A Cross-Linguistic Study of L3 Phonological Acquisition of Stop Contrasts

Jiaqi Liu\(^1\)\(^{1}\) and Jiayan Lin\(^2\)\(^{2}\)

Abstract
The research reported in this article investigated how students learning Japanese or Russian as a third language (L3) perceived and produced word-initial stops in their respective target language and the link between perception and production. The participants in the study were 39 Chinese university students who spoke Mandarin Chinese as their first language (L1), English as their second language (L2), and Japanese or Russian as their L3. An L3 identification task, an L3 reading task, and an L2 reading task were used to investigate the learners’ perception and production of word-initial stops. The results demonstrated that the phonetic similarity in different stop categories between L1, L2, and L3 contributed to learners’ confusion in perception. On the contrary, L3 learners could perceive the new acoustic feature voicing lead, but found it difficult to produce L3 voiced stops. In addition, the study found a positive relationship between the perception and production of voiceless stops in the initial stage of L3 acquisition, but there was no correlation between the perception and production of voiced stops. Pedagogical implications for L3 speech learning are discussed on the basis of the results.

Keywords
L3 phonological acquisition, cross-linguistic influence, stop contrasts, perception and production

Introduction
In the past, a number of researchers have observed and analyzed the similarities and commonalities between learners’ first languages and their target languages, and discussed the acquisition of sounds in a foreign language within the framework of second language (L2) acquisition. In the context of increasing diversity in language learning motivations, population movements, and the emergence of multilingualism, multilingual education has been introduced into school curricula (Cenoz, 2012).

The development of language other than English (LOTE) education is taking place at a remarkable speed in the evolving landscape of higher education in mainland China (Shen & Gao, 2019). However, these fast-growing programs face a range of challenges, including problems with unrealistic curricula, teacher shortages, and pedagogy (Han et al., 2019). Research on the teaching and acquisition of non-English languages is insufficient and is not keeping pace with the development of these programs in Chinese universities (Gu et al., 2011).

The acquisition of a third language (L3) will be influenced by a learner’s first language (L1) and L2 (Cenoz et al., 2001; Williams & Hammarberg, 1998); this is known as cross-linguistic influence (CLI). De Angelis (2007) defined CLI as the influence of prior linguistic knowledge on the production, comprehension, and development of a target language. Combined CLI occurs when the target language is concurrently influenced by two or more languages, or when the target language is influenced in turn by a language that has already been influenced by a prior language (De Angelis, 2007). Combined CLI is an important research topic in L3 phonological acquisition (Sypiańska, 2016b; Wrembel, 2014). The nature of combined CLI in L3 phonological acquisition is closely related to the combination of prior languages. For instance, if the two prior languages share some phonetic features, they will exert a “double interference” in L3 acquisition (Chamot, 1973).

Recent years have seen substantial progress in the study of L3 phonological acquisition as a subfield in the framework of multilingual acquisition (Cabrelli Amaro & Wrembel, 2016; Rothman et al., 2019). Cabrelli Amaro and Wrembel (2016)
hold that current findings in third language acquisition (TLA) research have generally demonstrated complex cross-linguistic influence between the native language and non-native languages, but they cannot fully explain L3 phonological models and the factors that determine or condition cross-linguistic influence in L3 speech. For instance, most studies investigated Indo-European languages as target languages (e.g., Cabrelli Amaro, 2017; Gut, 2010; Kopečková, 2015), but only a few studies have focused on non-Indo-European languages (e.g., Japanese by Tremblay (2007) and Liu et al. (2019), and Mandarin Chinese by Gabriel et al. (2014)). This imbalance in the amount of research on different target languages will impact our capability to discuss how typological distance and structural similarity will affect multilingual phonological acquisition; therefore, Cabrelli Amaro and Wrembel (2016) and Rothman et al. (2019) called for L3 phonological acquisition research to expand and include more language combinations and target languages. There is not much research into the link between speech perception and production from the TLA perspective, but such research is clearly necessary (Cabrelli Amaro & Wrembel, 2016). Due to the inherent complexity of L3 acquisition studies, the existing research findings cannot paint the whole picture of multilingual development.

This study was conducted taking into account the above issues with the existing research on L3 phonological acquisition. The participants in the study were L3 Japanese or Russian learners, who all had the same native language (Mandarin Chinese) and many years of experience of learning English as an L2. The study tested the perception and production of stops by L3 Japanese or Russian beginner learners, and analyzed the values and distribution of their voice onset time (VOT) in an effort to explore the perception and production of L3 word-initial stops and the link between the two modalities.

The research questions of this study are as follows:

**Research Question 1**: How do multilingual learners perceive L3 voiceless and voiced stops?

**Research Question 2**: How do multilingual learners produce L3 voiceless and voiced stops?

**Research Question 3**: Is there a relationship between the perception and production of L3 stops?

Findings around these questions will not only shed light on the perception–production link in L3 phonological acquisition but will also have implications for effective L3 speech training.

**Literature Review**

A review of the literature was carried out around the following four themes: cross-linguistic differences in word-initial stops, TLA of speech sounds, second language acquisition (SLA) of stops, and the perception–production link in SLA.

**Cross-Linguistic Differences in Word-Initial Stops**

Research on acoustic phonetics and perceptual phonetics shows that whether a sound is voiced or voiceless is related to a range of acoustic features, such as voicing lead, aspiration, closure duration, energy density, and F1 cutback. Among these features, voicing lead and aspiration are directly related to the interval between the release of a stop consonant and the onset of voicing, and therefore they are the key features that determine whether a sound is voiced or voiceless. These two acoustic features are captured by VOT (Abramson & Whalen, 2017; Lisker & Abramson, 1964).

Following the convention of marking the instant of release burst as zero-time, Lisker and Abramson (1964) stated that a negative value of VOT, which means that voicing begins before release, indicates voicing lead and that a positive value of VOT, which means voicing begins after release, indicates voicing lag. They studied 11 languages and classified word-initial stops into three categories on the basis of VOT values, as follows: (a) voicing lead or negative VOT, (b) a short-lag or zero onset, and (c) a long-lag. Following this categorization, Keating (1984) and Kingston and Diehl (1994) divided languages into voicing languages and aspirating languages on the basis of stop voicing. Voicing languages contrast voicing lead plosives with short-lag plosives. On the contrary, aspirating languages contrast short-lag plosives with long-lag plosives.

VOT can not only be used to describe cross-linguistic differences in stops (Cho & Ladefoged, 1999; Lisker & Abramson, 1964), but it is also an important acoustic parameter in studies on the acquisition of stops in L2 (Bohn & Flege, 1993; Liu, 2011) and L3 (Liu et al., 2019; Wrembel, 2014, 2015). Flege (2017) pointed out that VOT is a significant acoustic phonetic parameter to be adjusted in speech production and a meaningful auditory cue to the perception of stop consonants; therefore, VOT provides a bridge for research into the perception–production link. Accordingly, this study used VOT as a phonetic measure to investigate the perception and production of L3 speech sounds and the link between the two modalities.

**TLA of Speech Sounds**

The study of TLA can often provide a new perspective on the process of language learning that L1 or SLA research may not offer (e.g., Flynn et al., 2004). Earlier TLA studies were mostly based on observations of the complex patterns of phonetic acquisition in one or a small number of multilingual learners (Chamot, 1973; Rivers, 1979; Singh & Carroll, 1979). For instance, Chamot (1973) studied the acquisition of English as an L3 by a French-Spanish bilingual boy. One of the learning difficulties she discovered was that it was hard for the boy to learn English vowels which are different from French and Spanish. The vowels in the learner’s two earlier acquired languages, French and Spanish, are nearly
identical, which suggests that his prior knowledge of these two languages that share some phonetic features seemed to affect his acquisition of the L3, a problem termed “double interference” by Chamot. Although earlier studies mostly provided impressionistic accounts rather than using precise measurements, according to Cabrelli Amaro and Wrembel (2016), the language combinations studied and the results of the observations have important implications for current research.

