Journal of Phonetics 79 (2020) 100964

Contents lists available at ScienceDirect

Journal of Phonetics

journal homepage: www.elsevier.com/locate/Phonetics

Research Article

The effect of input prompts on the relationship between perception and production of non-native sounds

Misaki Kato *, Melissa Michaud Baese-Berk

Department of Linguistics, 1290 University of Oregon, Eugene, OR 97403, United States

ARTICLE INFO

Article history:
Received 21 December 2018
Received in revised form 21 January 2020
Accepted 29 January 2020
Available online 21 February 2020

Keywords:
Speech perception
Speech production
Second language acquisition

ABSTRACT

One factor known to affect a second language learner's pronunciation accuracy of non-native sounds is their perception accuracy of the same sounds. However, it is not clear how stable the relationship between the two modalities is when production is cued by perception or other input sources, such as orthography, which is also known to affect production of non-native sounds. We examined whether the relationship between perception and production of non-native sounds varies as a result of different types of input prompts (auditory vs. orthographic) for production, and whether this effect of input prompts on the perception-production relationship varies in different non-native sound contrasts, namely, English /ɹ/ vs. /l/ for native Japanese learners of English, and Japanese singleton vs. geminate consonants for native English learners of Japanese. The difference in the type of input prompt for production affected learners' perception-production relationship to a larger extent for the English contrast than for the Japanese contrast. This suggests that one factor that affects the relationship between perception and production of non-native sounds is the type of input prompt for production, and that this effect can vary for different non-native contrasts.

1. Introduction

The relationship between perception and production has been a controversial topic. Previous studies have reported mixed results showing evidence of both association (e.g., Goldinger, 1998; Nielsen, 2011) and dissociation (e.g., Baese-Berk & Samuel, 2016; Baese-Berk, 2019; Bradlow, Pisoni, Akahane-Yamada, & Tohkura, 1997; Schertz, Cho, Lotto, & Warner, 2015) between the two modalities. These studies have demonstrated that perception is one factor that affects production, in general. However, especially for production of non-native sounds, there are other sources that affect production performance. Specifically, orthography is an important source that impacts production patterns of non-native sounds (e.g., Bassetti, 2017; Bassetti, Sokolović-Perović, Mairano, & Cerni, 2018). Further, productions cued by orthography can be different from those cued by auditory stimuli (i.e., through a learner's perception: Davidson, 2010; de Jong, Hao, & Park, 2009; Hao & de Jong, 2016). These studies suggest that different types of input (e.g., auditory vs. orthographic) may affect productions of non-native sounds differently, and change how perception relates to production. However, it is not clear to what extent the difference in input prompts for production affects the relationship between perception and production of non-native sounds. Using two types of non-native contrasts, English /ɹ/ vs. /l/ and Japanese singleton vs. geminate consonants (e.g., /s/-/ss/), we examine the effect of input prompts on the relationship between perception and production of non-native sounds, and whether this effect varies for different non-native sound contrasts.

1.1. Association and dissociation between perception and production

Studies examining the relationship between perception and production have demonstrated mixed results. Several theories suggest a very close connection between speech perception and production systems (e.g., Motor Theory: Liberman & Mattingly, 1985; Direct Realism: Best, 1995; Fowler, 1986). For example, Direct Realism suggests that perception and production are both based on production gestures (i.e., perception and production share representations). Previous experimental studies have provided evidence for this close connection, demonstrating that perceptual exposure facilitates rapid

* Corresponding author.
E-mail address: misaki@uoregon.edu (M. Kato).

https://doi.org/10.1016/j.wocn.2020.100964
0095-4470/

changes to production (e.g., Goldinger, 1998; Houde & Jordan, 1998, 2002; Nielsen, 2011) and that speakers phonetically modify their speech to converge with their partner's speech in conversation (e.g., Pardo, 2006). Dominant models of second language (L2) phonological development, including the Perceptual Assimilation Model (PAM-L2; Best & Tyler, 2007) and the Speech Learning Model (SLM; Flege, 1995), also suggest a close connection between perception and production. PAM-L2 claims that speech perception and production share representations, and thus predicts that improvement in either perception or production of non-native sounds should correlate with the improvement in the other modality. SLM suggests that new, non-native sound categories are made in perception first, and that those same categories will be used in production.

However, other studies suggest some degree of dissociation between perception and production. For example, L2 learners' perception and production accuracy of non-native sounds are sometimes not correlated (Peperkamp & Bouchon, 2011) or only moderately correlated (Flege, MacKay, & Meador, 1999). Another type of evidence comes from training studies for perception and production of non-native sounds (e.g., Bradlow et al., 1997; Hanulíková, Dediu, Fang, Bašnaková, & Huettig, 2012; Kartushina, Hervais-Adelman, Frauenfelder, & Golestani, 2015; Tateishi, 2013; Wong, 2013). These studies have reported that while a learner's perception and production of non-native sounds could be improved effectively with appropriate methods targeting each skill, cross-modality training effects are not equally robust. That is, improvement in production does not necessarily transfer to perception, nor do production training effects necessarily transfer to perception improvements, providing evidence that an L2 learner's production does not necessarily reflect their perception of the same L2 sounds.

One challenge that is common across the studies comparing perception and production of non-native sounds is that these skills are assessed by very different tasks. That is, a learner's perception is often assessed by examining their responses to auditory stimuli (e.g., discrimination: Flege et al., 1999; Kartushina et al., 2015; Shin & Iverson, 2011, forced-choice identification: Schertz et al., 2015), whereas production is assessed by examining a learner's speech that is prompted by combinations of auditory and visual stimuli (e.g., Bradlow et al., 1997; Flege et al., 1999; Hanulíková et al., 2012; Sheldon & Strange, 1982). These variable methods examining perception and production of non-native sounds suggest that the relationship between the two modalities may vary depending on how production is prompted. Specifically, a learner's perception of a contrast and their production of that contrast may correlate with one another more or less strongly depending on how their productions are prompted (e.g., via auditory or orthographic forms). Thus, it is necessary to examine the effect of input prompts on production, particularly how auditory and orthographic prompts may impact production differently.

1.2. The effect of input prompts on non-native sound production

Previous studies have demonstrated that the availability of orthography affects L2 learners' production patterns of non-native sounds (Bassetti & Atkinson, 2015; Bassetti et al., 2018; Bassetti, 2017; Bassetti, Escudero, & Hayes-Harb, 2015). Availability of orthography also results in different production patterns of non-native sounds than auditory stimuli (Davidson, 2010; de Jong et al., 2009; Erdener & Burnham, 2005; Rafat, 2015; Steele, 2005). Particularly, L2 learners' production patterns are different when the productions are cued by orthography compared to when they are cued by auditory stimuli (i.e., when learners are relying on their perception). For example, de Jong and colleagues (2009) demonstrated that the overall accuracy in a mimicry test (where participants repeated non-words after auditory stimuli) was significantly lower than the accuracy in a reading test (where participants read aloud from orthography). That is, orthographic input alone (without auditory input) prompted more accurate production than auditory input. Further, productions of non-native sounds were more accurate when the learners were given both auditory and orthographic prompts compared to when they were given only auditory prompts (Davidson, 2010; Steele, 2005). These studies suggest that the explicit category information indexed by orthography can provide learners with targets for articulatory movements that perception alone does not necessarily provide. That is, while a learner still needs to use auditory perception to create sound-to-letter correspondences that would enable them to decode written information, they could obtain clearer production targets from category information via orthography than via auditory perception alone. In fact, Eckman (2004) suggests that orthography can be an effective tool to train and cue a learner's production of non-native sounds that are perceptually difficult to differentiate.

However, orthographic prompts are not always more helpful than auditory prompts. For example, auditory + orthographic prompts, as opposed to auditory only prompts, increased the number of errors made in productions of L2 Irish words for native Turkish participants (Erdener & Burnham, 2005). Further, Hao and de Jong (2016) demonstrated that the effect of orthographic vs. auditory prompts on learners' productions of non-native sounds can be different for different non-native sounds. They examined imitation (auditory prompt and production), reading aloud (orthographic prompt and production), and identification (perception) performed by native English learners with Mandarin tones and native Korean learners with English stops and fricatives. Native Korean learners' productions of English consonants were more accurate (judged by native English listeners) in reading aloud than imitation, but native English learners' productions showed the opposite pattern. The authors suggested that the difficulty of imitating non-native sounds may depend on the characteristics of the target non-native sounds; imitation of tones may be easier than consonants because tones involve salient acoustic features and less complex articulatory coordination than consonants.

These studies demonstrate that a learner's perception of non-native sounds is not the only source that affects their production of the same sounds; orthography also affects their production patterns. Given these studies, it is possible that the relationship between a learner's perception and production of non-native sounds could vary depending on how their production is assessed. That is, the auditory prompt (via learners' perception) may directly influence their production in some cases but may not in other cases where orthographic prompt is also present. In order to closely examine the perception-production

relationship, it is necessary to take the effect of input prompts into account, and examine how perception relates to production when production is based on different types of input prompts. In the present study, we compare a learner's perception to their performance on different types of production tasks, one involving perception but not orthography and the other involving orthography but not perception, to identify the influence of input prompts cuing production on the perception-production relationship of non-native sounds.

1.3. Similarity of processing involved in perception and production

The previous results demonstrate that the effect of input prompts (auditory vs. orthographic) on non-native sound production is not uniform (e.g., Erdener & Burnham, 2005; Hao & de Jong, 2016). This suggests that the effect of input prompts on the perception-production relationship may also vary for different non-native sounds. That is, the type of input prompt for production may influence how perception relates to production to different degrees depending on how the two modalities are related to one another. While one way to characterize the perception-production relationship is that these two systems utilize the same sound representations (e.g., Direct Realism: Best, 1995; Fowler, 1986), other studies suggest otherwise. Specifically, they suggest that perception and production are of a different nature from one another, and different strategies and mechanisms may be involved in the two modalities. For example, speech production can be described as a goal-oriented process (Houde & Jordan, 1998; Redford, 2015). Just as humans control their limbs to grasp an object, they control their articulators in order to produce an intended acoustic pattern. In this sense, accurate sound production requires that speakers control their articulators, using their internal representation of the sound to be produced, to generate the intended acoustic output. In this process, speakers have the control over the process of executing their intended articulatory movements. In contrast, speech sound perception requires listeners to cope with variability in acoustics (e.g., Diehl & Kluender, 1989; Stevens & Blumstein, 1981). Although perception still requires listeners' attentional control and is "active" in this sense (Heald & Nusbaum, 2014), the process relies not only on the listeners but also on others (i.e., the speakers). These studies suggest that speech production and perception may not utilize the same type of control during processing. That is, while perception involves a bottom-up process coping with variability in acoustics, production involves a top-down, goal-oriented process to produce the targeted articulatory movements. Thus, ease or difficulty of speech sound perception and production can be affected by different factors.

Assuming that perception is affected by acoustic characteristics and production involves articulatory control, the similarity of the cues that are critical in perception and production could vary across different sound contrasts. That is, cues that listeners rely on in perception and cues that speakers rely on in production may be more similar for one non-native sound contrast than for another. For example, for the English /ɹ/-/l/ contrast, what a learner must be able to do is quite different in perception and production. In production, speakers need to change their articulation (i.e., tongue position; Flege, Takagi, & Mann, 1995), because English /ɹ/ is a dorsal or retroflexed approximant and English /l/ is a lateral continuant. However, in perception, listeners need to detect the change in the relationship between formant frequencies, particularly the initial state and trajectory of the third formant (Miyawaki et al., 1975; Sheldon & Strange, 1982). On the other hand, for the Japanese singleton-geminate consonant contrast, duration is a primary cue in both perception and production[1]. That is, in production, the duration of the consonant (e.g., duration of stop closure for /t/-/tt/; duration of frication for /s/-/ss/) must be different between singleton and geminate consonants (Hayes, 2002). Similarly, in perception, listeners need to be sensitive to the duration of these consonants (Hirata, 2004). Thus, perception and production processing for the Japanese contrast, which relies on duration as a primary cue in both modalities, may be more similar to one another than those for the English contrast, which relies on the formant structure in perception vs. tongue movement in production.

These studies have suggested that the similarity of processing involved in perception and production could vary for different non-native sound contrasts. Given the variable effects of input prompts (auditory vs. orthographic) on non-native sound production (e.g., Davidson, 2010; de Jong et al., 2009; Erdener & Burnham, 2005; Hao & de Jong, 2016), a critical question is whether the effect of input prompts on the perception-production relationship differs depending on the similarity of the processing involved in the two modalities. It is possible that the prompt type (auditory vs. orthographic) impacts the perception-production relationship to a larger degree for a non-native sound contrast where the two modalities involve dissimilar processes (i.e., speaker and listener rely on different cues) than for another contrast where the two modalities involve similar processes (i.e., speaker and listener rely on similar cues). In other words, if a learner relies on similar cues in perception and production of non-native sounds (e.g., in the Japanese singleton-geminate consonant contrast), there may be little room for orthography to impact productions, resulting in little effect of the prompt type on the perception-production relationship. However, if a learner relies on different cues in the two modalities (e.g., the English /ɹ/-/l/ contrast), orthography may influence their productions differently than auditory prompts (i.e., via their perception), resulting in diverging patterns for the perception-production relationship depending on the form of the prompt for production.

1.4. Goals of the current study and hypotheses

In the current study, we examine the effect of input prompts on the relationship between perception and production of non-native sounds, and whether this effect varies depending on the similarity of processing involved in perception and production. We test this question with two types of non-native contrasts: English /ɹ/-/l/ and Japanese singleton-geminate consonant contrast. The English /ɹ/-/l/ contrast is difficult for native Japanese learners to perceive and produce (e.g., Cutler, Weber, & Otake, 2006; Goto, 1971; Yamada, 1995), as is the Japanese singleton-geminate contrast for native English learners (Han,

[1] Though do note that some work suggests that these differences in duration may actually be by-products of other articularly parameters (e.g., tenseness, stiffness, etc.; Parrell (2011)).

