Perception of Voiceless Bilabial Stop by Taiwan Mandarin Listeners

Naomi OGASAWARA

Abstract

It has been observed that Mandarin speakers tend to make errors in production and perception of the Japanese voiceless stops with their voiced counterparts, especially in the word-medial position. In this study, a perception experiment using the Japanese voiceless bilabial stop along a VOT continuum in the word-initial and word-medial positions was conducted with three groups of Taiwan Mandarin (TM) listeners at different Japanese proficiency levels (novice, beginning, and intermediate) and a group of native Japanese listeners. It was found that the TM listeners across Japanese proficiency levels were able to perceive [p] even with a short-lag VOT in the initial position, but had difficulty in the medial position when a VOT was short (0 to 25 ms). The results also revealed that the TM listeners' perception varied based on the phonetic contexts, and the interaction of VOT and the contexts might be a reason for the variability.

Keywords: VOT, Japanese stops, Taiwan Mandarin, L2 sound learning models.

1. Introduction

This study was launched by a simple question when Mandarin speakers encounter Japanese stops, why do they often confuse the voiceless stops with their voiced counterparts? Mandarin speakers tend to make errors in production and perception by substituting a voiceless stop with a voiced one in the middle of a word, such as [seito] 'student' as [seido] 'institution' or [kaito:] 'answer' as [kaido:] 'street' (Nishigōri, Komatsu, Ozaki, & Feng 2004). The primary reason for such errors may lie in the phonological difference in stop contrasts between the two languages (Chiang 2010; Fukuoka 1995; Kitamura 1999; Liu 2005; Liu & Nishigōri 2006; Nishigōri, Komatsu, Ozaki, & Feng 2004). Japanese has voicing contrasts (voiced stops /b, d, g/ and voiceless stops /p, t, k/) (Vance, 2008), while, Mandarin has no voiced stops, and voiceless stops contrast in aspiration (aspirated /p^h, t^h, k^h/ and unaspirated /p, t, k/) (Duanmu 2000; Lin 2007). Under the influence of L1 phonology, Mandarin speakers tend to link their native aspirated-unaspirated contrast to the Japanese voiceless-voiced contrast such that they produce aspirated stops for voiceless stops and unaspirated stops for voiced stops (Chiang 2010; Fukuoka 1995; Kitamura 1999; Liu 2005; Liu & Nishigōri 2006; Nishigōri, Komatsu, Ozaki, & Feng 2004).

When L2 listeners hear non-native sounds, phonetic processing of the sounds takes place in order to access phonological categories in the lexicon, and an assessment of phonetic similarities/dissimilarities between non-native and native sounds may play an important role in processing. Many studies have found evidence for such phonetic assessment corresponding to the learnability of non-native sounds (Aoyama, Flege, Guion, Akahane-Yamada, & Yamada 2004; Best, Hallé, Bohn, & Faber 2003; Flege, Schirru, & MacKay 2003; Hallé, Best, & Levitt 1999; Kingston 2003; Sharma & Dorman 2000; Winkler, Lehtoksoki, Alku, Vainio, Czugler, Csepe, Aaltonen, Raimo, Alho, Lang, Iivonen, & Nätäänen 1999). Frequently cited L2 sound learning models, such as the Speech Learning Model (SLM) proposed by Flege (1995; 1999; 2003), the Perceptual Assimilation Model (PAM and PAM-L2) by Best (Best, Hallé, Bohn, & Faber 2003; Hallé, Best, & Levitt 1999, Best & Tyler 2007), and the Native Language Magnet Model (NLM) by Kuhl (Iverson & Kuhl 1995), consider the importance of both contrastive and non-contrastive phonetic similarities and discrepancies in L2 acquisition.

SLM hypothesizes that when a non-native sound is close enough to a native sound in a phonetic space, and as long as it is identified as an exemplar of the native sound category, a new category for the non-native sound will not be formed. Instead, the existing category will be modified over time by merging the phonetic properties of the native and non-native sounds. The new sound is absorbed in the merged category, which makes the sound difficult to produce and perceive like native speakers do. For example, late bilinguals of English and French tended to produce /t/ with VOT length somewhere between the VOT length of French /t/ and English /t/ of monolinguals (Flege 1987).

On the other hand, the model predicts that when a non-native sound is phonetically distinct from an L1 category, a new category will be formed for the non-native sound, and the existing native categories will shift away. Hence, the new sound is mastered relatively well. Aoyama, Flege, Guion, Akahane-Yamada, and Yamada (2004) tested native Japanese speakers' acquisition of the /l/ and /r/ contrast in English which does not exist in Japanese. They found that Japanese children improved better in the production of English /r/ than /l/, and the advantage of English /r/ in learning was attributed to its perceptual dissimilarity from the Japanese rhotic.

NLM further considers phonetic distance within a category and predicts that phonetic variants around a prototype in a native category are difficult to discriminate. This is called a perceptual magnet effect. When non-native sounds are phonetically more similar to a native prototype, the discrimination of the non-native sounds is more difficult (Iverson & Kuhl 1995; Iverson, Kuhl, Akahane-Yamada, Diesch, Tohkura, Kettermann, & Siebert 2003).

Another model, PAM and PAM-L2 proposed by Best (Best, Hallé, Bohn, & Faber 2003; Hallé, Best, & Levitt 1999) has a common ground with SLM; that is, both models agree that adults are capable of perceptual learning of L1 speech sounds over time and of applying this ability to learning L2 sounds. There is, however, a critical difference in PAM from SLM, which is that PAM focuses on articulatory gestures. The model posits that when each sound of a non-native contrast is assimilated to a different native category, discrimination between the non-native sounds is easiest; when both sounds of a contrast are assimilated to a single native category, and if one of them fits phonetically better to the native category than the other does, then, discrimination is relatively easy; but if both sounds fit equally to the native category, then, discrimination is the most difficult (Best, Halle, Bohn, & Faber, 2003; Hallé, Best, & Levitt 1999).