Over the past few years, there has been substantial development in research on L3 phonological acquisition in terms of both research design and methods. Increasing attention has been paid to the influence of a non-native language or the combined influence of prior languages. A number of studies have demonstrated the influence of L2 on L3 (Kamiyama, 2007; Onishi, 2016; Tremblay, 2007; Wrembel, 2010), and some scholars have proposed that prior languages exert a combined CLI on L3 phonological acquisition (Sypiańska, 2016b; Wrembel, 2014, 2015). Previous empirical research suggests that multilingual learners tend to be more sensitive to new phonetic contrasts and learn them faster (Onishi, 2016; Wrembel, 2014, 2015; Wrembel et al., 2019).

Quite a few studies have focused on the VOT patterns in L3 acquisition of stop consonants (Liu, 2019; Liu et al., 2019; Llama et al., 2010; Llama & López-Morelos, 2016; Onishi, 2016; Shimizu, 2012; Sypiańska, 2013; Tremblay, 2007; Wrembel, 2014, 2015; Wunder, 2010; Zeng & Liu, 2019). These studies shed light on the cross-linguistic influence in phonological acquisition through examining the VOT patterns in stops across languages. Among these studies, Wrembel (2014, 2015) investigated the VOT patterns in the production of stops.

Wrembel’s (2014) study involved L1 Polish, L2 English learners who had learned French as L3 for 2 or 5 years or who had learned German as L3 for 5 or 7 years. The study revealed unique interlanguage VOT patterns produced by multilingual learners by comparing the VOT values the learners produced in all the three language systems. There was a distinct difference in voiceless stops between the learners’ L1 and L2. L3 learners were likely to have been influenced by the long VOT length in L2 English voiceless stops. In a similar study, Wrembel (2015) investigated the VOT patterns in the voiceless stops produced by L1 German, L2 English learners who had learned French as L3 for about 7 years. The results showed some evidence that multilingual learners restructured their phonetic space, but the author pointed out that the language combinations in the study made it impossible to tease apart the potential influences of L1 and L2 on the acquisition of voiceless stops in the L3.

The perception of stop consonants in L3 acquisition has not been as widely researched as production of stop consonants. Onishi (2016) studied the acquisition of L3 Japanese by L1 Korean and L2 English speakers in terms of perception. The participants in this study had been learning elementary Japanese as their L3 for less than two semesters. The accuracy in the learners’ perception of L3 was also found to be related to the proficiency of their L2. Onishi attributed this to L3 learners’ generally increased perceptual sensitivity after having studied an L2. Liu et al. (2019) investigated the perception of stop consonants by L3 Japanese, Russian, or Spanish beginners in the multilingual context of Chinese tertiary education and found that similarity in speech sounds had a negative influence on the perception of the target language.

The above studies explored the L3 production or perception of voiced and voiceless stops separately. Research on the link between production and perception is scant; Shimizu (2012) studied the perception and production of L3 stop consonants among L1 Thai, L2 English, and L3 Japanese learners who had learned Japanese for 2 to 4 years. The results of the test indicated that L1 transfer occurred in the production of L3 voiced stops /b, d/, whereas L2 transfer occurred in the production of L3 voiceless stops /p, t, k/. Besides, Shimizu found that learners could auditorily perceive the difference between languages, but had difficulty in articulating the target voiced stops.

The above review of the literature shows that there are not many findings about the link between the perception and production of L3 stop consonants in the multilingual context, and the existing research is insufficient to provide a full picture of the patterns in multilingual acquisition; this will impact our ability to effectively carry out speech training. However, there is more research on stop consonants in SLA than in TLA; in fact, extensive SLA research has been conducted about the link between the perception and production of stop consonants in SLA.

**SLA of Stops**

Previous research in SLA indicates that the stop consonant voicing contrast in the learners’ native language is closely related to the acquisition of the stop consonants in the target language (Flege, 1992) and that the phonological features of the L1 may be transferred to the L2 and impact the VOT values of the stops in the L2 (Flege, 1991; Flege & Hillenbrand, 1987). In the initial stage of acquisition, the accuracy of English language learners’ perception of word-initial stops is related to the mapping of stop categories between their L1 and L2 (Bohn & Flege, 1993). The voicing contrast in stop consonants in the target language is reported to present difficulties for Chinese native speakers (NSs) in the initial stage of acquisition (Flege, 1992; Liu, 2011; Wang, 2009). Chinese learners of Japanese are also likely to be confused in their perception of the voiced and voiceless stops in L2 Japanese, due to them being influenced in the initial stage by both the difference and similarity in phonological structures between L1 and L2 (Fukuoka, 1995; Liu, 2011). As aspiration is not an important distinctive feature in Russian, Russian learners of Chinese cannot discern whether a sound is aspirated or not as well as Chinese NSs do (Liang, 1963),
but with an increase in learning hours Russian learners of intermediate or higher level can approximate Chinese NSs in their perception of aspiration in Mandarin (Ran et al., 2016).

In the field of SLA, linguistic difference and similarity are drawn heavily on in the prediction and interpretation of L2 acquisition. Flege (1995) stated in his Speech Learning Model (SLM) that “the greater the perceived phonetic dissimilarity between an L2 sound and the closest L1 sound, the more likely it is that phonetic differences between the sounds will be discerned” (p. 239). However, learners are likely to assimilate a sound that is phonetically similar but not identical to an L1 category into that L1 category, a phenomenon termed “equivalence classification” (Flege, 1987). It is worth investigating whether SLM, a classic SLA model, can be extended to explain and predict multilingual acquisition. Sypiańska (2016a) investigated the applicability of SLM in multilingual acquisition, and the results supported the majority of the predictions regarding equivalence classification.

**Perception–Production Link in SLA**

In SLA, there is still no consensus on whether speech perception precedes speech production or vice versa. Major theories of L2 pronunciation learning such as the SLM (Flege, 1995, 2003) make predictions about the link between perception and production. According to Flege (1995), “many L2 sounds production errors have a perceptual basis” (p. 238). In other words, perception and production are closely related. In terms of speech training, some researchers argue that speech perception and production facilitate each other, so that training in one modality will also improve performance in the other (Hirata, 2004; Leather & James, 1991; Mathews, 1997). However, there are also researchers who believe that accurate L2 speech perception is not sufficient for accurate L2 speech production (Kartushina et al., 2015). Schertz et al. (2015) investigated L1 Korean, L2 English learners’ acquisition of stop consonants and found that their cue use in production and perception did not have equal weighting. Although the participants could successfully discriminate the stops in both perception and production, their ways of discriminating the sounds were not aligned at a subtler phonetic level. Perception and production relate to distinct cognitive and motor skills, which is likely to cause the two modalities to develop in an unsynchronized way (Nagle, 2018). Nagle (2018) studied the perception–production link by modeling English speakers’ perception and production of L2 Spanish stops over time and found a time-lagged relationship between the two modalities. The time-lagged change models showed that an increase in contrast sensitivity was significantly related to decreasing VOT in the production of L2 voiceless stop /p/ at a later testing time, but there was no significant relationship between the perception and production of the voiced stop /b/.

Nagle’s (2018) research on the temporal dynamics of the perception–production link also indicates that the acquisitional stage of a target language and the target language proficiency will have a direct impact on the link between the two modalities. Hanulíková et al.’s (2012) study also showed mixed results regarding the perception–production link; these researchers trained Dutch NSs on Slovak words with complex consonant clusters. The results led the researchers to hypothesize that perception and production may dissociate in the initial stage of acquisition or that production may only improve after perceptual accuracy has reached a certain point.

This review of the literature shows that previous studies on TLA of stop consonants have mainly focused on the cross-linguistic influence of prior languages on the target language in terms of perception or production. There is little research on the link between L3 speech perception and production, compared with a body of research on the link between the two modalities in SLA. In the context of China’s multilingual education, typical L3 learners in China speak Mandarin as their native language and have rich experience learning English as an L2. Targeting L1 Mandarin, L2 English learners, this study investigated their ability to perceive and produce stop consonants in L3 Japanese or Russian, and analyzed the perception–production link. Methods to improve speech sounds training are suggested on the basis of the results.