1992; Hirata, 2004). However, as discussed above, these non-native contrasts differ in terms of the similarity of the factors that are critical in perception and production. We hypothesize that the input prompts (auditory vs. orthographic) cuing production will modify how perception relates to production to a larger degree for the English /ɹ/-/l/ contrast, where critical factors in perception and production are more dissimilar, compared to the Japanese singleton-geminate consonant contrast, where the critical factors in the two modalities are more similar. If this is the case, it is also possible that these differences may manifest in distinct ways when presented with differing input. That is, if perception and production are dissociated for an individual, and a listener were to simply hear a word, they may not be able to accurately perceive it, and thus may not be able to accurately produce it, even if they are capable of producing the sound in general. However, if they are presented with orthography and asked to read the word allowed, one might expect that they would be able to produce the sound because they do not need to rely on their less-accurate perception.

The current study tests this hypothesis by comparing L2 learners' perception and production of the English and Japanese contrasts. Production is examined in conditions with different combinations of input (i.e., auditory and orthographic), where learners produce words in either imitation (using auditory prompts) or reading aloud (using orthographic prompts; Hao & de Jong, 2016). Learners' perception is tested in a two-alternative forced choice (2AFC) task. This perception task is used to examine learners' perception separately from their production, as well as to confirm that the target non-native contrasts (i.e., English /ɹ/-/l/ contrast and Japanese singleton-geminate contrast) are difficult contrasts for the learners to perceive as compared to other contrasts in the target language. We compare learners' productions between the two prompt conditions, and also examine how their perception relates to production in the two prompt conditions. We predict that the extent to which learners' orthography-based production deviates from their auditory-based production and from their perception will be larger for the English contrast than for the Japanese contrast. Specifically, for the English contrast, their production accuracy with orthographic prompts (i.e., when learners are given explicit category information) will be different from their production accuracy with auditory prompts (i.e., when learners rely on their perception) and from their perception accuracy in the 2AFC task. On the other hand, for the Japanese contrast, learners' production accuracy with orthographic prompts will be similar to their production accuracy with auditory prompts. Furthermore, their perception accuracy will relate to their production accuracy with auditory prompts and production accuracy with orthographic prompts to a similar extent.

In the following sections, we present two experiments: production and perception of the English /ɹ/-/l/ contrast (Section 2), and production and perception of the Japanese singleton-geminate consonant contrast (Section 3). In Section 4, we compare the results of the two experiments in order to examine how a learner's perception relates to their production in the two prompt conditions (auditory vs. orthographic), and whether the degree to which the prompt type affects the perception-production relationship differs for the two non-native contrasts.

2. Experiment 1: English /ɹ/-/l/ contrast

In this experiment, we examine production and perception accuracy of the English /ɹ/-/l/ contrast. Specifically, we examine whether native Japanese learners' production accuracy of the English /ɹ/-/l/ contrast changes depending on the type of input prompt they receive (auditory or orthographic). In order to confirm that the English /ɹ/-/l/ contrast is a difficult contrast to perceive for the native Japanese learners, we also examine learners' perception accuracy of the English /ɹ/-/l/ contrast as compared to other English contrasts.

2.1. Methods

2.1.1. Participants

Fifteen native Japanese learners of English (11 females) and 13 native speakers of English (6 females) participated. All participants were between the ages of 19 and 26. Native Japanese learners were students at the American English Institute (AEI) at the University of Oregon, which provides academic English support for international students. While these native Japanese learners started studying English at a similar age (from about the age of 12 in junior high school) in Japan, their length of stay in English-speaking countries varied (average = 14.4 months, range = 2–87 months). Native English speakers were recruited from the Psychology and Linguistics subject pool at the University of Oregon. All the participants received partial course credit for their participation. None of the participants reported any history of a speech or hearing impairment.

2.1.2. Stimuli

The stimuli consisted of 160 items: 80 target items (with the /ɹ/-/l/ contrast) and 80 distractor items (without the /ɹ/-/l/ contrast). Target items consisted of 20 real words and 60 nonwords (see Appendix A for the list of target items). Similarly, distractor items consisted of 20 real words and 60 nonwords. The distractor items were also minimal pairs. The sound contrasts that native Japanese speakers do not usually have difficulty with were embedded in the distractor items (e.g., *fenson/senson*). Following Strange and Dittmann (1984), the /ɹ/-/l/ sound contrast was embedded in multiple phonetic environments in the target items. Thirty-two minimal pairs (64 items) contrasted /ɹ/ and /l/ in four phonetic environments: onset singleton (e.g., *renk/lenk*), onset cluster (e.g., *brize/blize*), intervocalic (e.g., *neron/nelon*), and coda (e.g., *nare/nel*), and 8 pairs (16 items) contained both /ɹ/ and /l/ (e.g., *lorief/rolief*). Here, we analyzed only the items that contained either /ɹ/ or /l/, and not those that contained both, to simplify the analysis (e.g., examine one target consonant per item) and to have the item design comparable to Experiment 2; Experiment 2 did not include items that contained both singleton and geminate consonants. All the items were embedded in the carrier phrase, *"It says ___ here".* In all cases, the prosodic context was controlled, as target words received narrow focus and were likely to be produced with a nuclear pitch accent. A male native American English speaker provided the recordings of all the items in the carrier phrase to create the auditory stimuli.

2.1.3. Production and perception experiment procedure

The experiment consisted of two tasks: a production and a perception task. We conducted the production task first to prevent participants from being familiarized with the auditory forms of the items at the time of production. Because of this order of the two tasks, in the perception task, the participants had been familiarized with the minimal pairs that were presented via auditory prompts in the production task. Both production and perception tasks were conducted via E-Prime software (Schneider, Eschman, & Zuccolotto, 2002), using Sennheiser HD 202 II headphones. Auditory stimuli were presented at a comfortable listening level.

In the production task, participants produced the items in the carrier phrase in four conditions: Baseline, See-Say, Hear-Say, and Hear-Delay-Say conditions. Table 1 shows the production task procedure in the four different conditions. In the Baseline condition, participants saw a real word item in the carrier phrase (*"It says ___ here")* shown on the computer screen, heard the phrase, and repeated the phrase twice. This condition was intended to elicit participants' best possible productions of the target non-native sounds, by supporting their productions with multiple forms of input (orthographic and auditory). If participants have any articulatory difficulty producing the /ɹ/-/l/ contrast, it should manifest in the Baseline condition. Because real-word items were used only in the Baseline condition, we limited our analysis to the production data in the other three conditions that involved non-words, See-Say, Hear-Delay-Say, and Hear-Say conditions, in order to examine the effect of input prompts (auditory vs. orthographic) without lexical processing (see the results section for more details about how these conditions were compared to one another).

In the See-Say condition, only the orthographic prompt was available; participants saw a non-word item in the carrier phrase shown on the screen and read it aloud twice. In the Hear-Say and Hear-Delay-Say conditions, only the auditory prompt was available. In the Hear-Say condition, participants first heard a female (speaker 1) say, "*What does it say here?"*, then heard the target phrase in the male voice (speaker 2), and immediately repeated the phrase twice. However, production in this condition may not necessarily reflect participants' perceptual phonological categories because it is possible for them to simply imitate an acoustic exemplar of what they hear (Hao & de Jong, 2016). That is, they may not need sub-lexical representations (e.g., Ramus et al., 2010) to perform well in immediate repetition. Therefore, we included another condition with auditory-only prompts, the Hear-Delay-Say condition, where there was a short delay after participants heard the auditory prompt and before they were asked to repeat. The short delay was included in order to disrupt the effect of short-term echoic memory of the auditory prompt (e.g., Cowan, 1984; Goldinger, 1996) and to encourage the participants to use sub-lexical representations for repetition. Specifically, in this condition, participants first heard a female (speaker 1) say, *"What does it say here?"*, then heard the target phrase in the male voice (speaker 2), and then heard the same female (speaker 1) say, *"I'm sorry, could you say it again?"*. After this cue, participants repeated the target phrase twice. The participants who can imitate well but do not have established acoustic categories (i.e., those who do not necessarily know what they are producing) may perform better in the Hear-Say condition than in the Hear-Delay-Say condition.

The Baseline production block was always done first to familiarize participants with the experiment in a condition that gave them the most support for their productions (i.e., the condition with real English words with both orthographic and auditory support). The other three conditions were blocked and the order of the blocks was randomized for each participant. In order to ensure that items were not produced in the same condition, participants were divided into three groups. Everyone produced all the items (80 target items and 80 distractor items), but the condition in which they produced the items was different depending on the group. For example, a non-word item *ranby* was produced in the See-Say condition by a participant in one group, in the Hear-Say condition by a participant in a second group, and in the Hear-Delay-Say condition by a participant in the third group. At the beginning of each block, participants completed two practice trials with the distractor items. The experimenter was present in the testing room during the two practice trials to ensure that the participant understood the instructions.

Following the production task, the perception 2AFC task tested participants' abilities to perceptually discriminate between /ɹ/ and /l/ and choose the correct word from two options. Only the non-word items (60 target and 60 distractor non-words) were used in the perception task to examine their perception accuracy without any lexical knowledge. Participants first heard a recording of a phrase from the same male American English speaker that they heard in the Hear-Say and Hear-Delay-Say conditions in the production task, *"It says ___ here"*. They were simultaneously presented with two choices of /ɹ/-/l/ minimal pairs (e.g., *ranby/lanby*) in the target trials and other minimal pairs (e.g., *fenson/senson*) in the distractor trials on the computer screen. Then, they were instructed to choose one of the two choices by pressing A or B on the keyboard. The order of presentation of the items was randomized. After the perception task, participants answered a questionnaire about their language backgrounds.

Table 1
Production task procedure in four different conditions. X indicates the procedure that was not included. In all conditions, participants repeat the target word after presentation of the word in a carrier phrase. In the Baseline condition, participants both see an orthographic presentation of the word and hear the word auditorily. In the See-Say Condition, participants see the orthographic presentation of the word, but do not hear it. In the Hear-Delay-Say condition, participants hear the target token and hear an intervening phrase before producing the word. In the Hear-Say condition, participants hear the target token and immediately produce the word.

Participant's action	Baseline	See-Say	Hear-Say	Hear-Delay-Say
see orthography of	"It says REAL-WORD here"	"It says NON-WORD here"	X	X
hear speaker 1	X	X	"What does it say here"	"What does it say here?"
hear speaker 2	"It says REAL-WORD here"	X	"It says NON-WORD here"	"It says NON-WORD here"
hear speaker 1	X	X	X	"I'm sorry, could you say it again?"
repeat the phrase twice	"It says REAL-WORD here"	"It says NON-WORD here"	"It says NON-WORD here"	"It says NON-WORD here"

2.1.4. Acoustic analysis

Native English speakers' and native Japanese learners' productions were evaluated in acoustic analysis. Acoustic measurements were made for all the target items except for items that contained both /ɹ/ and /l/. Therefore, there were 64 target items for each speaker. Acoustic differences between /ɹ/ and /l/ are primarily and sufficiently manifested in the steady state and transition of the third formant (F3; Miyawaki et al., 1975; Yamada & Tohkura, 1992). Specifically, F3 values are generally lower for /ɹ/ than for /l/ (Lotto, Sato, & Diehl, 2004), and native English speakers rely dominantly on this cue (Miyawaki et al., 1975). However, native Japanese speakers are known to rely heavily on F2 as well (Iverson et al., 2003). Thus, we measured F1, F2, and F3 for /ɹ/ and /l/.

Using Praat (Boersma & Weenik, 2015), the steady state of /ɹ/ or /l/ was segmented, and average F1, F2, and F3 values from the segmented portion were collected. Following the segmentation criteria described in Flege et al. (1995), the duration of a steady state portion of /ɹ/ and /l/ was determined. In the onset position (e.g., *read/lead*), the release of lingual constriction was measured for /l/ where an increase in energy in F2 was obvious. When there was no release of lingual constriction for /ɹ/, the steady state was from the beginning of energy for the first three formants in the waveform until the point where F3 frequency started rising. In the onset cluster position (e.g., *broom/bloom*), the steady state was defined as the point of an increase in energy in F2 and F3. In the intervocalic position, the measurements were made where F2 rapidly dropped for /l/ and where F3 dropped for /ɹ/. In the coda position, the measurements were made where F2 rapidly dropped for /l/ and where F3 also dropped for /ɹ/ until the beginning of frication of *here* (as in the carrier phrase "*It says ___ here*") in the waveform.

Using these segmentation criteria, average F1, F2, and F3 values from the segmented portion were collected by the first author. The average formant values were calculated from points that were 6.25 milliseconds apart in the segmented regions. In addition to the three formant values, we also examined average F3 values divided by average F2 values. Of all measures collected, native English speakers produced the largest difference between /ɹ/ and /l/ for mean F3 values (mean F3 difference = 868.96, mean F2 difference = 287.93, mean F3/mean F2 difference = 1.18). Native Japanese speakers also produced the largest differences between /ɹ/ and /l/ for mean F3 values (mean F3 difference = 285.83, mean F2 difference = 80.96, mean F3/mean F3 difference = 0.33). Therefore, we chose F3 values as the indicator of production accuracy (i.e., whether the speakers made differences between their productions of /ɹ/ and /l/). A research assistant segmented 8% of the sound files, and the inter-rater agreement of mean F3 values was $r = 0.95$ ($p < 0.001$).

