10, Saarbrücken, Germany, Proceedings, 2007*, pp. 773-776.
[17] E. Janse, S. G. Nooteboom, and H. Quene, “Coping with gradient forms of /t/-deletion and lexical ambiguity in spoken word recognition,” *Language and Cognitive Processes*, vol. 22, no. 2, pp. 161–200, 2007.
[18] C. T. McLennan, P. A. Luce, and J. Charles-Luce, “Representation of lexical form,” *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 29, pp. 539–553, 2003.
[19] C. T. McLennan, P. A. Luce, and J. Charles-Luce, “Representation of lexical form: Evidence from studies of sublexical ambiguity,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 31, pp. 1308–1314, 2005.
[20] B. V. Tucker and N. Warner, “Inhibition of processing due to reduction of the American English flap,” *ICPhS 2017 – 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, proceedings*, 2007, pp. 1949-1952.
[21] M. A. Pitt, “The strength and time course of lexical activation of pronunciation variants,” *Journal of Experimental Psychology: Human Perception and Performance*, vol. 35, pp. 896- 910, 2009.
[22] Psychology Software Tools, Inc. [E-Prime 2.0]. Retrieved from http://www.pstnet.com, 2012.
[23] Ministry of Education, Culture, Sports, Science and Technology, *Shin Gakushu Shido Yoryo, [New Courses of Study]*, http://www.mext.go.jp/a_menu/shotou/new-cs/1384661.htm, 2017..
[24] A. Sheldon and W. Strange, “The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception,” *Applied Psycholinguistics*, vol. 3, no. 3, pp. 243-261, 1982.