**The Study**

**Participants**

The participants in this study were 20 L3 Japanese learners and 19 L3 Russian learners, with the same L1 Mandarin and L2 English. The participants were aged between 18 and 20 years, all having sound hearing and speaking abilities. They were recruited from a university in China on a voluntary basis and each received RMB 50 (US$7.3) as compensation. They had learned Japanese or Russian for 2 months (or 80 contact hours), spending 10 hours per week in the language class over 8 weeks. By the time of participation, they had learned either the Japanese Kana or the Russian alphabet, and could be regarded as beginner learners. They had not taken any courses that targeted pronunciation training. All their teachers were Chinese and had not introduced to them the differences in pronunciation between their native language or English and the target language. The study restricted participation to those who were born and raised in northern China or northwestern China, as the dialects spoken in these regions have the same stop system as in Mandarin Chinese. None of the participants had been to their target language country.

The participants were asked to report their English learning experience. They had all started learning English at the age of 9 or 10, equivalent to Year 3 of primary school in
China. They had studied English for 10.30 years on average ($SD = 1.56$), spending an average of 3.5 hours per week in class during the semester. Almost all of their English teachers were Chinese, but two participants reported having had weekly one-hour English classes taught by NSs of English for 3 years when they were high school students. None of the participants had stayed in any English-speaking country for over a month.

Also recruited in the study were NSs of Japanese and Russian, with 24 participants for each language. These participants were language teachers working in China or students studying in China. Two of them provided the phonetic stimuli for the perception test, 10 participated in the perception test, and 12 in the production test.

**Target Structure: VOT in Mandarin, English, Japanese, and Russian**

The participants of this study spoke Mandarin Chinese as their L1. Wu (1986) carried out a comprehensive analysis of the voice and voiceless stops in Mandarin from physiological and acoustic perspectives. He pointed out that the key to Chinese people’s distinguishing between aspirated and unaspirated stops was the intensity of airflow. The VOT values of voiceless aspirated stops range from 85 to 104 ms (long-lag), and the VOT values of voiceless unaspirated stops range from 8 to 15 ms (short-lag). According to the categorization described in section “Cross-Linguistic Differences in Word-Initial Stops,” Mandarin is regarded as an aspirating language.

Findings from acoustic experiments show that pronunciation of word-initial voiced stops in English is usually without voicing lead—that is, vocal cords do not vibrate during the period of articulatory closure for the stops (Lisker & Abramson, 1964). Word-initial voiced stops pronounced by native English speakers do not have voicing lead, and VOT values are positive (Klatt, 1975; Kopczyński, 1977). Ladefoged and Keith (2011) stated that the word-initial /p, t, k/ in English are voiceless aspirated sounds and that the key difference between /p/ and /b/ was not voicing but aspiration. Therefore, word-initial stops in English fall into two categories: short-lag for voiced and long-lag for voiceless stops, which is characteristic of an aspirating language. The VOT values of voiceless aspirated stops range from 47 to 70 ms (long-lag), and the VOT values of voiceless unaspirated stops range from 11 to 27 ms (short-lag) (Klatt, 1975). Although English NSs occasionally prevoice /b, d, g/, it is very uncommon (Beckman et al., 2011; Davidson, 2016).

The study focused on Japanese and Russian as target languages, partly because such language combinations have rarely been investigated and partly because these are the two languages with the largest number of learners except English in China. Among the more than 400 million foreign language learners in China, 93.6% are learning English, 7.1% are learning Russian, and 2.5% are learning Japanese (Wei & Su, 2012). The stop systems in Japanese and Russian are different from those in Mandarin and English. The word-initial stops in Japanese fall into two groups: voiceless stops [p, t, k] and voiced stops [b, d, g] (The International Phonetic Association, 1999, p. 117). The average VOT values of word-initial labial stops are [b]-89 ms and [p]41 ms (Shimizu, 1993). Voiceless stops in Russian are voiceless unaspirated sounds, and vocal cords vibrate during the closure for voiceless stops (Kulikov, 2012). Ringen and Kulikov’s (2012) study shows that more than 97% of Russian word-initial voiced stops are prevoiced, mean VOT values being [b]–70 ms and [p]18 ms respectively. According to the categorization described in “Cross-Linguistic Differences in Word-Initial Stops,” both Japanese and Russian are regarded as voicing languages.

When speech sounds are compared across languages, the same phonetic category would be classified differently at the phonological level (Piccinini & Arvaniti, 2018). Although both Mandarin and English are regarded as contrasting voiced and voiceless stops at the phonological level (/p/~/b/, /t/~/d/, /k/~/g/), there are differences between the abstract phonological categories and phonetic categories, and this can be captured by VOT. To summarize, both Mandarin and English are aspirating languages as their word-initial stops are distinguished in terms of aspiration, whereas both Japanese and Russian are voicing languages, as they discriminate between voiced and voiceless stops on the basis of whether there is voicing lead during the closure.

**Materials**

The participants were asked to complete an L3 identification task and an L3 reading task. Both tasks had the same set of target items but different fillers. There were 24 target items and 12 fillers for each L3 (i.e., Japanese or Russian). The target items were bilabial consonants /p, b/, alveolar consonants /t, d/, and velar consonants /k, g/, with each consonant in the word-initial position of a monosyllable or disyllable word. The vowel that immediately followed the stop was [a]. There was no syllable stress in the target item. Refer to Appendix A for the target items used in this study.

The stimuli for identification task were recorded by four Japanese and Russian NSs (two speakers, one of each sex, for each language). The NSs were aged between 25 and 45 years. The Japanese NSs grew up in Tokyo and the Russian NSs grew up in Moscow.

For each L3, the two NSs each read 36 linguistic stimuli (including 24 target items and 12 fillers), creating a total of 72 stimuli. All the stimuli were inserted into the same sentence structure in each L3. The Japanese sentence structure was “このは_______” and the Russian structure was “Бор_______,” both meaning “This is_______.” Each sentence was read three times by the NSs, and the pronunciations in the second reading were used as the stimuli for the identification task. We used a TASCAM DR44WL linear PCM recorder (with 44.1 kHz sampling rate, 16-bit quantization
level) and AKG C544L headset microphones. Praat 6.0 (Boersma & Weenink, 2009) was used to create the phonetic stimuli for the identification task. Each stimulus was preceded by a 400-ms period of silence to space out the stimuli. A synthetic sound with a fundamental frequency of 500 Hz and a duration of 400 ms was added as a start signal to alert the participants. Finally, a 1000-ms period of silence was added to the end of each stimulus so that the participants would have enough time to understand the provided options on the computer screen.

In terms of the L3 reading task, similar to the identification task, a total of 36 stimuli (including 24 target items and 12 fillers) were inserted into the same sentence structure for each L3. The sentences in Japanese hiragana or the Russian alphabet were shown on PowerPoint slides and presented on a computer screen in front of the participants.

To understand the learners’ acquisition of stops in L2 English, the learners were also asked to complete an English reading task targeting voiced and voiceless stops. The target items were monosyllabic words with /p, t, k/ and /b, d, g/ in the word-initial position. The procedures and equipment used for this task were the same as those for the L3 reading task.

Procedures

The identification task and the reading tasks took place at different times. The first time the participants met, they were asked to complete a questionnaire about their native language background and foreign language study experience, and then they completed the identification task in L3 Japanese or Russian. A week after the identification task, the participants attended again to complete the reading tasks. They first completed the L3 reading task, and after a short break they completed the L2 reading task.

Before each task, the participants were asked to introduce themselves in the target language (saying, for example, their name, university, and year of study). The purpose of the self-introduction was to ease them into their target language mode. The items in the L3 identification and reading tasks were spelled out using the target language orthography (i.e., Japanese hiragana or the Russian alphabet) because the participants had finished learning the target language alphabet, and using the target language orthography would also enable the identification task and the reading task to be presented in the same style.