2.2. Results

2.2.1. Production accuracy of English /ɹ/-/l/ contrast

In this analysis, our primary interest was whether native Japanese learners' production accuracy changed depending on the type of input prompt they received (auditory or orthographic). Fig. 1 shows the mean F3 values by speakers' native language (English or Japanese), condition (Baseline, See-Say, Hear-Delay-Say, or Hear-Say), and consonant (/ɹ/ or /l/). Examining the figure, native Japanese learners generally made a smaller difference in F3 between /ɹ/ and /l/ than native English speakers did. Further, native Japanese learners made a larger difference in F3 in the See-Say condition, where orthographic prompts were available, compared to the conditions where only auditory prompts were available (Hear-Say and Hear-Delay-Say).

These observations were tested using linear mixed-effects regression models using lme4 (Bates, Maechler, Bolker, & Walker, 2015) in R (R Core Team, 2016) with F3 values in the speakers' productions as the dependent variable. The fixed factors were consonant (/ɹ/ or /l/), two-way interaction between consonant and speakers' native language background (English or Japanese), two-way interaction between consonant and production condition (Baseline, See-Say, Hear-Delay-Say, and Hear-Say), and three-way interaction among consonant, speakers' native language background, and production condition. Speakers' native language background was contrast coded such that positive beta values were associated with native English speakers (0.5) and negative values were associated with native Japanese learners (−0.5). Consonant was also contrast coded to compare between items with /l/ (0.5) and items with /ɹ/ (−0.5). Production condition was contrast coded to compare between Baseline and See-Say (0.5, −0.5, 0, 0), between Hear-Delay-Say and Hear-Say (0, 0, 0.5, −0.5), and between See-Say and auditory conditions (Hear-Delay-Say and Hear-Say; 0, 0.5, −0.25, −0.25). The maximal random effects structure that would converge was implemented, which included random intercepts for speakers and items. The random effects structure also included random slope for speakers by the interaction between consonant and production condition. Significance was determined using model comparisons to examine whether each fixed factor

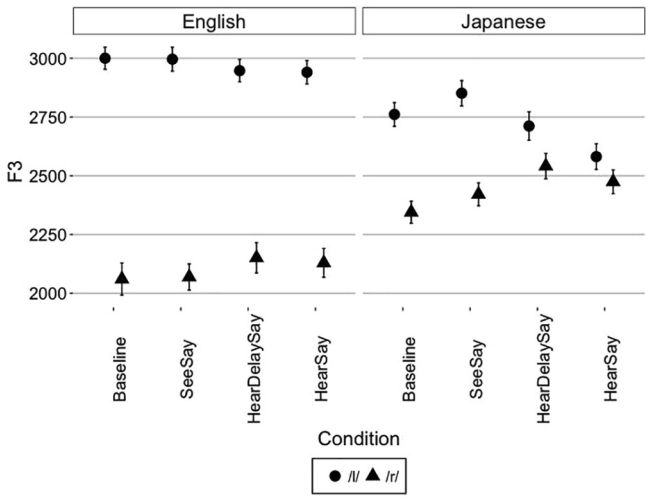

Fig. 1. Mean F3 values by speakers' native language background, condition, consonant. The error bars represent the 95% confidence interval. In all conditions, participants repeat the target word after presentation of the word in a carrier phrase. In the Baseline condition, participants both see an orthographic presentation of the word and hear the word auditorily. In the SeeSay Condition, participants see the orthographic presentation of the word, but do not hear it. In the HearDelaySay condition, participants hear the target token and hear an intervening phrase before producing the word. In the HearSay condition, participants hear the target token and immediately produce the word.

contributed significantly to the model fit. The full model is summarized along with its R syntax in the Appendix B.

The results revealed a significant main effect of consonant (/ɹ/ or /l/; χ^2 (1) = 177.92, $p < 0.001$), indicating that F3 values were higher for productions of /l/ than for /ɹ/ in general. The two-way interaction between speakers' native language (English or Japanese) and consonant was also a significant predictor of the model fit (χ^2 (1) = 545.58, $p < 0.001$), indicating that native English speakers made a larger F3 difference between /ɹ/ and /l/ than native Japanese learners. In terms of the effect of the production condition, the two-way interaction between the See-Say vs. auditory prompt (Hear-Delay-Say and Hear-Say) condition term and consonant (/ɹ/ or /l/) was a significant predictor (χ^2 (1) = 24.57, $p < 0.001$), and so was the three-way interaction among these two factors and speakers' native language (χ^2 (1) = 4.69, $p < 0.05$), indicating that the F3 difference made between /ɹ/ and /l/ was generally larger in the See-Say than in the auditory prompt conditions, and this effect was larger for native Japanese learners' productions than for native English speakers'. The two-way interaction between the Hear-Delay-Say vs. Hear-Say condition term and consonant was not a significant predictor (χ^2 (1) = 0.21, $p = 0.65$), and neither was the three-way interaction among these two factors and speakers' native language (χ^2 (1) = 0.49, $p = 0.48$), indicating that speakers in both language groups made a similar amount of F3 difference between /ɹ/ and /l/ in the two auditory prompt conditions. The two-way interaction between Baseline vs. See-Say condition term and consonant was a significant predictor (χ^2 (1) = 10.05, $p < 0.01$), but the three-way interaction among these two factors and speakers' native language was not (χ^2 (1) = 2.12, $p = 0.15$), indicating that the amount of F3 difference made between /ɹ/ and /l/ was different between the Baseline and See-Say conditions in productions for both language groups.

2.2.2. Perception accuracy of English /ɹ/-/l/ contrast

The purpose of this analysis was to examine native Japanese learners' perception accuracy of the English /ɹ/-/l/ contrast separately from the production task, and to examine whether perceiving the English /ɹ/-/l/ contrast was more difficult than other English contrasts for the learners. Fig. 2 shows the mean proportion correct on the perception task by listeners' native language and consonant. The figure shows that native English listeners' perception was generally more accurate than native Japanese learners' perception. Native Japanese learners' perception of the target items was less accurate than the distractor items, but native English listeners showed a much smaller difference between perception accuracy of the distractor and target items. Further, while native Japanese learners seemed to be biased towards the items with /ɹ/, there was no such tendency for native English listeners.

The perception data were analyzed using mixed-effects logistic regression models with listeners' accuracy in the perception task as the dependent variable (i.e., correct or incorrect). The fixed factors were listeners' native language background (English or Japanese), consonant (distractor, /ɹ/, or /l/), and the interaction between these two factors. Listeners' native language was contrast coded such that positive beta values were associated with native English listeners (0.5) and negative values were associated with native Japanese learners (−0.5). Consonant was also contrast coded to compare between distractor and target items (0.5, −0.25, −0.25) and between items with /ɹ/ and /l/ (0, 0.5, −0.5). The maximal random effects structure allowing the model to converge was included in the model, and included random intercepts for listeners and items, as well as random slope for listeners by item type (distractor vs. target items). Significance was determined via model comparison as described above. The full model is summarized along with its R syntax in the Appendix B.

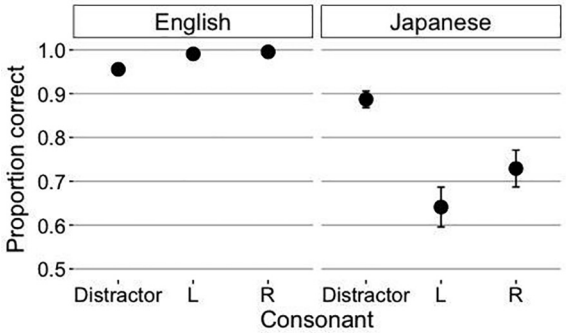

Fig. 2. Mean proportion correct on the perception task by listeners' native language background and consonant. The error bars represent the 95% confidence interval. Distractor words did not contain an /ɹ/, or /l/. Accuracy on /l/ tokens is lower than on /ɹ/ for native Japanese learners.

The results revealed that listeners' native language was a significant predictor of the model fit (χ^2 (1) = 49.34, $p < 0.001$), indicating that native English listeners' perception accuracy was higher than native Japanese learners'. The main effect of item type (distractor or target items) was not a significant predictor (χ^2 (1) = 0.45, $p = 0.5$), but its interaction with listeners' native language was (χ^2 (1) = 46.61, $p < 0.001$), indicating that the effect of item type (distractor or target items) was larger in native Japanese learners' perception than in native English listeners'. Finally, while the proportion correct for the items with /ɹ/ seemed higher than the items with /l/ in native Japanese learners' perception, consonant (/ɹ/ or /l/) was not a significant predictor (χ^2 (1) = 2.8, $p = 0.09$), and neither was its interaction with listeners' native language (χ^2 (1) = 0.08, $p = 0.77$), indicating that listeners' perception accuracy did not significantly differ for items that contained /ɹ/ or /l/.

2.3. Summary of Experiment 1

Overall, the English /ɹ/-/l/ production results demonstrated that the speakers in both language groups generally made a larger F3 difference between /ɹ/ and /l/ in the orthographic prompt condition (See-Say) than in the auditory prompt conditions (Hear-Say and Hear-Delay-Say). However, this effect was larger for native Japanese learners' productions than for native English speakers' productions. Thus, native Japanese learners' production accuracy of the English /ɹ/-/l/ contrast was higher when given the orthographic prompts than when given the auditory prompts. This suggests that native Japanese learners were able to produce the English /ɹ/-/l/ contrast accurately given the explicit category information (i.e., orthographic prompt) but did not demonstrate the same production accuracy when they had to rely only on their perception in the auditory prompt conditions.

In perception, native Japanese learners' perception of the English /ɹ/-/l/ contrast was less accurate than that of native English listeners. Further, the learners' perception of the English /ɹ/-/l/ contrast was less accurate than their perception of other English contrasts. The learners also showed a slight bias towards perception of /ɹ/ than /l/.

3. Experiment 2: Japanese singleton-geminate consonant contrast

In this experiment, we examine production and perception accuracy of the Japanese singleton-geminate consonant contrast. Specifically, we examine whether native English learners' production accuracy of the Japanese singleton-geminate consonant contrast changes depending on the type of prompt they receive (auditory or orthographic). Also, in order to confirm that the Japanese singleton-geminate contrast is a difficult contrast to perceive for the native English learners, we also examine learners' perception accuracy of the Japanese singleton-geminate consonant contrast as compared to other Japanese contrasts. We test these questions in two types of consonants: fricative (/s/-/ss/) and stop (/t/-/tt/). Previous studies have shown that native English learners of Japanese have more difficulty perceiving the fricative singleton-geminate contrast (/s/-/ss/) compared to the stop contrast (/t/-/tt/) because the difference in frication duration between /s/ and /ss/ is generally smaller than the difference in stop closure duration between /t/ and /tt/ in native Japanese speakers' productions (Hardison & Saigo, 2010; Hayes, 2002). Given these studies, we predict that perceiving and producing the fricative singleton-geminate consonant contrast will be more difficult than the stop contrast for native English learners. However, we predict that the type of prompt (auditory and orthographic) will affect these consonant types similarly because duration is part of the critical cues for perception and production of the singleton-geminate distinction in both consonant types.

3.1. Methods

3.1.1. Participants

Sixteen native English learners of Japanese (11 females) and 14 native speakers of Japanese (7 females) participated. All participants were between the ages of 19 and 43 (mean: 22.9 years old). Native English learners were students in an intermediate level Japanese class at the University of Oregon. These native English learners had studied Japanese for about 10 years (average = 9.06 years, range = 1–28 years). Five learners reported that they had lived in Japan; the duration ranged from 5 to 9 years; their data were comparable to those of other native English learners, thus these five learners' data were included in the analyses. While one of the native English learners reported a hearing impairment and use of hearing aids, this particular participant's data did not deviate from other native English learners' data, and thus the participant was included in our analyses. Native Japanese speakers were recruited from the AEI at the University of Oregon. All the participants received partial course credit for their participation.

3.1.2. Stimuli

The stimuli consisted of 120 items of disyllabic words: 60 target (with a singleton-geminate contrast) and 60 distractor items (without singleton-geminate contrasts). The list of target items is shown in Appendix A. The 60 target items consisted of items of a fricative contrast /s/-/ss/ (15 minimal pairs; 30 items) and a stop contrast /t/-/tt/ (11 minimal pairs; 22 items). We also included items of an affricate contrast /ts/-/tts/ (4 minimal pairs; 8 items), but these were not included in the current analysis because the number of the affricate items was small. Of the 30 fricative and 22 stop target items, 6 items (3 minimal pairs) in each consonant type were real Japanese word items. These real word items were used in the Baseline condition in the production task. We used two different pitch-accent patterns in the stimuli: high-low (HL) and low-high (LH), given previous studies demonstrating that pitch-accent can impact identification accuracy of Japanese singleton-geminate contrasts (Minagawa & Kiritani, 1996; Tsukada, Cox, Hajek, & Hirata, 2018). All the items were embedded in the carrier phrase, *"korewa ___ to yomimasu" ("This is read ___")*. A male native Japanese speaker provided the recordings of all the items in the carrier phrase to create the auditory stimuli.

3.1.3. Production and perception experiment procedure

The same paradigm from Experiment 1 was used in this experiment. In the production task, the general instruction displayed on the computer screen was in English, but the orthographic prompts of the stimuli were in Japanese. In the Baseline condition, participants saw a real Japanese word item in the carrier phrase (*"korewa ___ to yomimasu" "This is read___.")* shown on the computer screen, heard the phrase, and repeated the phrase twice. While the auditory stimuli used in the Baseline and auditory conditions (Hear-Say and Hear-Delay-Say) contained the information about both segmental (e.g., singleton and geminate consonants) and pitch-accent (HL and LH pitch-accent), the Japanese orthography (used in the Baseline and See-Say conditions) only signals segmental information. Therefore, in order to signal the pitch-accent information visually in the See-Say (with only the orthographic prompt) condition, an arrow was displayed over the target word. An arrow pointing upward signaled the LH pitch-accent and an arrow pointing downward signaled the HL pitch-accent. The orthographic prompts were displayed in the same manner in the Baseline condition.