Identification task. Praat 6.0’s ExperimentMFC 6 script was used for the identification task. Before starting the task, the participants completed a short training session to become familiar with the experiment equipment and procedures. Each participant spent about 10 minutes doing the identification task. The task environment was quiet, with participants sitting in front of a computer and wearing a SONY MDR-ZX110NC noise-canceling headset. They started the task by pressing the space bar, and then minimal pairs of phonetic stimuli randomized by ExperimentMFC 6 appeared on the computer screen (Appendix B). The participants were instructed to choose the stimuli they heard as quickly and accurately as possible. They were given a chance to take a short break between every five stimuli and could decide when to continue independently.

Reading tasks. During the L3 reading task, the participants could move on to the next slide by clicking the mouse when they were ready. Each slide randomly presented one test item. The participants were instructed to read each test item three times at a normal speed. If they found themselves making a mistake in pronunciation, they were allowed to start again and repeat until they were satisfied. There were three breaks in this task. The task took place in a sound-proof environment. A TASCAM DR44WL linear PCM recorder (44.1 kHz sampling rate, 16-bit quantization level) and AKG C544L head microphones were used to record the speech. Following a short break after the L3 reading task, the participants completed the L2 English reading task. The task procedures and recording equipment were the same as for the L3 reading task.

Analysis

The perception test (i.e., identification task) was analyzed to investigate how accurately L3 learners and NSs could perceive voiceless and voiced stops in Japanese or Russian, and the relationship between perceptual accuracy and the VOT values of the stimuli. The production test (i.e., the reading tasks) was analyzed to find out the VOT distribution of the voiceless and voiced stops in learners’ L2 English, and the ratio of positive and negative VOTs and the distribution of VOTs produced by L3 learners and NSs of Japanese or Russian. To investigate the perception–production link in L3, we first analyzed the relationship between the perceptual accuracy of voiceless stops and VOT values produced by the learners, and then analyzed the relationship between the perceptual accuracy of voiced stops and the proportion of negative VOTs.

The study followed the method in Lisker and Abramson (1964) to measure the VOT values. We annotated the stops produced by L3 learners and NSs using SPPAS (Version 1.8.6) and manually checked the annotation. The Praat script (Version 2012/10/30) was used to extract the acoustic parameters from each annotation tier. Programming software R (R Core Team, 2014) was used to analyze the data and generate figures and graphs.

Results

Results Regarding the Acquisition of L2 (English) Stops

Both L3 Japanese and Russian learners completed an L2 English reading task, which was used to test how they
produced voiced and voiceless stops in English. The VOT values of voiced and voiceless stops produced by L3 learners are displayed in Figure 1. The two groups of learners almost overlapped in the distribution of VOT values for word-initial stops in L2. The VOT values for voiceless stops /p, t, k/ ranged from 58 to 104 ms, and the VOT values for voiced stops /b, d, g/ ranged from 12 to 25 ms. In learners’ production, all the VOTs for word-initial stops in English were above zero, and the main difference between voiced and voiceless stops lay in the value of VOT. The cutoff where learners separated voiced and voiceless stops lay between 30 and 40 ms. Most human languages have a VOT of 30 to 40 ms as the boundary between aspirated and unaspirated stops (Keith, 2003), which means it is a universal threshold. In other words, similar to English NSs, the word-initial stops produced by both L3 groups showed a contrast between aspirated (long-lag) and unaspirated (short-lag) stops, typical of an aspirating language.

**Results Regarding the Perception of L3 Stops**

L3 Japanese and L3 Russian learners’ accuracy in perceiving voiced and voiceless stops is shown in Figure 2. We used the t.test function in R to perform an independent-samples t-test of the perceptual accuracy scores produced by the non-native speakers (NNSs) and NSs. The results showed that there was no significant difference in the accuracy of perceiving voiced stops between L3 Japanese learners and Japanese NSs, $t(33) = 0.36, p = .72$, n.s. However, the NSs’ accuracy in perceiving Japanese voiceless stops was significantly higher than that of the L3 Japanese learners, $t(30) = 3.48, p < .01$. Russian NSs were significantly more accurate than L3 Russian learners in perceiving both voiced stops, $t(23) = 5.29, p < .001$, and voiceless stops, $t(23) = 8.42, p < .001$. In addition, a paired sample t-test showed that L3 Japanese learners and L3 Russian learners had a significantly higher accuracy in perceiving voiced stops than perceiving voiceless stops: $t(19) = 5.25, p < .001$ for Japanese and $t(18) = 3.35, p < .01$ for Russian.

To explore the relationship between learners’ perceptual accuracy and the VOTs of the phonetic stimuli, we used the glm function in R to conduct a logistic regression, with learners’ accuracy in perceiving L3 stops as a dependent binary variable and the VOTs of the stimuli as an independent variable. Results showed that there was a positive correlation between L3 Japanese and L3 Russian learners’ accuracy in perceiving voiceless stops and the VOTs of the stimuli ($β = 0.137, SE = 0.018, z = 7.543, p < .001$ for L3 Japanese learners; $β = 0.049, SE = 0.008, z = 6.28, p < .001$ for L3 Russian learners). To be more specific, the bigger the VOT value, the higher the accuracy in perception. By contrast, there was no significant correlation between learners’ accuracy in perceiving voiced stops in L3 and the VOTs of the stimuli.

**Results Regarding the Production of L3 Stops**

The descriptive statistics for L3 Japanese and L3 Russian learners’ production of voiced and voiceless stops are displayed in Table 1 and Figure 3. Table 1 shows the proportions
of positive VOTs and negative VOTs and the means and standard deviations of VOTs for voiced and voiceless stops produced by L3 learners and NSs. As shown in Table 1, there was a higher proportion of positive VOTs than negative VOTs in the production of L3 voiced stops in both L3 (Japanese: 79.17% vs. 20.83%; Russian: 59.21% vs. 40.79%). However, NSs produced a higher proportion of negative VOTs than positive VOTs for voiced stops (Japanese: 63.89% vs. 36.11%; Russian: 99.31% vs. 0.69%). Figure 3 portrays the distribution of VOTs produced by learners and NSs for voiced and voiceless stops. The vertical line in Figure 3 marks the release of plosives and indicates a division separating positive VOTs to the right and negative VOTs to the left. Figure 3 also shows a considerable difference in the distribution of VOTs for voiced stops between learners and NSs.

Table 1. Proportions of Positive and Negative VOTs and VOT Means.

| Language | Group | Voice | Positive VOT | Negative VOT |
|----------|-------|-------|--------------|--------------|
|          |       |       | %+VOT | Mean VOT (SD) | %−VOT | Mean VOT (SD) |
| Japanese | NNS   | Voiced| 79.17 | 21.08 ms (11) | 20.83 | −99.78 ms (34) |
|          |       | Voiceless| 99.58 | 60.78 ms (26) | 0.42 | −86.00 ms (NA) |
|          | NS    | Voiced| 36.11 | 20.27 ms (7) | 63.89 | −74.05 ms (24) |
|          |       | Voiceless| 100 | 39.07 ms (22) | 0 |             |
| Russian  | NNS   | Voiced| 59.21 | 18.10 ms (7) | 40.79 | −120.14 ms (41) |
|          |       | Voiceless| 79.39 | 19.02 ms (9) | 20.61 | −108.38 ms (44) |
|          | NS    | Voiced| 0.69 | 16.00 ms (NA) | 99.31 | −90.55 ms (36) |
|          |       | Voiceless| 100 | 23.97 ms (13) | 0 |             |

Note. VOT = voice onset time; NNS = non-native speakers; NS = native speakers.

Figure 3. Distribution of VOTs in the production test.
Note. VOT = voice onset time.
To further understand the difference between learners and NSs in the production of voiced and voiceless stops, we used linear mixed-effects model (the lme function from the nlme package in R) to analyze the production data. The independent variables were Voice (whether the stop is voiced or voiceless), VOT type (whether the VOT value is positive or negative), and Group (NNS vs. NS), and the dependent variable was the proportion of positive and negative VOTs.

Analysis shows that there was interaction between the three variables Voice, VOT type, and Group for both L3 Japanese, $\chi^2(1) = 50.57, p < .0001$, and L3 Russian, $\chi^2(1) = 65.82, p < .0001$. The emmeans function in R was used to further analyze the interaction effects. It was found that the proportion of negative VOTs for voiced stops produced by Japanese or Russian NSs was significantly higher than that produced by L3 Japanese or Russian learners—Japanese: $t(30) = 7.72, p < .0001$; Russian: $t(29) = 6.81, p < .0001$. This indicates that L3 Japanese or Russian learners were often confused about the voicing contrast when they produced voiced stops. In contrast, similar to NSs, L3 Japanese or Russian learners predominantly produced positive VOTs for voiceless stops. To be more specific, there was no significant difference in VOT type between Japanese NSs and learners in the production of voiceless stops, $t(30) = 0.08, p = .94$, n.s. L3 Russian learners also produced a high proportion of positive VOTs for voiceless stops (79.39%).

The Correlation Between Perception and Production

We further analyzed the results from the identification task and the L3 reading task to explore the relationship between learners’ perception and production of stops in L3 Japanese and Russian. Due to the repeated measures design, we used the linear mixed-effects model to test the relationship between the perception and production of voiceless stops.

First, we used the lme function (from the nlme package in R) to test whether there was a correlation between learners’ perceptual accuracy and VOT production of voiceless stops. The independent variable was learners’ average accuracy in perceiving the voiceless stops, and the dependent variable was VOT values of the voiceless stops. Results showed that there was a positive correlation between the perceptual accuracy and VOT production of voiceless stops for both groups of L3 learners—Japanese: $\chi^2(1) = 6.76, p < .01$; Russian: $\chi^2(1) = 19.98, p < .0001$. In other words, learners’ accuracy in perceiving voiceless stops can predict VOT values: the higher the perceptual accuracy, the larger the VOT value. Then we used the effect package in R to plot the relationship between perception and production (Figure 4). Each plot in Figure 4 has learners’ perceptual accuracy of voiceless stops on the $x$-axis and VOT values of voiceless stops on the $y$-axis. It is clear from the two effect plots that there was a positive correlation between the perception and production of voiceless stops for both L3 Japanese and Russian learners.

Next, we used the lme function (from the nlme package in R) to test whether there was a correlation between learners’ accuracy in perceiving voiced stops and the proportion of negative VOTs, the independent variable being learners’ average perceptual accuracy of voiced stops and the dependent variable being the proportion of negative VOTs of voiced stops. Results showed that there was no correlation between learners’ accuracy in perceiving voiced stops and their production of negative VOTs for both groups of L3 learners—Japanese: $\chi^2(1) = 1.78, p = .18$, n.s.; Russian: $\chi^2(1) = 0.068, p = .79$, n.s. This means that learners’ perceptual accuracy of voiced stops cannot predict the proportion of negative VOTs.

Discussion

Perception of L3 Stops

The perception test of L3 voiceless stops showed that NSs were significantly more accurate than learners. An analysis of the word-initial stop contrasts across L1, L2, and L3 indicated that the difference in performance was mainly related to the cross-linguistic difference in the distribution of VOTs.
The L3 target languages Japanese and Russian are voicing languages, which contrast short-lag and voicing lead in the stops system, whereas the learners’ L1 and L2 are aspirating languages which distinguish stops in terms of aspiration (long-lag vs. short-lag). The voiceless aspirated stops in learners’ L1 and L2 had clearly longer VOTs and a wider distribution of VOTs than the voiceless stops in both L3s, whereas the voiceless stops in both L3s were close in VOTs to the voiceless unaspirated stops in the learners’ L1 and L2. Therefore, in terms of VOTs, L3 voiceless stops are similar to the voiceless unaspirated stops in L1 and L2. When the learners were perceiving the L3 voiceless stops, they may have been affected by the similarity between phonemes and mistakenly mapped L3 voiceless stops onto the voiceless unaspirated stops in L1 or L2. Hence, there was confusion in the perception of L3 voiceless stops and a low accuracy.

The above finding was in line with the previous results in the SLA field. For example, Bohn and Flege (1993) found that in the initial stage of acquisition, the accuracy of Spanish learners of English in perceiving word-initial stops was related to the mapping between L1 and L2 stop categories, rather than being purely determined by the VOT values of stops. Both Fukuoka (1995) and Liu (2011) conducted empirical studies on the perception of voiceless and voiced stops by L1 Chinese, L2 Japanese beginners and found that learners had a low accuracy in perceiving word-initial voiceless stops in Japanese. Liu (2011) proposed that the difficulty in discriminating voiceless and voiced stops was probably due to the combined influence from the phonological difference and similarity between the native language and L2.

It is generally acknowledged that L3 learners are more sensitive to phonetic contrasts in the target language that closely resemble those in prior languages (Gut, 2010; Onishi, 2016; Wrembel et al., 2019). Onishi (2016) argued that L3 learners have a general advantage in phonological perception after having learned an L2. Wrembel et al.’s (2019) study indicated that it seems to be easier for L3 beginners to perceive the subtle phonetic differences in new phonological contrasts. However, different from their findings, the L3 learners in this study did not seem to have an advantage in phonological perception over L2 learners. The learners still made mistakes in perceiving the voiceless stops in Japanese or Russian, probably due to inaccurate mapping of stop categories that are phonetically similar between languages. This result may be explained by the fact that the learners’ L1 and L2 are both aspirating languages which share phonetic contrasts that are different from L3 Japanese and Russian, two voicing languages; this indicates the presence of combined cross-linguistic influence (De Angelis, 2007). The L3 learners in this study were susceptible to double interference (Chamot, 1973) from similar phonetic contrasts in L1 Mandarin and L2 English, which seemed to have exerted a negative impact on their perception of L3 voiceless stops.

Results from the perception test showed that learners were more accurate in perceiving voiced stops in L3 Japanese and L3 Russian than in perceiving voiceless stops. The prevoicing feature in L3 Japanese and Russian does not exist in the word-initial stops in learners’ L1 and L2, but Chinese learners’ perception of L3 voiced stops was more accurate than that of voiceless stops. This demonstrates that learners were more likely to perceive the novel phonological feature in L3 (i.e., prevoicing) that was absent from the phonetic contrasts in L1 and L2, and therefore they were able to achieve a high perceptual accuracy.

The results from the perception test showed that learners’ acquisition of L3 voiceless and voiced stops was in line with the prediction made by the SLM (Flege, 1995). According to SLM, an L2 linguistic feature that is similar to an L1 feature is hard to learn because learners may put this in the same category as in the L1 and stop learning the new feature. Compared with subtle cross-linguistic similarity, significant difference between languages is more conducive to the phonetic learning of the target language. This is because a linguistic feature that is novel or markedly different from L1 is more likely to be noticed by learners and is therefore learned more easily. Earlier analysis showed that the VOT distributions of the voiceless stops in L3 Japanese and Russian were similar to the VOT distributions of the voiceless unaspirated stops in L1 and L2, but the two sets of phones belong to different phonological categories. Chinese learners were likely to have equated the voiceless stops in L3 Japanese or Russian with the category of voiceless unaspirated stops in L1 and L2, which made it difficult for them to perceive the voiceless stops in Japanese or Russian. On the contrary, the perception test of L3 voiced stops showed that L3 learners were more likely to perceive the novel phonological feature in L3 (i.e., prevoicing) that was absent from the phonetic contrasts in L1 and L2, and had a higher perceptual accuracy. If learners can distinguish a target feature in L3 from prior languages, it will help them to establish a new L3 category. On the contrary, if learners have difficulty in discriminating target features that are similar to prior languages, they are not likely to further test and establish the cross-linguistic difference, which will lead to perceptual difficulty (Kingston, 2003). SLM is a widely recognized model of L2 phonetic acquisition, but it also seems to be able to explain or predict the acquisition of phonetic contrasts in L3 under the circumstances of this study (i.e., when L1 and L2 share a phonetic contrast that is different from that of L3 and subject L3 phonetic acquisition to double interference).

It will be challenging but theoretically and practically rewarding to investigate whether SLA models can be applied in TLA. Sypiańska (2016a) investigated the applicability of SLM in multilingual acquisition and produced findings that were in line with most of the predictions of equivalence classification in SLM.