In the Hear-Say condition, participants first heard a female (speaker 1) say, "*korewa nanto yomimasuka?"* ("*What does it say here?"*), then heard the target phrase in the male voice (speaker 2), and immediately repeated the phrase twice. In the Hear-Delay-Say condition, participants first heard a female (speaker 1) say, "*korewa nanto yomimasuka?*" ("*What does it say here?"*), then heard the target phrase in the male voice (speaker 2), and then heard the same female (speaker 1) say, *"sumimasen, moo ichido itte kudasai."* (*"I'm sorry, could you say it again?"*). After this cue, participants repeated the target phrase twice. Following the production task, participants completed the perception 2AFC task, where we tested their abilities to perceptually discriminate between singleton and geminate consonants. The stimuli were 96 items, which consisted of 48 target non-word items and 48 distractor non-word items that were used in the production task.

3.1.4. Acoustic analysis

Native Japanese speakers' and native English learners' productions were evaluated in acoustic analysis. Acoustic measurements were made for all the target items except for items that contained the affricate contrast. Therefore, there were 56 target items for each speaker. The durations were measured from spectrograms and waveforms using Praat. The first author measured durations of the target words, frication of the intervocalic fricatives, closures of the intervocalic stops, and VOTs of the intervocalic stops. The consonant duration of the intervocalic stops was defined as the duration of the closure and VOT (following Idemaru & Guion-Anderson, 2010). Therefore, the intervocalic stop duration was measured from the offset of the preceding vowel, which was indicated by the last complete periodic cycle in the waveform with reference to the F2 energy, to the onset of the following vowel, which was indicated by the onset of the first complete periodic cycle in the waveform with reference to the F2 energy. Following Idemaru and Guion-Anderson (2010), the duration of the intervocalic fricatives was measured from the left edge of the frication noise to the onset of the following vowel, which was defined as the left zero crossing of the first complete periodic cycle. Following Hirata and Whiton (2005), we defined the normalized duration to be the ratio of the intervocalic consonant duration to the duration of the whole target word. Thus, to calculate the normalized duration of the consonant for each production, we divided the intervocalic consonant duration by the word duration.

3.2. Results

3.2.1. Production accuracy of Japanese singleton-geminate consonant contrasts

In this analysis, we were interested in whether native English learners' production accuracy changed depending on the type of input prompt they received (auditory or orthographic). We used the amount of normalized duration difference between singleton and geminate consonants as an indicator of production accuracy. Fig. 3 shows the mean normalized duration by consonant type (fricative or stop), speakers' native language (Japanese or English), condition (Baseline, See-Say, Hear-Delay-Say, or Hear-Say), and consonant (singleton or geminate). The figure shows that native English learners generally made a smaller difference in normalized duration between singleton and geminate consonants than native Japanese speakers did. Furthermore, for both language groups' productions, the normalized duration difference for the fricative consonants was smaller than for the stop consonants. However, for both language groups' productions, the normalized duration difference did not seem to differ in the See-Say condition compared to the auditory conditions (Hear-Say and Hear-Delay-Say). Additionally, the normalized duration difference, especially for the fricative contrast, seems smaller in the Baseline than in other three conditions for native Japanese and native English learners' productions. It is possible that the lexical access for the real-word items in the Baseline condition, acting on top of phonological processing, made it more difficult to process the stimuli compared to the other three conditions, which involved non-word items with either the auditory or orthographic prompt.

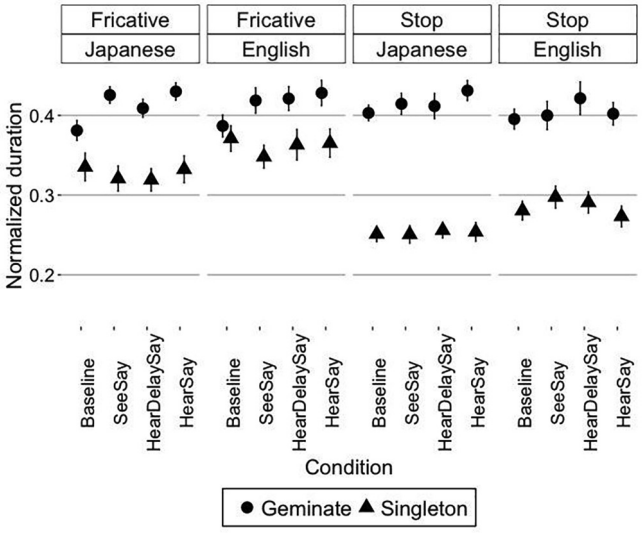

Fig. 3. Mean normalized duration by consonant type, speakers' native language background, condition, and consonant. The error bars represent the 95% confidence interval. In all conditions, participants repeat the target word after presentation of the word in a carrier phrase. In the Baseline condition, participants both see an orthographic presentation of the word and hear the word auditorily. In the SeeSay Condition, participants see the orthographic presentation of the word, but do not hear it. In the HearDelaySay condition, participants hear the target token and hear an intervening phrase before producing the word. In the HearSay condition, participants hear the target token and immediately produce the word.

These observations were tested using linear mixed-effects regression models with normalized duration in the speakers' productions as the dependent variable. The fixed factors were consonant (singleton or geminate), two-way interactions between consonant and speakers' native language background (Japanese or English), between consonant and production condition (Baseline, See-Say, Hear-Delay-Say, and Hear-Say), and between consonant and consonant type (fricative or stop). The fixed factors also included three-way interactions among consonant (singleton or geminate), consonant type (fricative or stop), and speakers' native language, among consonant, consonant type, and production condition, and among consonant, speakers' native language, and production condition. All the fixed factors were contrast coded: speakers' native language background (native Japanese: 0.5, native English: −0.5); consonant (singleton: 0.5, geminate: −0.5); consonant type (stop: 0.5, fricative: −0.5). Production condition was also contrast coded to compare between Baseline and See-Say (0.5, −0.5, 0, 0), between Hear-Delay-Say and Hear-Say (0, 0, 0.5, −0.5), and between See-Say and auditory conditions (Hear-Delay-Say and Hear-Say; 0, 0.5, −0.25, −0.25). The maximal random effects structure that would converge was implemented, which included random intercepts for speakers and items. The random effects structure also included random slope for speakers by the interaction among consonant, consonant type, and production condition. Significance was determined via model comparison as described above. The full model is summarized along with its R syntax in the Appendix B.

The results revealed a significant main effect of consonant (singleton or geminate; χ^2 (1) = 45.3, $p < 0.001$), indicating that normalized duration was shorter for singleton consonants than for geminate consonants in general. The two-way interaction

between speakers' native language (Japanese or English) and consonant was also a significant predictor (χ^2 (1) = 76.98, $p < 0.001$), indicating that native Japanese speakers made a larger normalized duration difference between singleton and geminate consonants than native English learners. The two-way interaction between consonant type (fricative or stop) and consonant was also a significant predictor (χ^2 (1) = 7.97, $p < 0.01$), while the three-way interaction among these two factors and speakers' native language was not (χ^2 (1) = 2.59, $p = 0.11$), indicating that the normalized duration difference between singleton and geminate consonants was larger in stop than fricative consonants for both language groups.

In terms of the effect of the condition, the two-way interaction between the Baseline vs. See-Say condition term and consonant (singleton or geminate) did not significantly improve the model fit (χ^2 (1) = 1.07, $p = 0.3$), indicating that the normalized duration difference made between singleton and geminate consonants was similar in these conditions. This pattern was similar in both language groups and in both consonant types, as the three-way interaction among these two factors and speakers' native language was not a significant predictor (χ^2 (1) = 0.07, $p = 0.8$), nor was the three-way interaction among these two factors and consonant type (χ^2 (1) = 0.53, $p = 0.47$). The comparison between the two auditory prompt conditions revealed similar patterns; the normalized duration difference made between singleton and geminate consonants was similar between the Hear-Delay-Say vs. Hear-Say conditions (χ^2 (1) = 0.72, $p = 0.4$). This pattern was similar in both language groups (χ^2 (1) = 3.51, $p = 0.06$) and in both consonant types (χ^2 (1) = 0.03, $p = 0.87$). Finally, the normalized duration difference made between singleton and geminate consonants was similar between the See-Say vs. auditory prompt (Hear-Delay-Say and Hear-Say) conditions (χ^2 (1) = 0.08, $p = 0.78$). The three-way interaction among the consonant (singleton or geminate), See-Say vs. auditory prompt condition term, and consonant type (fricative or stop) was a significant predictor (χ^2 (1) = 7.24, $p < 0.01$). This indicates that the effect of See-Say vs. auditory prompt conditions was different between the productions of fricative (normalized duration difference in See-Say condition: 0.08; in auditory conditions: 0.07) and stop consonants (normalized duration difference in See-Say condition: 0.13; in auditory conditions: 0.15).

3.2.2. Perception accuracy of Japanese singleton-geminate consonant contrasts

The purpose of this analysis was to examine native English learners' perception accuracy of the Japanese singleton-geminate consonant contrasts separately from the production task, and to examine whether perceiving the Japanese singleton-geminate contrast was more difficult than other Japanese contrasts for the learners. Fig. 4 shows the mean proportion correct on the perception task by listeners' native language (Japanese or English), consonant type (distractor, fricative or stop), and consonant (singleton or geminate). The figure suggests that while native English listeners' perception of the target items (fricative and stop consonants) was less accurate than the distractor items, their perception was generally quite accurate (over 80% correct). Native English learners also showed a bias to choose geminate rather than singleton for the stop contrast.

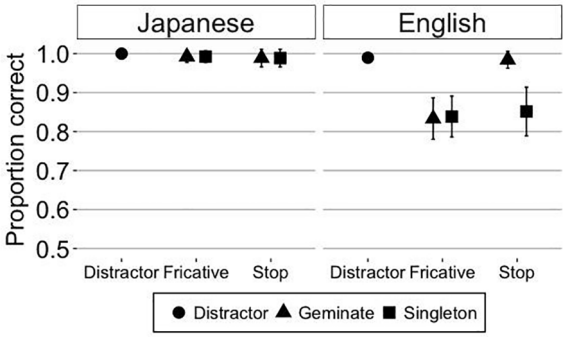

Fig. 4. Mean proportion correct on the perception task by listeners' native language background, consonant type, and consonant. The error bars represent the 95% confidence interval. Distractor items did not contain the singletons or geminates used in the Fricative and Stop conditions.

The perception data were analyzed using mixed-effects logistic regression models with listeners' accuracy in the perception task as the dependent variable (i.e., correct or incorrect). The fixed factors were listeners' native language (Japanese or English), consonant type (distractor, fricative, or stop) and consonant (singleton or geminate). The fixed factors also included the interactions between consonant type and listeners' native language, between consonant and listeners' native language, as well as the three-way interaction among consonant type, consonant, and listeners' native language. All the fixed factors were contrast coded: speakers' native language background (native Japanese: 0.5, native English: −0.5); consonant (singleton: 0.5, geminate: −0.5). Consonant type (distractor, fricative, or stop) was also contrast coded to compare between distractor and target items (0.5, −0.25, −0.25) and between fricative and stop items (0, 0.5, −0.5). The maximal random effects structure allowing the model to converge was included in the model, and included random intercepts for listeners and items, as well as random slope for listeners by consonant (distractor vs. target items). Significance was determined via model comparison as described above. The full model is summarized along with its R syntax in the Appendix B.

The results revealed a significant main effect of listeners' native language (χ^2 (1) = 18.24, $p < 0.001$), indicating that native Japanese listeners' perception accuracy was higher than native English learners'. The main effect of distractor vs. target (fricative and stop) item term was significant (χ^2 (1) = 8.67, $p < 0.01$), while its interaction with listeners' native language was not (χ^2 (1) = 0.73, $p = 0.39$), indicating that proportion correct for the distractor items was higher than the proportion correct for the target items across the two language groups. The main effect of the consonant type (fricative or stop) was not significant (χ^2 (1) = 0.55, $p = 0.46$), while its interaction with listeners' native language was (χ^2 (1) = 12.89, $p < 0.001$), indicating that the consonant type affected perception accuracy for native Japanese and native English listeners differently. The main effect of the consonant (singleton or geminate) was not significant (χ^2 (1) = 1.6, $p = 0.21$), while its interaction with listeners' native language was (χ^2 (1) = 8.19, $p < 0.01$), indicating that the consonant affected perception accuracy differently across the two language groups. These patterns (different effects of consonant types and consonant

across the two language groups) were most likely driven by the native English learners' bias towards geminate in the stop contrast; in fact, the three-way interaction among consonant type (fricative or stop), consonant (singleton or geminate), and listeners' native language was a significant predictor (χ^2 (1) = 8.69, p < 0.01).

3.3. Summary of Experiment 2

Overall, in production, native English learners generally made a smaller difference between singleton and geminate consonants than native Japanese speakers. Consonant type also affected the amount of duration difference; both language groups made a larger duration difference in the stop contrast than in the fricative contrast. However, the normalized duration difference was similar between the See-Say and auditory conditions for both language groups. That is, the type of input prompt (orthographic or auditory) did not affect the native English learners' production accuracy of the singleton-geminate consonant contrasts.

In perception, native English learners' perception of the singleton-geminate consonant contrasts was less accurate than that of native Japanese listeners, while the learners' perception of the contrasts was still quite accurate. Further, native English learners' perception was more accurate for the stop contrast than for the fricative contrast; particularly, native English learners showed a bias towards geminate in the stop contrast but not in the fricative contrast.

In the following analysis, we examine correlation patterns of the perception and production results for the English /ɹ/-/l/ and Japanese singleton-geminate contrasts together. Specifically, we examine how a learner's perception accuracy of the target non-native contrast relates to their production accuracy in different prompt conditions, and whether this pattern differs for different non-native contrasts. Because the goal of the current study is to examine the influence of perception on production in relatively difficult L2 contrasts, we limit the analysis to the English /ɹ/-/l/ and the Japanese fricative contrast, for which native English learners made a smaller singleton-geminate normalized duration difference, and perceived the difference less accurately than the stop contrast.

4. Comparison of Experiment 1 and 2

To ask whether the type of input prompt for production affected the perception-production relationship differently for the two non-native contrasts, we examined whether a learner's perception accuracy correlated with their production accuracy in a similar way in the auditory and orthographic prompt conditions, and whether different non-native contrasts showed different patterns. Specifically, we examined if a learner's production accuracy, measured by either differences in F3 or normalized duration, was predicted by their perception accuracy in a similar way across the orthographic prompt (See-Say) and the auditory prompt (Hear-Say and Hear-Delay-Say) conditions. We expected that the amount of variation in production accuracy predicted by perception accuracy would differ between the two prompt (orthographic vs. auditory) conditions for the English /ɹ/-/l/ contrast, but not for the Japanese singleton-geminate consonant contrast.

Since we were interested in how a learner's perception accuracy related to their production accuracy in different prompt conditions, we limited this analysis to production and perception data of native Japanese learners and native English learners. For native Japanese learners' production data, the F3 difference was calculated by subtracting the average F3 value for /ɹ/ from the average F3 value for /l/ for each condition for each learner. For native English learners' production data, the normalized duration difference for fricative consonants was calculated by subtracting the average normalized duration for the fricative singleton consonant from the average normalized duration for the fricative geminate consonant. The left panel in Fig. 5 shows the F3 differences made by native Japanese learners in different prompt conditions (auditory and orthographic prompt: y-axis) by their proportion correct in the perception task (x-axis). The right panel shows the normalized duration differences made by native English learners in different prompt conditions (y-axis) by their proportion correct in the perception task (x-axis). Each data point indicates one learner in the auditory (a triangle) or in the orthographic prompt condition (a circle). The solid lines are the best fitting linear regression lines.

As the left panel shows, native Japanese learners' perception accuracy generally correlated with their production accuracy of the English /ɹ/-/l/ contrast in the auditory (r = 0.77, t (12) = 4.24, p < 0.01) and in the orthographic prompt conditions (r = 0.54, t (12) = 2.25, p = 0.04). However, learners' productions were generally better in the orthographic prompt condition than in the auditory prompt conditions, even given the same values of perception accuracy (x-axis). Furthermore, the data points around the best fitting regression line in the orthographic prompt condition seem to be more spread out than the data points around the best fitting regression line in the auditory prompt conditions. To test whether this pattern was statistically significant, a paired t-test was conducted comparing residuals of the data points from the best fitting linear regression line in the orthographic prompt condition and residuals of the data points from the best fitting linear regression line in the auditory prompt conditions. The paired t-test confirmed that the residuals in the auditory prompt conditions were smaller than in the orthographic prompt condition (t (13) = −2.73, p < 0.05), indicating that native Japanese learners' perception accuracy more strongly correlated with production accuracy in the auditory prompt conditions than production accuracy in the orthographic prompt condition.

We conducted a similar analysis to examine the relationship between native English learners' production and perception accuracy for the Japanese singleton-geminate fricative contrast (see the right panel in Fig. 5). The correlation between native English learners' perception accuracy and their production accuracy in the auditory prompt conditions was statistically significant (r = 0.49, t (14) = 2.1, p = 0.05), but not in the orthographic prompt condition (r = 0.44, t (14) = 1.82, p = 0.09). It should be noted that these correlation coefficients indicate medium-to-large correlation, but do not necessarily reach significance because of the small number of data points (Field, Miles, & Field, 2012). Furthermore, a paired t-test was used comparing residuals of the data points from the best fitting linear regression line in the orthographic prompt condition and residuals of the data points from the best fitting linear regres-

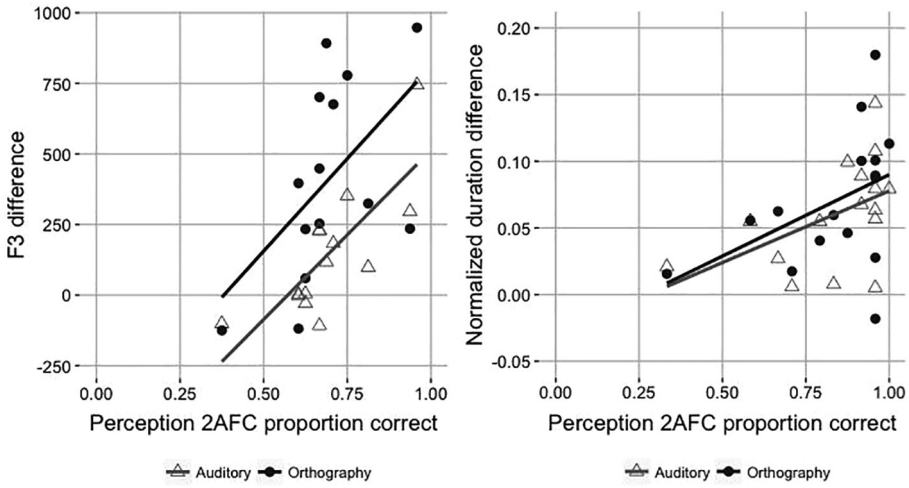

Fig. 5. Scatterplots of correlation between native Japanese learners' F3 differences made in auditory and orthographic prompt conditions and their proportion correct in the perception task (left panel), and between native English learners' normalized duration differences made in auditory and orthographic prompt conditions and their proportion correct in the perception task (right panel).

sion line in the auditory prompt conditions. It indicated that the residuals in the auditory and orthographic prompt conditions were not significantly different (t (15) = −0.75, p = 0.47), supporting that native English learners' perception accuracy correlated with production accuracy to a similar extent in the two types of prompt conditions.

These analyses indicated that native Japanese learners' perception of the English /ɹ/-/l/ contrast correlated with their production in the auditory prompt conditions more strongly than in the orthographic prompt condition, whereas native English learners' perception of the Japanese singleton-geminate contrast correlated with their productions in the two prompt conditions to a similar extent. However, because of the different types of the acoustic characteristics examined as a measure of production accuracy for the English and Japanese contrast (i.e., formant values for the English contrast, duration values for the Japanese contrast), the effect of input prompts on L2 learners' perception-production relationship was not directly compared between the two contrasts. In order to complement this, in a further analysis, we numerically compared the difference between the English and Japanese contrasts. Specifically, using the linear regression models discussed in the above analyses, we compared the difference in the model fit between the perception-auditory production model and perception-orthography production model for each contrast, and examined whether the difference in the model fit between the two models was numerically larger for one contrast than for the other. For the English contrast, the perception-auditory production model explained 33% more variance than the perception-orthography production model (ΔR^2 = 0.33[2]). However, for the Japanese contrast, the perception-auditory production model explained 5% more variance than the perception-orthography production model (ΔR^2 = 0.05). Thus, the difference in how well perception predicted production between the two prompt conditions (auditory vs. orthographic) was larger for the English contrast than for the Japanese contrast.

Together, these analyses indicated that L2 learners' perception related to their production more strongly in the auditory prompt conditions than in the orthographic prompt condition for the English contrast, whereas L2 learners' perception related to their production in the two prompt conditions to a similar extent for the Japanese contrast. These results suggest that there is some difference between the English and Japanese contrasts, or perhaps between the English and Japanese learner groups, in terms of how the input prompts cuing production impacted L2 learners' perception-production relationship.

5. Discussion

5.1. Summary of main findings

In the present study, we examined the effect of input prompts on the relationship between perception and production of non-native sounds, and whether it varied for different non-native sound contrasts: English /ɹ/-/l/ contrast for native Japanese learners of English and Japanese singleton-geminate consonant contrasts for native English learners of Japanese. The effect of input prompts cueing production was predicted to be larger for the English contrast, where the critical aspects that learner needs to focus on in perception and production are largely different (i.e., formant structure in perception and tongue movement in production), compared to the Japanese contrast, where the critical aspects in the two modalities are similar (i.e., duration as one of the primary cues in both perception and production).

In the production task, we investigated how input prompts cuing production (auditory vs. orthographic) affected learners' production accuracy of the non-native sound contrasts. Learners' production accuracy of the English /ɹ/-/l/ contrast was higher in the orthographic prompt condition than in the auditory prompt condition. However, learners' production accuracy of the Japanese singleton-geminate contrast was similar

[2] ΔR^2 indicates the size of increase in predictive power (e.g., Kleemans et al. (2014)) from model 1 (perception-auditory production model) to model 2 (perception-orthography production model). ΔR^2 was calculated by subtracting the adjusted R^2 for the perception-orthography production model from the adjusted R^2 for the perception-auditory production model.

between the two prompt conditions. Additionally, we separately measured learners' perception accuracy of the non-native contrasts and examined how their perception accuracy related to production accuracy in the different production conditions (with auditory vs. orthographic prompts). The results showed that, for the English /ɹ/-/l/ contrast, learners' perception accuracy more strongly correlated with production accuracy with auditory prompts compared to production accuracy with orthographic prompts. However, for the Japanese singleton-geminate contrast, learners' perception accuracy correlated with their production accuracy in the two prompt conditions to a similar extent. Though these results need to be interpreted with caution because direct comparisons were not made between the two contrasts, a series of analyses indicate that the effect of input prompts for the perception-production relationship was larger for the English /ɹ/-/l/ contrast than for the Japanese singleton-geminate contrast.

In the following sections, we first discuss the present results in relation with our hypothesis, in terms of why similarity of processing between perception and production may influence the effect of input prompts on the relationship between the two modalities, as well as what these results might suggest regarding how non-native sounds are represented in the two modalities. Furthermore, we discuss several accounts of alternative explanations for the different effects of input prompts on the perception-production relationship for the two non-native contrasts. While the present data cannot completely differentiate these alternative explanations, it should be noted that these explanations are not mutually exclusive from one another. That is, the current results could be explained by combinations of the possibilities presented below. In addition to discussing these possibilities in detail based on the current results and relevant literature, we suggest possible future directions to examine how each of these possibilities may impact the perception-production relationship of non-native sounds.

5.2. Perception- and production-based phonological categories

The current results demonstrated that L2 learners' production accuracy in the auditory and orthographic conditions align to a smaller extent for the English /ɹ/-/l/ contrast than for the Japanese singleton-geminate contrast. While a previous study has also demonstrated that productions in imitation and word reading do not align for a difficult non-native contrast (English /æ/-/ɛ/ contrast for native German learners of English: Llompart & Reinisch, 2019), this previous result and the current results cannot be accounted for by the same explanation. That is, while Llompart and Reinisch (2019) attributed the weak relationship between imitation and word reading to lexical difficulties (i.e., difficult L2 sounds were not robustly mapped on to learners' L2 lexical representations), this does not explain the current results because the stimuli used in the present auditory (Hear-Say & Hear-Delay-Say) and orthographic (See-Say) prompt conditions were non-words. The current results were, however, consistent with our hypothesis that the input prompts cuing production would affect the perception-production relationship to a larger degree for a non-native contrast where the critical aspects during processing are largely different between the two modalities (i.e., the English /ɹ/-/l/ contrast) than for a contrast where the critical aspects are more similar (e.g., the Japanese singleton-geminate contrast). Given these patterns, we suggest that the structures of non-native phonological categories that are learned through perception and production may not necessarily be the same. Particularly, a learner's production of non-native sounds can be influenced by auditory source (i.e., via perception) and/or by category information indexed by orthography, but the structure of the representations influenced by these sources may differ from one another.

One way to establish sound representations is to practice articulation of the sounds based on a combination of auditory and orthographic input sources. For example, in instructional settings, L2 learners are often exposed to orthography in early stages of learning (Bassetti, 2017), where pronunciation instruction commonly involves heavy use of phonetic transcription and decontextualized practice (e.g., repetition) so that learners can map their articulatory movements properly to the target non-native sounds (Saito & Lyster, 2012). In this type of practice, a language instructor may enunciate the target sounds very clearly to maximize the acoustic difference for their students while pointing at a matching orthographic representation, and then have their students repeat after them. Consequently, learners may associate their articulatory gestures with the very "best" (or "extreme") examples of the non-native sounds (e.g., /ɹ/ and /l/ in the present study). In other words, the internal sound representations to which they map their articulatory movements onto (along with orthographic forms) may favor "hyperarticulated" targets (Johnson, Flemming, & Wright, 1993). Another source that can shape the structures of non-native sound representations is variability in the auditory source. Specifically, sound representations can be formed when learners are exposed to a wide variety of sounds in the auditory input, such as in naturalistic environments or in lab-based training paradigms that utilize input variability (e.g., high-variability training: Bradlow et al., 1997, 1999; Logan, Lively, & Pisoni, 1991, or distributional learning: Escudero & Williams, 2014; Wanrooij, Escudero, & Raijmakers, 2013). The phonological categories formed based on these perceptual experiences may be broad, including a wide variety of sound representations. Thus, a learner could establish phonological sound representations through repeatedly practicing the articulation of the sounds (production) and/or through exposure to a variety of auditory input (perception), but the sound representations established based on one modality may not necessarily correspond to those established based on the other modality.

This potential mismatch between perception- and production-based sound representations may be larger for a non-native contrast for which a learner relies on dissimilar cues in perception and production (e.g., the English /ɹ/-/l/ contrast), than for a contrast for which a learner relies on more similar cues in the two modalities (e.g., the Japanese singleton-geminate contrast). That is, for example, because perception may not be very helpful to form production targets for the English contrast, native Japanese learners may heavily utilize the pronunciation training with orthography using 'clear' /ɹ/ and /l/ sounds, potentially mapping their representations of /ɹ/ and /l/ onto hyperarticulated sounds. One consequence of such a large mismatch between perception- and production-based

sound representations is that hyperarticulated production-based representations may not be broad enough to cope with the natural variability in acoustics for accurate perception. Although the auditory stimuli used in the current study were recorded in a laboratory setting (i.e., without reductions that are often present in natural speech), it is possible that they were not hyperarticulated enough for the native Japanese learners (i.e., outside of the learners' hyperarticulated representations of English /ɹ/ and /l/), making it difficult for the learners to perceptually differentiate; whereas the auditory stimuli were sufficient for the native English learners of the Japanese contrast to form production targets, leaving less room for orthography to impact productions. Thus, compared to native Japanese learners' productions of the English contrast, native English learners' production targets of Japanese singleton and geminate consonants may more directly reflect their perceptual learning via acoustic information.