**Production of L3 Stops**

The results from the L3 reading task showed that the VOTs of the voiceless stops produced by both groups of L3 learners were mainly positive, and this pattern was not different from
the VOT production by both groups of NS. This means that confusion between voiced and voiceless stops was not evident in learners’ production of voiceless stops. It is worth noting that Wrembel (2014) investigated combined CLI by exploring the interaction between three phonological systems in terms of the VOT patterns of stops. The two parallel studies she conducted involved different language combinations: L1 Polish, L2 English, and L3 French or L3 German. The results showed a significant difference in voiceless stops between the learners’ L1 and L2. It was found that L3 learners were likely to have been influenced by the long-lag VOT categories of voiceless stops in L2 English. In a similar study, Wrembel (2015) focused on a different language combination: L1 German, L2 English, and L3 French. Some evidence of multilingual learners restructuring their phonetic space was detected, but the author acknowledged that the influence of L1 and L2 on the acquisition of VOT patterns in L3 could not be teased apart. In this study, learners’ two prior languages share the features of stop consonants in aspirating languages, giving rise to double interference (Chamot, 1973), a scenario that is different from the combined CLI documented by Wrembel (2014) but similar to the situation in Wrembel (2015).

Our results also showed that although L3 beginners had a high accuracy in perceiving L3 voiced stops, production was not easy for them. Both groups of L3 learners produced a significantly lower proportion of negative VOTs for voiced stops than NS. This result was also obtained in Shimizu (2012). Shimizu found that learners could auditorily discern the difference between the target language (L3 Japanese) and two prior languages (L1 Thai and L2 English), but it was difficult for them to produce the target sounds correctly. Fukuoka (1995) also found that Chinese L2 Japanese beginners had difficulty in producing the voiceless stops with negative VOT in Japanese. Nagle (2018, 2019) investigated the dynamic development of L1 English, L2 Spanish learners’ perception and production of stops. He found that most learners achieved near-native perceptual performance by the third or fourth session of data collection (about 1 or 1.5 semesters of Spanish). However, the group trajectory indicated that although learners began to prevoice Spanish stops, they did not approximate a native-like level of prevoicing.

The clear dissociation between the perception and production of stops can also be related to articulatory factors. Ohala’s (1983, 1997) research demonstrates that it is difficult to initiate and maintain voicing (such as in the pronunciation of voiced stops) due to the Aerodynamic Voicing Constraint (AVC), a universal physiological mechanism. From the perspective of language universals, the existence of voiceless obstruents in a language is a prerequisite for voiced obstruents, which means if a language has voiced obstruents, it must have voiceless obstruents—but the opposite is not true (Kubozono, 2003). These findings indicate that the production of voiced stops is more marked than that of voiceless stops in the light of the universal patterns in speech production. This is consistent with L3 learners’ production performance in this study. The VOTs of L3 voiceless stops were mostly positive, whereas the proportion of negative VOTs for L3 voiced stops was low, significantly different from that produced by NS.

**Perception–Production Link of L3 Stops**

There is not a lot of existing research into the link between the perception and production of target language phones in the multilingual context. This study was an attempt in this regard; we aimed to investigate the relationship between the perception and production of L3 word-initial voiceless and voiced stops among learners whose L1 and L2 shared phonetic contrasts that were different from those in L3.

Our analysis showed that there was a positive correlation between the perceptual accuracy of L3 voiceless stops and VOT production: the higher the perceptual accuracy, the longer the VOT produced. In other words, perceptual accuracy could predict the length of VOT. Also note that there was a positive correlation between learners’ perceptual accuracy of L3 voiceless stops and the VOT values of the stimuli used in the perception test (as shown in section “Results Regarding The Perception of L3 Stops”). These results indicated that both learners’ perception and production of voiceless stops were closely related to VOT values. Learners were likely to use the length of VOT as a clue to help themselves perceive and produce L3 voiceless stops. The same clue also serves to distinguish aspirated and unaspirated sounds in terms of long-lag versus short-lag in L1 Mandarin and L2 English. This finding has therefore confirmed our expectation that in the initial stage of acquisition, L3 learners may utilize the stop systems in L1 and L2 to perceive and produce L3 voiceless stops.

The correlation analysis also showed that there was no correlation between the perceptual accuracy of L3 voiced stops and the proportion of negative VOTs. In other words, learners’ perceptual accuracy of L3 voiced stops could not predict the production of negative VOTs.

The results also showed that both groups of L3 learners had a high accuracy in perceiving L3 voiced stops, but produced a much lower proportion of negative VOTs for L3 voiced stops than Japanese or Russian NS. Findings about naturalistic language learners indicate that the link between the perception and production of speech sounds may vary depending on learners’ language proficiency and the properties of the target sounds (Flege et al., 1997, 1999; Saito & van Poeteren, 2018). Learners with a high proficiency in the target language or monolinguals display a high correlation between perception and production accuracy (Flege, 1995, 2003; Flege et al., 1997; Piccinini & Arvaniti, 2018). At the initial stages of L2 acquisition, perception and production may dissociate, which means that perceptual accuracy may not be sufficient to facilitate accurate production (Hanuliková et al., 2012). Interestingly, Nagle (2018) demonstrated that...
although learners were getting closer to native norms in the production of Spanish voiced stops with the passage of time, their performance still did not attain a native-like level. In addition, no synchronous, time-lagged, or asymptotic relationships were observed between learners’ perception and production of voiced stops.

**Pedagogical Implications**

Research on L3 learning and teaching in China’s multilingual context is lagging behind (Gu et al., 2011; Han et al., 2019). The teaching and learning of L3 stops is an area of difficulty, as Chinese L3 learners face challenges in both the perception and production of stops. It is suggested that language teachers first understand the mapping of phonetic contrasts between learners’ prior languages and the target language. Learners in the initial stages of acquisition are likely to assimilate phonetic contrasts in the target language to existing categories in their prior languages. In the case of the stop system, first a distinction should be made between prior languages and the target language in terms of whether they are aspirating languages or voicing languages, on the basis of which cross-linguistic or combined cross-linguistic influences may be anticipated. Apart from this, to facilitate the teaching and learning of L3 stops, we make the following suggestions on the basis of the findings of this study.

Regarding the acquisition of voiceless stops in L3 Japanese and Russian, our study indicated that both learners’ perception and production of voiceless stops seemed to be closely related to VOT. In the initial stages of acquisition, L3 learners in this study had not yet established a phonetic space for the new sounds in the target language; instead, they were drawing on the short-lag versus long-lag contrasts in their L1 and L2, two aspirating languages, to perceive and produce the voiceless stops in L3 Japanese or Russian. Considering that beginner learners are perceptually sensitive to VOT, it is suggested that we help learners to shift their perceptual VOT boundary to fit the characteristics of L3 (e.g., by associating short-lag VOT with voiceless stops in the target language). Targeted training on L3 perceptual ability should be in place to alert learners to the difference between the target language and prior languages, and to help learners reduce the intensity of airflow in the production of voiceless unaspirated stops in L3 Japanese or Russian. The training would help learners establish a phonetic space for new phones as soon as possible.

Regarding the acquisition of L3 voiced stops, our study indicated that learners had a high accuracy in perceiving L3 voiced stops but experienced difficulty in producing voiced stops. This could be related to a physical and phonetic constraint on speech production (e.g., an aerodynamic challenge). In addition, no correlation was detected between the perceptual accuracy of L3 voiced stops and the production of negative VOT. This is consistent with the results from previous studies (Hanulíková et al., 2012; Nagle, 2018, 2020) which also indicated that perception and production may dissociate during the initial stages of acquisition. The L3 perceptual accuracy of voiced stops could not predict the level of production. Learners’ high accuracy in perceiving voiced stops was not sufficient to promote production performance. Therefore, language teachers should not be surprised if they find that perception training in L3 voiced stops may not effectively improve production.

Meanwhile, Nagle (2020) draws our attention to the benefits of imitation. Imitation could become a bridge between the perception and production of speech sounds; teachers could increase imitation training to help learners improve the production of voiced stops. It is worth noting that target language proficiency will also directly affect the training outcome. For example, perception and production may dissociate in the case of beginner learners, whereas the two modalities may be synchronous (or at least there may be a correlation between perceptual accuracy and production accuracy) in the case of more advanced learners (Flege, 1995, 2003; Flege et al., 1997) Therefore, learners’ target language proficiency should be taken into account when we decide what to teach and how to teach it.