Though L2 learners' sound representations can be influenced differently by auditory and orthographic input sources, it should be emphasized that any information encoded into orthographic representations of sounds needs to be mediated by perception. In fact, in the current results, it is clear that L2 learners' perception accuracy was related to their production accuracy to some extent given the correlation between their perception accuracy and production accuracy in the orthographic prompt condition (native Japanese learners with the English contrast: r = 0.54; native English learners with the Japanese contrast: r = 0.44). However, based on the different patterns of the perception-production relationship for different non-native contrasts, we suggest that the influence of perception on production may be gradient depending on the similarity of the critical aspects in perception (influenced by acoustic factors) and production (governed by articulatory factors). For example, perception and production of F0 contours (e.g., Mandarin tones as in Hao & de Jong, 2016) may be more closely connected than those for the English /ɹ/-/l/ contrast. To perceive and produce F0 contours, a learner needs to track a movement of a particular frequency (perception) and have their vocal folds vibrate faster or slower (production). This type of perception-production relationship may be more straightforward than that for the /ɹ/-/l/ contrast, where a learner needs to pay attention to F3 in relation to F2 (perception) and manipulate vocal tract configurations (production), and these could vary depending on phonetic environments (e.g., word onset, intervocalic, or coda). As another example, processing involved in perception and production of the English pre-voiced vs. short-lag contrast (Baese-Berk, 2019) may be more dissimilar than the Japanese /t/-/tt/ and /s/-/ss/ contrasts (current study). That is, while the pre-voiced vs. short-lag contrast uses a timing cue in both perception and production as in the Japanese singleton-geminate contrast, the complexity of the processing may be different. Coordinating both the timing of stop release and vocal fold vibration in production for the pre-voiced vs. short-lag contrast may make the connection between perception- and production-based sound representations less straightforward, compared to the Japanese /t/-/tt/ and /s/-/ss/ contrasts that does not include voicing, and rely largely on duration differences in perception and production.

While duration is a primary cue in both perception and production for the Japanese singleton-geminate consonant contrast, it is possible that duration representations have different natures in the two modalities. Specifically, duration differences may not be the primary production targets for the singleton and geminate consonants, but are rather a by-product of differences in some other dimension (e.g., constriction degrees in productions of Spanish /p/ vs. /b/ in the phrase-medial position; Parrell, 2011), whereas perceptual representations may be based primarily on duration. In other words, while duration is a primary cue to distinguish the Japanese singleton-geminate contrast in acoustics, this may not necessarily mean that duration is a part of speakers' representations of the production targets. Thus, differences in processing in perception and production may not necessarily be manifested in observable differences in acoustic characteristics. It is an open question whether varying similarity between cues used in perception and production impacts how acoustic vs. orthographic input is utilized during learning; this question needs to be asked in relation with the observable differences in acoustics and how those are represented in perception and production mechanisms.

However, it should also be noted that given the lack of a formal framework that allows us to predict the degree of similarity between perception- and production-based sound representations based on acoustic characteristics and articulatory movements, it is difficult to determine that the observed differences between the two non-native contrasts in the current study originate from the degree of similarity between processing involved in perception and production. Thus, until there is such a framework, our interpretation of the current results in relation with how sounds are represented in the two modalities remains tentative.

5.3. The role of perception accuracy in the variable effects of input prompts

The difference in the effect of input prompts on the perception-production relationship between Experiment 1 and 2 may also stem from differences in learners' perception accuracy more generally. Specifically, the results demonstrated that native English learners' perception of the Japanese contrast was more accurate (mean proportion correct: 0.84) than native Japanese learners' perception of the English contrast (mean proportion correct: 0.69). It is possible that native English learners of Japanese did not have to rely on the explicit category information provided by the orthographic prompts for accurate production because their perception was already helpful for their production, resulting in little effect of the prompt type on the perception-production relationship, whereas perception of the English contrast was difficult, making native Japanese learners of English rely on the orthographic prompts more for accurate production. There are two possible reasons for why perception of the Japanese contrast may have been easier for the native English learners than the perception of the English contrast for the native Japanese learners. One possibility is that native English learners' experience with durational cues that signal phonological and phonetic differences (e.g., voice onset time for stops) in L1 English helped them learn the durational contrast in L2 Japanese; the second pos-

sibility is that processing durational cues is inherently easier than spectral cues. Below, we discuss both possibilities in more detail.

One source of the difference between native English and native Japanese learners' perception accuracy of the non-native sound contrasts is their native language experience. Specifically, perceiving the Japanese duration differences may have been easier for native English listeners because duration contrasts exist in English, as opposed to Japanese listeners perceiving English spectral differences that are not found in Japanese. In English, for example, differences in voice onset timing signal phonological contrasts in consonants (e.g., as in initial consonants of *pill vs. bill*). Further, various phonetic differences in consonants and vowels are cued by duration (e.g., Flege, 1993; Mermelstein, 1978; Whalen, 1989). For example, 'fake' gemination of consonants occurs at morpheme boundaries (e.g., *topic vs. top pick*: Hayes, 2002; *innate vs. unnamed*; Kaye, 2005; Oh & Redford, 2012). However, in Japanese, the English /ɹ/ and /l/ are both mapped onto the Japanese flap phoneme, and native Japanese listeners are not used to attending to F3, the primary acoustic cue distinguishing the English /ɹ/ from /l/ (Iverson et al., 2003). Thus, native English learners may have been more well-equipped to learn the Japanese singleton-geminate consonant contrasts because of the duration differences in English, than native Japanese learners were to learn the English /ɹ/ and /l/ contrast.

These patterns are in line with previous studies demonstrating the effect of native language experience on learning of non-native sound contrasts (Choi, Kim, & Cho, 2016; McAllister, Flege, & Piske, 2002). For example, in production of English coda voicing contrasts (e.g., *bet* vs. *bed*), native Korean learners did not manipulate spectral information as much as they did the temporal information of the preceding vowel, while native English speakers manipulated both types of features, suggesting the influence of the learners' use of spectral vs. temporal features in L1 Korean on L2 English (Choi et al., 2016). That is, Korean has a smaller vowel inventory than English which may reduce a Korean listener's sensitivity to spectral cues in the signal as compared to temporal cues which are used more robustly in their L1. Therefore, it is possible that the difference we observed in the effect of input prompts on the perception-production relationship of non-native sound contrasts was partly due to the relative difficulty of perceiving the target contrast; native English learners may have been more sensitive to the duration-based Japanese singleton-geminate contrast than native Japanese learners were to the F3-based English /ɹ/-/l/ contrast.

The effect of native language experience may also explain the perceptual bias observed in both types of non-native contrasts in the current study. Specifically, native Japanese learners tended to identify items that contained English /ɹ/ correctly more often than the items that contained /l/ in the perception task, suggesting that they were biased towards the items that contained /ɹ/. Similarly, native English learners tended to identify items that contained the Japanese geminate stop consonant (/tt/) correctly more often than the items that contained the singleton stop consonant (/t/). It is possible that learners were perceptually biased towards the L2 sound that was more dissimilar to their closest L1 sound (English /ɹ/ for native Japanese learners and Japanese /tt/ for native English learners). Previous studies have demonstrated that learners' productions were more accurate for an L2 sound that is dissimilar from their closest L1 sound than for an L2 sound that is similar to their L1 sound (e.g., Flege, 1987). Similarly, for English /æ/-/ɛ/, native German learners of English were perceptually biased towards identifying /æ/, which is the 'new' vowel, than /ɛ/ (Bohn & Flege, 1990). For native Japanese learners, the Japanese flap /ɾ/ is perceptually more similar to English /l/ than to /ɹ/ (Takagi, 1993). For native English learners, /t/ is a phoneme that exists in English, but not /tt/. Thus, it is possible that learners' experience with the sound patterns of their native language influenced their perceptual bias in the non-native contrasts.

An alternative explanation for the asymmetry in the L2 learners' baseline perception may be that temporal information is inherently easier to perceive than spectral information regardless of learners' native language experience (e.g., Bohn, 1995; Choi et al., 2016). Bohn (1995) demonstrated that, even though Spanish does not have duration contrasts in vowels or other segments, native Spanish learners of English relied more heavily on duration than spectral quality to distinguish English tense-lax vowels. Similarly, native Mandarin listeners relied more on duration than on spectral cues in spite of the fact that duration is not a reliable cue in general in Mandarin (Bohn, 1995). Further, other studies have shown that, when learning to use prosodic features (e.g., final lengthening, F0 movement) for word segmentation in an artificial language, learners of different L1s rely on temporal information similarly regardless of language background, but rely on spectral information in L1-specific ways (e.g., Kim, Broersma, & Cho, 2012; Tyler & Cutler, 2009), suggesting that temporal cues have a more language-universal effect on listeners' perception than spectral information. Given these studies, it is possible that perceiving the duration-based Japanese singleton-geminate consonant contrast may have been easier not only for native English learners (in the current study) but also for learners of other L1 backgrounds than perceiving the F3-based English /ɹ/-/l/ contrast.

For the reasons pointed out in these studies, perceiving the Japanese singleton-geminate contrast may have been relatively easy for the native English learners, possibly contributing to the near ceiling effects in their perception in the current results. It is possible that this is partly responsible for the numerically weaker correlations between learners' perception and production for the Japanese contrast than for the English contrast. However, given that the native English learners' production of the Japanese contrast was not at ceiling (i.e., smaller differences between singleton and geminate consonants made by the learners than by the native Japanese speakers), it is difficult to determine the influence of the learners' perception accuracy on the relationship between perception and production for this non-native contrast. Perhaps, examining this question in different contexts, such as with learners of different proficiency levels or with different non-native contrasts that have more or less similar processing for perception and production, would help us identify the role that perception plays in the relationship between the two modalities.

5.4. The role of L1 orthographic system in the variable effects of the input prompts

The current results demonstrated that the orthographic prompts were a major factor influencing productions of the English contrast but less so for the Japanese contrast. A possible explanation for this difference is that the spectral contrast (the English /ɹ/-/l/ contrast) is more likely to be hyperarticulated given orthography than the duration contrast (the Japanese singleton-geminate contrast), regardless of the speakers' native language status. This explanation is in line with the current result that, for the English contrast, speakers in general (both native English speakers and native Japanese learners) made a larger F3 difference between /ɹ/ and /l/ in the orthographic prompt condition than in the auditory prompt conditions, whereas for the Japanese contrast, native Japanese speakers and native English learners made similar amount of duration differences between singleton and geminate consonants across the two types of production conditions. Furthermore, for the English contrast, the effect of orthographic vs. auditory prompts on the size of the F3 difference was larger for productions of native Japanese learners than those of native English speakers. These results suggest that, the orthographic prompts induced hyperarticulation of the English contrast more than the Japanese contrast in general, and this orthographically-induced hyperarticulation was more evident in native Japanese learners' productions of the English contrast than in native English speakers' productions.

The difference in the effect of orthographic prompts on L2 learners' productions of the two target contrasts possibly stems from the orthographic system of the learners' native language. That is, learners' use of orthography may have been different between L1 Japanese and L1 English, and this affected how they utilized the orthographic prompts for non-native sound production. It has been suggested that speakers of a language with a transparent orthographic system, where sound-to-letter mappings are direct and consistent (e.g., one sound corresponds to one and only one letter or one combination of letters as in Turkish), rely more heavily on the orthography than speakers of a language with a rather opaque orthographic system (e.g., English), where the sound-to-letter mappings are less consistent (orthographic depth hypothesis: Katz & Frost, 1992). Such differences in the use of orthography can apply to L2 learning. That is, the effect of non-native orthography on non-native sound production can be larger for native speakers of a language that has a transparent orthography than for native speakers of a language that has an opaque orthography (e.g., Bassetti, 2017; Bassetti et al., 2018; Erdener & Burnham, 2005). Particularly, the effect of input prompts (auditory only vs. auditory + orthography) on production of L2 non-words was stronger for native speakers of Turkish, which has a transparent orthography, than for native speakers of English, which has a rather opaque orthography (Erdener & Burnham, 2005).

Given these studies, it is possible that the native English learners in the current study did not rely heavily on the orthographic prompts for their productions of the Japanese contrast because they do not do so in their native language (e.g., Van den Bosch, Content, Daelemans, & De Gelder, 1994). Furthermore, they may not have relied on the written information when learning speech sounds in L2 Japanese, being biased to rely more on auditory information than written information when forming sound categories in Japanese. On the other hand, native Japanese learners' experience with the Japanese orthographic system, particularly, with *romaji* (the Roman alphabet letters), may have encouraged them to rely on orthography when learning English sounds. While the *kanji* writing system in Japanese, which involves ideograms of Chinese origin, does not always contain phonological information (thus an opaque orthography; Akamatsu, 2006), *romaji* does. Specifically, *romaji* transcribes each Japanese phoneme (a vowel or a combination of a consonant and a vowel) into an alphabetic sequence, presenting Japanese phonemes via one-to one mapping (Goetry, Urbain, Morais, & Kolinsky, 2005; Umemuro, 2004). With *romaji* being increasingly used in modern Japanese writing, such as for representing proper nouns in railway station signs and for operating computerized word processor (Goetry et al., 2005), it is possible that native Japanese learners of English has applied their experience with *romaji* to establishing sound-to-letter mappings in English. That is, because of the stable sound-to-letter mappings in *romaji*, native Japanese learners of English may have relied heavily on the written information when learning speech sounds in L2 English, especially for those sounds that are perceptually difficult to differentiate (e.g., English /ɹ/ and /l/).