**Conclusion and Future Directions**

In this study we investigated how L1 Mandarin Chinese, L2 English, and L3 Japanese or Russian learners perceived and produced word-initial stops in their respective L3. We aimed to explore the patterns in the perception and production of L3 speech sounds, and the link between the two modalities. Pedagogical suggestions for teaching speech sounds in the multilingual context were put forward on the basis of the results of the study. The learners’ L1 and L2 were both aspirating languages, whereas their L3s were both voicing languages. The results of the study indicated that phonetic similarity in different stop categories between the target language and prior languages was likely to cause confusion in terms of the learners’ perception of the target phones. On the contrary, L3 learners were able to perceive the novel phonetic feature in L3 (i.e., voicing lead), but they had difficulty producing L3 voiced stops, their production being significantly different from that of NS.

Under the conditions of this study, we found that the widely recognized SLA phonetic learning model SLM could also be applied to explain or predict patterns in L3 phonological perception. The correlation analysis showed that there was a positive correlation between the perception and production of L3 voiceless stops in the initial stages of acquisition, but no correlation was found between the perception and production of L3 voiced stops. It is therefore proposed that pronunciation training in the multilingual context should be carried out on the basis of our empirical findings, which will help to accelerate learners’ establishment of a new phonetic space for target phones.

There are some limitations with this study which at the same time suggest directions for future research. First, with the development of target language ability over time, the link between the perception and production of L3 stops may also
undergo change. Perception and production may dissociate (e.g., manifesting a time-lagged relationship as found in Nagle (2018)). A longitudinal study is necessary to shed light on the process of reconstructing phonetic spaces. Second, as the reviewers pointed out, to properly investigate the link between perception and production, it is important to use a battery of appropriate tasks and test both production and perception, as the two modalities are not always linked in a straightforward manner (Hattori & Iverson, 2010; Nagle, 2018, 2020; Piccinini & Arvaniti, 2018). As this study aimed to explore patterns in the perception and production of L3 speech sounds and the link between the two modalities, we tested L2 production and L3 perception and production. We acknowledge the importance of testing all the languages spoken by the learners, as this will provide more insight into the cross-linguistic influences and phonological permeability (e.g., Cabrelli Amaro, 2017). This will be a direction for our future work.

Appendix A

Examples of Phonetic Materials in the Experiment.

| Language | Monosyllable | Disyllable |
|----------|--------------|------------|
| Japanese | ば [pa] ば [ba] ばさ [pasa] ばさ [bas] | は [ka] は [ka] はさ [kasa] はさ [gas] |
| Russian | ぱ [pa] ぱ [ba] ぱさ [pasa] ぱさ [bas] | か [ka] か [ka] かさ [kasa] かさ [gas] |

Note. The phonetic stimuli are in the word-initial syllables.

Appendix B

Screenshot of the perception test using Praat’s ExperimentMFC script.

Acknowledgments

We are very grateful to our lab members Dr. Xiuchuan Lu, Dr. Ting Zeng, and our multilingual research team members Prof. Yongyan Zheng, Dr. Xin Li, and Dr. Yuli Feng for their constructive and insightful comments. We would also like to express our heartfelt gratitude to our four reviewers who provided detailed and specific feedback on our manuscript and encouraged us to revise our work.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research is supported by the National Social Science Foundation of China (Grant Number 18BYY227).

Ethical Statement

The research reported is in compliance with Fudan University’s ethical standards involving human participants.

ORCID iDs

Jiaqi Liu https://orcid.org/0000-0003-3335-7164
Jiayan Lin https://orcid.org/0000-0003-0789-7845

References

Abramson, A. S., & Whalen, D. H. (2017). Voice onset time (VOT) at 50: Theoretical and practical issues in measuring voicing distinctions. Journal of Phonetics, 63, 75–86.

Beckman, J., Helgason, P., McMurray, B., & Ringen, C. (2011). Rate effects on Swedish VOT: Evidence for phonological over-specification. Journal of Phonetics, 39(1), 39–49.

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (Version 6.0) [Computer program]. http://www.fon.hum.uva.nl/praat

Bohn, O. S., & Flege, J. E. (1993). Perceptual switching in Spanish/English bilinguals. Journal of Phonetics, 21(3), 267–290.

Cabrelli Amaro, J. (2017). Testing the phonological permeability hypothesis: L3 phonological effects on L1 versus L2 systems. International Journal of Bilingualism, 21(6), 698–717.

Cabrelli Amaro, J., & Wrembel, M. (2016). Investigating the acquisition of phonology in a third language: A state of the science and an outlook for the future. International Journal of Multilingualism, 13(4), 395–409.

Cenoz, J. (2012). Bilingual and multilingual education: Overview. In C. A. Chapelle (Ed.), The Encyclopedia of applied linguistics (pp. 1–8). Blackwell.

Cenoz, J., Hufeisen, B., & Jessner, U. (2001). Bilingual education and bilingualism: Vol. 31. Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives. Multilingual Matters.

Chamot, A. (1973). Phonological problems in learning English as a third language. International Review of Applied Linguistics, 11(3), 243–250.

Cho, T., & Ladeffoged, P. (1999). Variation and universals in VOT: Evidence from 18 languages. Journal of Phonetics, 27(2), 207–229.

Davidson, L. (2016). Variability in the implementation of voicing in American English obstruents. Journal of Phonetics, 54, 35–50. https://doi.org/10.1016/j.wocn.2015.09.003

De Angelis, G. (2007). Third or additional language acquisition. Multilingual Matters.

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. (1991). Age of learning affects the authenticity of voice onset time (VOT) in stop consonants produced in a second language. Journal of the Acoustical Society of America, 89(1), 395–411.
Flege, J. E. (1995). Second language speech learning: Theory, findings and problems. In W. Strange (Ed.), Speech perception and linguistic experience: Issues in cross-language research (pp. 233–277). York Press.

Flege, J. E. (2003). Assessing constraints on second-language segmental production and perception. In N. O. Schiller & A. S. Meyer (Eds.), Phonetics and phonology in language comprehension and production: Differences and similarities (pp. 319–355). Mouton de Gruyter.

Flege, J. E. (1992). Speech learning in a second language. In C. S. Ferguson & L. Menn (Eds.), Phonological development: Models, research, and application (pp. 565–604). York Press.

Flege, J. E. (1997). Effects of experience on non-native speakers’ production and perception of English vowels. Journal of Phonetics, 25(4), 437–470.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers’ production and perception of English vowels. Journal of Phonetics, 25(4), 437–470.

Flege, J. E., & Hillenbrand, J. (1987). A differential effect of release bursts on the stop voicing judgments of native French and English listeners. Journal of Phonetics, 15(2), 203–208.

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

Flynn, S., Foley, C., & Vin Nichols, K. E. (2004). The cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.

Fukuoka, M. (1995). A longitudinal and cross-sectional study of the acquisition of Japanese voiced-voiceless plosives by speakers of Peking and Shanghai dialects. Journal of Japanese Language Teaching, 87, 40–53.

Gabriel, C., Thulke, J., & Kupisch, T. (2014, September 11–14). A cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.

Flege, J. E., & Hillenbrand, J. (1987). A differential effect of release bursts on the stop voicing judgments of native French and English listeners. Journal of Phonetics, 15(2), 203–208.

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

Flynn, S., Foley, C., & Vin Nichols, K. E. (2004). The cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.

Fukuoka, M. (1995). A longitudinal and cross-sectional study of the acquisition of Japanese voiced-voiceless plosives by speakers of Peking and Shanghai dialects. Journal of Japanese Language Teaching, 87, 40–53.

Gabriel, C., Thulke, J., & Kupisch, T. (2014, September 11–14). A cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers’ production and perception of English vowels. Journal of Phonetics, 25(4), 437–470.