This relatively strong effect of orthography on native Japanese speakers' learning of L2 English sounds may also explain the perceptual bias towards /ɹ/ than /l/ (discussed in an earlier section). That is, it is possible that the orthographic system influenced learners' judgment of ambiguous liquid sounds. The *romaji* system uses the symbol "r" to represent the Japanese flap phoneme /ɾ/. Due to this influence, it is possible that whenever the learners heard an ambiguous liquid sound (e.g., *?anby*) they interpreted the signal was an /ɹ/ (e.g., *ranby*) rather than an /l/ (e.g., *lanby*). Thus, transparency of L1 orthographic system is one potential source of perceptual bias in L2. However, whether the transparency of L1 orthography affects the effect of auditory vs. written information on L2 sound category formation needs to be investigated further.

6. Conclusion

We examined whether the effect of input prompts on the perception-production relationship of non-native sounds varied for different non-native contrasts. This question was tested with two types of non-native contrasts (English /ɹ/-/l/ and Japanese singleton-geminate consonant contrasts). It was predicted that a difference in input prompts (auditory vs. orthographic) would influence the perception-production relationship to a larger degree for the English contrast, where the critical cues during processing are more distinct in perception and production, than for the Japanese contrast, where the critical cues in the two modalities are more similar to one another. Overall, the results demonstrated that the difference in the prompt type impacted how perception related to production for the English contrast but less so for the Japanese contrast. Specifically, for the English contrast, learners' production accuracy was higher when given orthographic prompts compared to when given auditory prompts. However, for the Japanese contrast, learners' production accuracy was similar in

different prompt conditions. These results suggest that one of the factors that characterizes the complex perception-production relationship is the input type for production, and this influence of input type also depends on the similarity of the factors that are critical for perception and production. Though future research may be able to distinguish between alternative explanations in different research contexts, such as by comparing different non-native contrasts and different L1-L2 pairings, we suggest that the perception-production relationship of non-native sounds can be influenced by combinations of these explanations rather than a single factor.

CRediT authorship contribution statement

Misaki Kato: Conceptualization, Methodology, Software, Formal analysis, Investigation, Visualization, Writing - original draft. **Melissa Michaud Baese-Berk:** Writing - review & editing, Funding acquisition, Supervision.

Acknowledgements

This work was supported by National Science Foundation (NSF) BCS grant 1734166 to MMBB. The authors would like to thank Melissa Redford, Charlotte Vaughn, Vsevolod Kapatsinski, Taehong Cho, and three anonymous reviewers for providing comments on an earlier version of this manuscript. We also thank Maiko Hata and Kaori Idemaru for their help recruiting participants, and Cydnie Davenport for her assistance with acoustic analysis.

Appendix A. List of stimuli in Experiment 1 and 2

	Position				
	Onset	Onset CL	Intervocalic	Coda	L/R Both (not included in analysis)
Real words	read	grow	correct	dear	clarify
	lead	glow	collect	deal	role
	room	broom	berry	war	Laura
	loom	bloom	belly	wall	Mallory
Non-words	renk	brize	neron	nare	lorief
	lenk	blize	nelon	nel	rolief
	reab	prost	surive	osfire	lyrome
	leab	plost	sulive	osfile	rylome
	ryfee	freep	herry	skeer	leira
	lyfee	fleep	helly	skeel	reila
	ranby	breen	sorash	zore	roless
	lanby	bleen	solash	zall	loress
	rof	grep	teerow	jeer	roully
	lof	glep	teeolow	jeel	loury
	reimon	groon	zerard	speer	ryless
	leimon	gloon	zelard	speel	lyress

List of English target items used in Experiment 1

	Fricative (pitch-accent)		Stop (pitch-accent)		Affricate (not included in the analysis)	
Real words	asa (HL)	sassa (HL)	mate (HL)	matte (HL)		
	hosu (HL)	hossu (LH)	heta (LH)	hetta (LH)		
	kasai (LH)	kassai (LH)	wata (LH)	watta (LH)		
Non-words	rusa (HL)	russa (HL)	ruta (HL)	rutta (HL)	datsu (HL)	dattsu (HL)
	besa (HL)	bessa (HL)	reta (HL)	retta (HL)	hetsu (HL)	hettsu (HL)
	wasu (HL)	wassu (HL)	bate (HL)	batte (HL)		
	zesu (HL)	zessu (HL)	mute (HL)	mutte (HL)		
	nase (HL)	nasse (HL)				
	huse (HL)	husse (HL)				
	musa (LH)	mussa (LH)	muta (LH)	mutta (LH)	gatsu (LH)	gattsu (LH)
	desa (LH)	dessa (LH)	teta (LH)	tetta (LH)	metsu (LH)	mettsu (LH)
	rasu (LH)	rassu (LH)	sate (LH)	satte (LH)		
	hesu (LH)	hessu (LH)	zute (LH)	zutte (LH)		
	mase (LH)	masse (LH)				
	zuse (LH)	zusse (LH)				

List of Japanese target items used in Experiment 2

Appendix B

Model summaries and results of likelihood ratio testing for the mixed-effects regression models. The change in degrees of freedom for the Chi-square values is 1.

Experiment 1 (English /ɹ/-/l/ contrast): Average F3 in production

R syntax:

F3.model = lmer(f3Mean ~ /ɹ/ vs. /l/ + /ɹ/ vs. /l/: Baseline vs. See-Say + /ɹ/ vs. /l/: Hear-Delay-Say vs. Hear-Say + /ɹ/ vs. /l/: See-Say vs. Auditory + Native language: /ɹ/ vs. /l/ + Native language: /ɹ/ vs. /l/: Baseline vs. See-Say + Native language: /ɹ/ vs. /l/: Hear-Delay-Say vs. Hear-Say + Native language: /ɹ/ vs. /l/: See-Say vs. Auditory + (1 + /ɹ/ vs. /l/: Baseline vs. See-Say + /ɹ/ vs. /l/: Hear-Delay-Say vs. Hear-Say + /ɹ/ vs. /l/: See-Say vs. Auditory | speaker) + (1|word), data = F3.dat, REML = F)

Fixed Effects	Estimate	Std Err	*t*-value	Chi Square (*p*-value)
(Intercept)	2565.71	38.64	66.41	
/ɹ/ vs. /l/	576.78	17.79	30.69	177.92 ($p < 0.001$)
/ɹ/ vs. /l/: Baseline vs. See-Say	204.65	64.98	3.15	10.05 ($p < 0.01$)
/ɹ/ vs. /l/: Hear-Delay-Say vs. Hear-Say	13.04	35.31	0.37	0.21 ($p = 0.65$)
/ɹ/ vs. /l/: See-Say vs. Auditory	414.26	72.78	5.69	24.57 ($p < 0.001$)
Native language: /ɹ/ vs. /l/	586.79	24.01	24.44	545.58 ($p < 0.001$)
Native language: /ɹ/ vs. /l/: Baseline vs. See-Say	−120.36	82.67	−1.46	2.12 ($p = 0.15$)
Native language: /ɹ/ vs. /l/: Hear-Delay-Say vs. Hear-Say	−49.82	70.85	−0.7	0.49 ($p = 0.48$)
Native language: /ɹ/ vs. /l/: See-Say vs. Auditory	−284.37	126.39	−2.25	4.69 ($p < 0.05$)

Experiment 1 (English /ɹ/-/l/ contrast): Perception 2AFC task

R syntax:

English2AFC.model = glmer(Correct ~ Native language + Distractor vs. target + /ɹ/ vs. /l/ + Distractor vs. target: Native language + Native language: /ɹ/ vs. /l/ + (1 + Distractor vs. target| listener) + (1| word), family = "binomial", data = English2AFC.dat)

Fixed Effects	Estimate	Std Err	*z*-value	Chi Square (*p*-value)
(Intercept)	3.4	0.26	13.17	
Native language	3.65	0.46	7.97	49.34 ($p < 0.001$)
Distractor vs. target items	−0.33	0.47	−0.69	0.45 ($p = 0.5$)
/ɹ/ vs. /l/ consonants	−0.68	0.53	−1.27	2.8 ($p = 0.09$)
Distractor vs. target items: Native language	−4.83	0.72	−6.69	46.61 ($p < 0.001$)
Native language: /ɹ/ vs. /l/ consonants	−0.26	0.9	−0.29	0.08 ($p = 0.77$)

Experiment 2 (Japanese singleton-geminate consonant contrast): Normalized duration in production

R syntax:

Dur.model = lmer (normalized duration ~ Singleton vs. geminate + Singleton vs. geminate: Fricative vs. Stop + Native language: Singleton vs. geminate + Singleton vs. geminate: Baseline vs. See-Say + Singleton vs. geminate: Hear-Delay-Say vs. Hear-Say + Singleton vs. geminate: See-Say vs. Auditory + Singleton vs. geminate: Fricative vs. Stop: Native language + Singleton vs. geminate: Fricative vs. Stop: Baseline vs. See-Say + Singleton vs. geminate: Fricative vs. Stop: Hear-Delay-Say vs. Hear-Say + Native language: Singleton vs. geminate: See-Say vs. Auditory + Native language: Singleton vs. geminate: Baseline vs. See-Say + Native language: Singleton vs. geminate: Hear-Delay-Say vs. Hear-Say + Native language: Singleton vs. geminate: See-Say vs. Auditory + (1 + Singleton vs. geminate: Baseline vs. See-Say conditions: Fricative vs. Stop + Singleton vs. geminate: Hear-Delay-Say vs. Hear-Say conditions: Fricative vs. Stop + Singleton vs. geminate: See-Say vs. Auditory conditions: Fricative vs. Stop | speaker) + (1|word), data = Dur.dat, REML = F)

Fixed Effects	Estimate	Std Err	*t*-value	Chi Square (*p*-value)
(Intercept)	0.36	0.008	47.29	
Singleton vs. geminate	−0.1	0.01	−8.51	45.3 ($p < 0.001$)
Singleton vs. geminate: Fricative vs. Stop	−0.07	0.02	−2.94	7.97 ($p < 0.01$)
Native language: Singleton vs. geminate	−0.04	0.004	−8.89	76.98 ($p < 0.001$)
Singleton vs. geminate: Baseline vs. See-Say	0.04	0.04	1.04	1.07 ($p = 0.3$)
Singleton vs. geminate: Hear-Delay-Say vs. Hear-Say	0.005	0.006	0.86	0.72 ($p = 0.4$)
Singleton vs. geminate: See-Say vs. Auditory	0.03	0.03	1.08	0.08 ($p = 0.78$)
Singleton vs. geminate: Fricative vs. Stop: Native language	−0.01	0.008	−1.65	2.59 ($p = 0.11$)
Singleton vs. geminate: Fricative vs. Stop: Baseline vs. See-Say	−0.06	0.09	−0.68	0.53 ($p = 0.47$)
Singleton vs. geminate: Fricative vs. Stop: Hear-Delay-Say vs. Hear-Say	0.004	0.01	0.31	0.03 ($p = 0.87$)
Singleton vs. geminate: Fricative vs. Stop: See-Say vs. Auditory	0.001	0.06	0.03	7.24 ($p < 0.01$)
Native language: Singleton vs. geminate: Baseline vs. See-Say	0.006	0.01	0.41	0.07 ($p = 0.8$)
Native language: Singleton vs. geminate: Hear-Delay-Say vs. Hear-Say	0.02	0.01	1.87	3.51 ($p = 0.06$)
Native language: Singleton vs. geminate: See-Say vs. Auditory	−0.02	0.16	−0.95	2.07 ($p = 0.15$)

Experiment 2 (Japanese singleton-geminate consonant contrast): Perception 2AFC task

R syntax:

Japanese2FC.model = glmer (Correct ~ Distractor vs. target + Native language + Fricative vs. Stop + Singleton vs. geminate + Distractor vs. target: Native language + Native language: Fricative vs. Stop + Native language: Singleton vs. geminate + Native language: Fricative vs. Stop: Singleton vs. geminate + (1 + Distractor vs. target| listener) + (1| word), family = "binomial", data = Japanese2AFC.dat)

Fixed Effects	Estimate	Std Err	*z*-value	Chi Square (*p*-value)
(Intercept)	7.18	3.2	2.25	
Distractor vs. target items	12.21	12.71	0.96	8.67 ($p < 0.01$)
Native language	7.47	6.39	1.17	18.24 ($p < 0.001$)
Fricative vs. Stop	−0.49	0.63	−0.77	0.55 ($p = 0.46$)
Singleton vs. geminate	−0.61	0.63	−0.97	1.6 ($p = 0.21$)
Distractor vs. target items: Native language	19.43	25.4	0.77	0.73 ($p = 0.39$)
Native language: Fricative vs. Stop	1.83	1.23	1.49	12.89 ($p < 0.001$)
Native language: Singleton vs. geminate	1.22	1.23	1.0	8.19 ($p < 0.01$)
Native language: Fricative vs. Stop: Singleton vs. geminate	−4.7	1.56	−3.02	8.69 ($p < 0.01$)

References

Akamatsu, N. (2006). Literacy acquisition in Japanese-English bilinguals. In R. M. Joshi & P. G. Aaron (Eds.), *Handbook of orthography and literacy* (pp. 481–496). Mahwah, NJ: Lawrence Erlbaum.

Baese-Berk, M. M. (2019). Interactions between speech perception and production during learning of novel phonemic categories. *Attention, Perception, & Psychophysics, 81*(4), 981–1005.

Baese-Berk, M. M., & Samuel, A. G. (2016). Listeners beware: Speech production may be bad for learning speech sounds. *Journal of Memory and Language, 89*, 23–36.

Bassetti, B. (2017). Orthography affects second language speech: Double letters and geminate production in English. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(11), 1835–1842.

Bassetti, B., & Atkinson, N. (2015). Effects of orthographic forms on pronunciation in experienced instructed second language learners. *Applied Psycholinguistics, 36*(2), 67–91.

Bassetti, B., Escudero, P., & Hayes-Harb, R. (2015). Second language phonology at the interface between acoustic and orthographic input. *Applied Psycholinguistics, 36*, 1–6.

Bassetti, B., Sokolović-Perović, M., Mairano, P., & Cerni, T. (2018). Orthography-induced length contrasts in the second language phonological systems of L2 speakers of English: Evidence from minimal pairs. *Language and Speech, 61*(4), 577–597.

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

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

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In O.-. S. Bohn & M. Munro (Eds.), *Language experience in second language speech learning: In honor of James Emil Flege* (pp. 13–34). Amsterdam, The Netherlands: John Benjamins Publishing.

Boersma, P., & Weenik, D. (2015). Praat: Doing phonetics by computer.

Bohn, O. S. (1995). Cross-language speech perception in adults: First language transfer doesn't tell it all. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 279–304). Timonium, MD: York Press.

Bohn, O. S., & Flege, J. E. (1990). Interlingual identification and the role of foreign language experience in L2 vowel perception. *Applied Psycholinguistics, 11*(3), 303–328.

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English /r/ and /l/: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*(5), 977–985.

Bradlow, A. R., Pisoni, D. B., Akahane-Yamada, R., & Tohkura, Y. I. (1997). Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. *The Journal of the Acoustical Society of America, 101*(4), 2299–2310.

Choi, J., Kim, S., & Cho, T. (2016). Phonetic encoding of coda voicing contrast under different focus conditions in L1 vs. L2 English. *Frontiers in Psychology, 7*, 1–17.

Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin, 96*, 341–370.

Cutler, A., Weber, A., & Otake, T. (2006). Asymmetric mapping from phonetic to lexical representations in second-language listening. *Journal of Phonetics, 34*(2), 269–284.

Davidson, L. (2010). Phonetic bases of similarities in cross-language production: Evidence from English and Catalan. *Journal of Phonetics, 38*(2), 272–288.

de Jong, K., Hao, Y.-C., & Park, H. (2009). Evidence for featural units in the acquisition of speech production skills: Linguistic structure in foreign accent. *Journal of Phonetics, 37*(4), 357–373.

Diehl, R. L., & Kluender, K. R. (1989). On the objects of speech perception. *Ecological Psychology, 1*(2), 121–144.

Eckman, F. R. (2004). From phonemic differences to constraint rankings: Research on second language phonology. *Studies in Second Language Acquisition, 26*(4), 513–549.

Erdener, V. D., & Burnham, D. K. (2005). The role of audiovisual speech and orthographic information in nonnative speech production. *Language Learning, 55*, 191–228.

Escudero, P., & Williams, D. (2014). Distributional learning has immediate and long-lasting effects. *Cognition, 133*(2), 408–413.

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Thousand Oaks, CA: Sage Publications.

Flege, J. E. (1987). The production of "new" and "similar" phones in a foreign language: Evidence for the effect of equivalence classification. *Journal of Phonetics, 15*(1), 47–65.

Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *The Journal of the Acoustical Society of America, 93*(3), 1589–1608.

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

Flege, J. E., MacKay, I. R., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *The Journal of the Acoustical Society of America, 106*(5), 2973–2987.

Flege, J. E., Takagi, N., & Mann, V. (1995). Japanese adults can learn to produce English /ɹ/ and /l/ accurately. *Language and Speech, 38*(1), 25–55.

Fowler, C. A. (1986). An event approach to the study of speech perception from a direct-realist perspective. *Journal of Phonetics, 14*(1), 3–28.

Goetry, V., Urbain, S., Morais, J., & Kolinsky, R. (2005). Paths to phonemic awareness in Japanese: Evidence from a training study. *Applied Psycholinguistics, 26*(2), 285–309.

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 22*(5), 1166–1183.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review, 105*(2), 251–278.

Goto, H. (1971). Auditory perception by normal Japanese adults of the sounds. *Neuropsychologia, 9*(3), 317–323.

Han, M. S. (1992). The timing control of geminate and single stop consonants in Japanese: A challenge for nonnative speakers. *Phonetica, 49*(2), 102–127.

Hanulíková, A., Dediu, D., Fang, Z., Bašnaková, J., & Huettig, F. (2012). Individual differences in the acquisition of a complex L2 phonology: A training study. *Language Learning, 62*(s2), 79–109.

Hao, Y. C., & de Jong, K. (2016). Imitation of second language sounds in relation to L2 perception and production. *Journal of Phonetics, 54*, 151–168.

Hardison, D. M., & Saigo, M. M. (2010). Development of perception of second language Japanese geminates: Role of duration, sonority, and segmentation strategy. *Applied Psycholinguistics, 31*(1), 81–99.

Hayes, R. L. (2002). The perception of novel phoneme contrasts in a second language: A developmental study of native speakers of English learning Japanese singleton and geminate consonant contrasts. *Coyote Papers, 12*, 28–41.

Heald, S. L., & Nusbaum, H. C. (2014). Speech perception as an active cognitive process. *Frontiers in Systems Neuroscience, 8*, 1–5.

Hirata, Y. (2004). Training native English speakers to perceive Japanese length contrasts in word versus sentence contexts. *The Journal of the Acoustical Society of America, 116*(4), 2384–2394.

Hirata, Y., & Whiton, J. (2005). Effects of speaking rate on the single/geminate stop distinction in Japanese. *The Journal of the Acoustical Society of America, 118*(3), 1647–1660.

Houde, J. F., & Jordan, M. I. (1998). Sensorimotor adaptation in speech production. *Science, 279*, 1213–1216.

Houde, J. F., & Jordan, M. I. (2002). Sensorimotor adaptation of speech I: Compensation and adaptation. *Journal of Speech, Language, and Hearing Research, 45*(2), 295–310.

Idemaru, K., & Guion-Anderson, S. (2010). Relational timing in the production and perception of Japanese singleton and geminate stops. *Phonetica, 67*(1–2), 25–46.

Iverson, P., Kuhl, P. K., Akahane-Yamada, R., Diesch, E., Tohkura, Y. I., Kettermann, A., & Siebert, C. (2003). A perceptual interference account of acquisition difficulties for non-native phonemes. *Cognition, 87*(1), B47–B57.

Johnson, K., Flemming, E., & Wright, R. (1993). The hyperspace effect: Phonetic targets are hyperarticulated. *Language, 69*(3), 505–528.

Kartushina, N., Hervais-Adelman, A., Frauenfelder, U. H., & Golestani, N. (2015). The effect of phonetic production training with visual feedback on the perception and production of foreign speech sounds a). *The Journal of the Acoustical Society of America, 138*(2), 817–832.

Katz, L., & Frost, R. (1992). The reading process is different for different orthographies. In R. Frost & L. Katz (Eds.). *Orthography, phonology, morphology, and meaning: Advances in psychology* (Vol. 94, pp. 67–84). Oxford: North Holland.

Kaye, A. S. (2005). Germination in English. *English Today, 21*, 43–55.

Kim, S., Broersma, M., & Cho, T. (2012). The use of prosodic cues in learning new words in an unfamiliar language. *Studies in Second Language Acquisition, 34*(3), 415–444.

Kleemans, T., Segers, E., & Verhoeven, L. (2014). Cognitive and linguistic predictors of basic arithmetic skills: Evidence from first-language and second-language learners. *International Journal of Disability, Development and Education, 61*(3), 306–316.

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition, 21*(1), 1–36.

Llompart, M., & Reinisch, E. (2019). Imitation in a second language relies on phonological categories but does not reflect the productive usage of difficult sound contrasts. *Language and Speech, 62*(3), 594–622.

Logan, J. S., Lively, E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of Acoustical Society of America, 89*(2), 874–886.

Lotto, A. J., Sato, M., & Diehl, R. L. (2004). Mapping the task for the second language learner: The case of Japanese acquisition of /r/ and /l/. In J. Slifka, S. Manuel, & M. Matthies (Eds.), *From sound to sense: 501 years of discoveries in speech communication [Online conference proceedings]*. Cambridge, MA: MIT.

McAllister, R., Flege, J. E., & Piske, T. (2002). The influence of L1 on the acquisition of Swedish quantity by native speakers of Spanish, English and Estonian. *Journal of Phonetics, 30*(2), 229–258.

Mermelstein, P. (1978). On the relationship between vowel and consonant identification when cued by the same acoustic information. *Perception & Psychophysics, 23*(4), 331–336.

Minagawa, Y., & Kiritani, S. (1996). Discrimination of the single and geminate stop contrast in Japanese by five different language groups. *Annual Bulletin of Research Institute of Logopedics and Phoniatrics, 30*, 23–28.

Miyawaki, K., Jenkins, J. J., Strange, W., Liberman, A. M., Verbrugge, R., & Fujimura, O. (1975). An effect of linguistic experience: The discrimination of [r] and [l] by native speakers of Japanese and English. *Perception & Psychophysics, 18*(5), 331–340.

Nielsen, K. (2011). Specificity and abstractness of VOT imitation. *Journal of Phonetics, 39*(2), 132–142.

Oh, G. E., & Redford, M. A. (2012). The production and phonetic representation of fake geminates in English. *Journal of Phonetics, 40*(1), 82–91.

Pardo, J. S. (2006). On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America, 119*(4), 2382–2393.

Parrell, B. (2011). Dynamical account of how /b, d, g/ differ from /p, t, k/ in Spanish: Evidence from labials. *Laboratory Phonology, 2*(2), 423–449.

Peperkamp, S., & Bouchon, C. (2011). The relation between perception and production in L2 phonological processing. In *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech 2011)* (pp. 161–164). Rundle Mall: Causal Productions.

R Core Team (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rafat, Y. (2015). The interaction of acoustic and orthographic input in the acquisition of Spanish assibilated/fricative rhotics. *Applied Psycholinguistics, 36*(1), 43–66.

Ramus, F., Peperkamp, S., Christophe, A., Jacquemot, C., Kouider, S., & Dupoux, E. (2010). A psycholinguistic perspective on the acquisition of phonology. *Laboratory Phonology, 10*(3), 311–340.

Redford, M. A. (2015). Unifying speech and language in a developmentally sensitive model of production. *Journal of Phonetics, 53*, 141–152.

Saito, K., & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning, 62*(2), 595–633.

Schertz, J., Cho, T., Lotto, A., & Warner, N. (2015). Individual differences in phonetic cue use in production and perception of a non-native sound contrast. *Journal of Phonetics, 52*, 183–204.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime: User's guide*. Psychology Software Incorporated.

Sheldon, A., & Strange, W. (1982). The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics, 3*(3), 243–261.

Shin, D., & Iverson, P. (2011). Individual differences in vowel epenthesis among Korean learners of English. In *Proceedings of the International Congress of Phonetic Sciences* (pp. 1822–1825). Hong Kong, China.

J. Steele (2005). Assessing the role of orthographic versus uniquely auditory input in acquiring new L2 segments. Paper presented at 7èmes rencontres internationales du réseau français de phonologie (pp. 2–4).

Stevens, K. N., & Blumstein, S. E. (1981). The search for invariant acoustic correlates of phonetic features. In P. D. Eimas & J. L. Miller (Eds.), *Perspectives on the study of speech* (pp. 1–38). Hillsdale, NJ: Erlbaum.

Strange, W., & Dittmann, S. (1984). Effects of discrimination training on the perception of /r-l/ by Japanese adults learning English. *Perception & Psychophysics, 36*, 131–145.

Takagi, N. (1993). *Perception of American English /r/ and /l/ by adult Japanese learners of English: A unified view* Ph.D. Dissertation. Irvine: University of California.

Tateishi, M. (2013). *Effects of the use of ultrasound in production training on the perception of English /r/ and /l/ by native Japanese speakers* Unpublished Masters thesis. University Of Calgary.

Tsukada, K., Cox, F., Hajek, J., & Hirata, Y. (2018). Non-native Japanese learners' perception of consonant length in Japanese and Italian. *Second Language Research, 34*(2), 179–200.

Tyler, M. D., & Cutler, A. (2009). Cross-language differences in cue use for speech segmentation. *The Journal of the Acoustical Society of America, 126*(1), 367–376.

Umemuro, H. (2004). Lowering elderly Japanese users' resistance towards computers by using touchscreen technology. *Universal Access in the Information Society, 3*, 276–288.

Van den Bosch, A., Content, A., Daelemans, W., & De Gelder, B. (1994). Measuring the complexity of writing systems. *Journal of Quantitative Linguistics, 1*(3), 178–188.

Wanrooij, K., Escudero, P., & Raijmakers, M. E. (2013). What do listeners learn from exposure to a vowel distribution? An analysis of listening strategies in distributional learning. *Journal of Phonetics, 41*(5), 307–319.

Whalen, D. H. (1989). Vowel and consonant judgments are not independent when cued by the same information. *Perception & Psychophysics, 46*(3), 284–292.

Wong, J. W. S. (2013). The effects of perceptual and/or productive training on the perception and production of English vowels /ɪ/ and /iː/ by Cantonese ESL learners. *Proceedings of Interspeech*, 2113–2117.

Yamada, R. A. (1995). Age and acquisition of second language speech sounds: Perception of American English /ɹ/ and /l/ by native speakers of Japanese. In W. Strange (Ed.), *Speech perception and linguistic experience: Issues in cross-language research* (pp. 305–320). Timonium, MD: York Press.

Yamada, R. A., & Tohkura, Y. I. (1992). The effects of experimental variables on the perception of American English /r/ and /l/ by Japanese listeners. *Perception & Psychophysics, 52*(4), 376–392.