Flege, J. E., & Hillenbrand, J. (1987). A differential effect of release bursts on the stop voicing judgments of native French and English listeners. Journal of Phonetics, 15(2), 203–208.

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

Flynn, S., Foley, C., & Vin Nichols, K. E. (2004). The cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.

Fukuoka, M. (1995). A longitudinal and cross-sectional study of the acquisition of Japanese voiced-voiceless plosives by speakers of Peking and Shanghai dialects. Journal of Japanese Language Teaching, 87, 40–53.

Gabriel, C., Thulke, J., & Kupisch, T. (2014, September 11–14). A cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers’ production and perception of English vowels. Journal of Phonetics, 25(4), 437–470.

Flege, J. E., & Hillenbrand, J. (1987). A differential effect of release bursts on the stop voicing judgments of native French and English listeners. Journal of Phonetics, 15(2), 203–208.

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

Flynn, S., Foley, C., & Vin Nichols, K. E. (2004). The cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.

Fukuoka, M. (1995). A longitudinal and cross-sectional study of the acquisition of Japanese voiced-voiceless plosives by speakers of Peking and Shanghai dialects. Journal of Japanese Language Teaching, 87, 40–53.

Gabriel, C., Thulke, J., & Kupisch, T. (2014, September 11–14). A cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.

Flege, J. E., Bohn, O. S., & Jang, S. (1997). Effects of experience on non-native speakers’ production and perception of English vowels. Journal of Phonetics, 25(4), 437–470.

Flege, J. E., & Hillenbrand, J. (1987). A differential effect of release bursts on the stop voicing judgments of native French and English listeners. Journal of Phonetics, 15(2), 203–208.

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

Flynn, S., Foley, C., & Vin Nichols, K. E. (2004). The cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.

Fukuoka, M. (1995). A longitudinal and cross-sectional study of the acquisition of Japanese voiced-voiceless plosives by speakers of Peking and Shanghai dialects. Journal of Japanese Language Teaching, 87, 40–53.

Gabriel, C., Thulke, J., & Kupisch, T. (2014, September 11–14). A cumulative-enhancement model for language acquisition: Comparing adults’ and children’s patterns of development in L1, L2 and L3 acquisition of relative clauses. The International Journal of Multilingualism, 1(1), 3–16.
Llama, R., & López-Morelos, L. P. (2016). VOT production by Spanish heritage speakers in a trilingual context. *International Journal of Multilingualism*, 13(4), 444–458.

Mathews, J. (1997). The influence of pronunciation training on the perception of second language contrasts. *International Review of Applied Linguistics*, 35(2), 223–229.

Nagle, C. L. (2018). Examining the temporal structure of the production–perception link in second language acquisition: A longitudinal study. *Language Learning*, 68(1), 234–270.

Nagle, C. L. (2019). A longitudinal study of voice onset time development in L2 Spanish stops. *Applied Linguistics*, 40(1), 86–107. https://doi.org/10.1093/applin/amx011

Nagle, C. L. (2020). Revisiting perception-production relationships: Exploring a new approach to investigate perception as a time-varying predictor. *Language Learning*, 60(1), 242–268.

Ohala, J. J. (1997). Aerodynamics of phonology. In *Proceedings of the 4th Seoul international conference on linguistics* (pp. 92–97). Linguistic Society of Korea. http://www.linguistics.berkeley.edu/~ohala/papers/SEOUL2-aerodynamics_of_phonology.pdf

Onishi, H. (2016). The effects of L2 experience on L3 perception. *International Journal of Multilingualism*, 13(4), 459–475.

Piccinini, P., & Arvaniti, A. (2018). Dominance, mode, and individual variation in bilingual speech production and perception. *Linguistic Approaches to Bilingualism*, 9(4–5), 1–31.

Ran, B., Yang, P., & Hu, Z. (2016). A comparative study on category perception of Chinese aspirated and unaspirated bilabial plosives by Russian Chinese learners and Chinese speakers. *TCSOL Studies*, 4, 24–30.

R Core Team. (2014). *R: A language and environment for statistical computing* (Version 3.4.0) [Computer program]. https://www.r-project.org

Ringen, C., & Kulikov, V. (2012). Voicing in Russian stops: Cross-linguistic implications. *Journal of Slavic Linguistics*, 20(2), 269–286. https://doi.org/10.1353/jsl.2012.0012

Rivers, W. M. (1979). Learning a sixth language: An adult learner’s longitudinal study. *Language Learning*, 29(1), 234–270.

Rothman, J., Gonzales Alonso, C., & Puig-Mayenko, E. (2019). *Third language acquisition and linguistic transfer*. Cambridge University Press.

Saito, K., & van Poeteren, K. (2018). The perception–production link revisited: The case of Japanese learners’ English /r/ performance. *International Journal of Applied Linguistics*, 28(1), 3–17.

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.

Shen, Q., & Gao, X. (2019). Multilingualism and policy making in greater China: Ideological and implementational spaces. *Language Policy*, 18, 1–16. https://doi.org/10.1007/s10993-018-9473-7

Shimizu, K. (1993). A cross-language study of phonetic characteristics of stop consonants: With reference to voicing contrasts. *Journal of Asian and African Studies*, 45, 163–176.

Shimizu, K. (2012). A cross-linguistic study on VOT in third language learning by Thai speakers. *Journal of Nagoya Gakuin University: Language and Culture*, 24(1), 61–72.

Singh, R., & Carroll, S. (1979). L1, L2 and L3. *Indian Journal of Applied Linguistics*, 3, 51–63.

Sybińska, J. (2013). Quantity and quality of language use and attrition of L1 Polish due to L2 Danish and L3 English [Unpublished doctoral dissertation]. Adam Mickiewicz University.

Sybińska, J. (2016a). L1 vowels of multilinguals: The applicability of SLM in multilingualism. *Research in Language*, 14(1), 79–94. https://doi.org/10.1515/relna-2016-0003

Sybińska, J. (2016b). Multilingual acquisition of vowels in L1 Polish, L2 Danish and L3 English. *International Journal of Multilingualism*, 13(4), 476–495.

Tremblay, M. (2007, August 3–4). *L2 influence on L3 pronunciation: Native-like VOT in the L3 Japanese of English-French bilinguals* [Paper presentation]. The L3 Phonology Satellite Workshop of ICPhS XVI, Freiburg, Germany.

Wang, M. (2009). A study of Chinese learners’ pronunciation of English stops at intervocalic position. *Modern Foreign Languages*, 2, 186–194.

Wei, R., & Su, J. (2012). The statistics of English in China: An analysis of the best available data from government sources. *English Today*, 28(3), 10–14.

Williams, S., & Hammarbeg, B. (1998). Language switches in L3 production: Implications for a polyglot speaking model. *Applied Linguistics*, 19(3), 295–333.

Wrembel, M. (2010). L2-accented speech in L3 production. *International Journal of Multilingualism*, 7(1), 75–90.

Wrembel, M. (2014). VOT patterns in the acquisition of third language phonology. *Concordia Working Papers in Applied Linguistics*, 5, 750–770.

Wrembel, M. (2015). Cross-linguistic influence in second vs. third language acquisition of phonology. In U. Gut, R. Fuchs, & E.-M. Wunder (Eds.), *Universal or diverse paths to English phonology* (pp. 41–70). Mouton de Gruyter.

Wrembel, M., Marecka, M., & Kopečková, R. (2019). Extending perceptual assimilation model to L3 phonological acquisition. *International Journal of Multilingualism*, 16(4), 513–533.

Wu, Z. (1986). *The spectrographic album of mono-syllables of standard Chinese*. China Social Sciences Press.

Wunder, E.-M. (2010). Phonological cross-linguistic influence in third or additional language acquisition. In K. Dziubalska-Kolaczyn, M. Wrembel, & M. Kul (Eds.), *New sounds 2010: Proceedings of the 6th international symposium on the acquisition of second language speech* (pp. 566–571). Adam Mickiewicz University.

Zeng, T., & Liu, J. (2019). A perception experiment of third language stop consonants: An analysis based on perception model. *Fudan Forum on Foreign Languages and Literature*, 1, 37–45.