No matter what listeners perceive is phonetic cues or invariant articulatory gestures, one primary phonetic property for Mandarin listeners to process Japanese stops may be voice onset time (VOT), cross-linguistically a primary phonetic feature for stop categorization (Lisker & Abramson 1964; Shimizu 1989). In a cross-linguistic study by Lisker and Abramson (1964), it was reported that VOT of voiceless unaspirated stops was 0-25 ms (short-lag), VOT of voiceless aspirated stops was 60-100 ms (long-lag), and some voiced stops exhibited negative VOT (prevoicing). For the Taiwan Mandarin (TM) stops, VOT of aspirated stops ranged from 75 to 92 ms, which belongs to a long-lag category, and VOT of unaspirated stops ranged from 14 to 27 ms, which belongs to a short-lag category (Chao & Chen 2008; Chen, Chao, & Peng 2007). Since the aspiration contrast is phonemic in their native language, Mandarin speakers must be sensitive to the amount of VOT and are likely to apply the two-way categorization (short-lag or long-lag) in their native language to discriminate voicing in Japanese stops. Listeners may tend to anchor short-lag stops with voiced stops and to anchor long-lag stops with voiceless stops. This anchoring has been observed both in production (Fukuoka 1995; Kitamura 1999) and perception (Fukuoka 1995; Liu 2005; Liu & Nishigōri 2006; Nishigōri, Komatsu, Ozaki, & Feng 2004).

However, this strategy might not always work well for the correct discrimination of non-native stops. Rather, it may cause some confusion due to the complicated nature of VOT in Japanese stops. First, VOT of Japanese voiceless stops, ranging from 25 ms to 57 ms reported in the previous studies, does not fall exclusively into a short-lag or a long-lag category (Harada 2003; Homma 1980; 1981; Shimizu 1996; Riney, Takagi, Ota, & Uchida 2007). Riney, et al. (2007) called this VOT range 'intermediate.' Moreover, the VOT values of voiceless stops are inconsistent and influenced by position in a word. Ogasawara (2011) reported that VOT became almost half as long in the middle of a word (ranging from 13 ms to 23 ms) as at the beginning of a word (ranging from 22 ms to 42 ms). She also examined TM stops and found that VOT of unaspirated stops ranging from 9 ms to 21 ms. Apparently, VOT of word-medial Japanese voiceless stops is closer to that of TM unaspirated stops than aspirated stops with VOT ranging from 50 ms to 78 ms. This would lead Mandarin speakers to categorize word-medial voiceless stops as voiced.

In Fukuoka's perception study (1995), Mandarin speakers and Shanghainese speakers had good performance in discriminating word-initial Japanese stops, but they often confused word-medial voiceless stops with voiced ones (e.g., mishearing [papa] as [paba]). Liu (2005) also reported a similar observation that the discrimination of Japanese voiceless stops by Mandarin speakers and Shanghainese speakers was not as accurate in the medial position as in the initial position. She also found that among the voiceless stops, [k] was the easiest, and [p] was the most difficult for Mandarin speakers because cross-linguistically, as the place of articulation moves further back, VOT becomes longer and easier to recognize as voiceless. The results from these previous studies suggest that VOT is a strong cue for Mandarin (and Shanghainese) speakers when processing Japanese stops, and they apply the strategy of linking relatively long VOT to voiceless stops and short VOT to voiced stops.

For word-medial stops, the ratio of the whole stop length including closure to the adjacent vowel length may be another important cue for discriminating voicing of stops. Lisker (1957) pointed out that the duration of closure and VOT can be a main cue for voicing distinction. He reported that the duration of the intervocalic English [p] (130 ms) was longer than that of [b] (75 ms). Port (1979) reported that the perceptual boundary between [b] and [p] in the words *rabid* and *rapid* was placed at around 75 ms of closure duration in slow speech, and in fast speech the boundary shifted to about 8 ms shorter. For Japanese stops, Muraki and Nakaoka (1990) measured the duration, including the closure of the intervocalic [k] in *jiken* 'incident', at 85.6 ms and its ratios to the preceding and following vowels were 0.93 and 0.63, respectively. On the contrary, those ratios in the same word produced by Mandarin speakers were 2.3 ([k] to the preceding vowel) and 1.75 ([k] to the following vowel) due to their much longer closure duration (186.2 ms) compared with native Japanese speakers' (80.4 ms). In the current study, the closure duration for the word-medial [p] was set to 100 ms. [p] with 100 ms closure sounds natural, as the stop might be heard as a geminate [pp] with a longer closure; according to Homma (1981), the average closure for Japanese /p/ was 77 ms and for /pp/ was 183 ms. Since the main purpose of the current study is to discover the perception of a voiceless stop along with a VOT continuum by Japanese and Taiwan Mandarin listeners, fixing the closure duration lets us focus on the VOT cue.

Although word-initial Japanese voiceless stops are not perfect exemplars of the category of voiceless stops with long-lag VOT because of their 'intermediate' state in VOT (Riney, Takagi, Ota, & Uchida 2007), a relatively sufficient amount of aspiration seems to help Mandarin speakers hear stops as voiceless. As previously mentioned, many studies have found word-medial voiceless stops are more difficult to recognize as voiceless because of a short VOT, but little is known as yet regarding how much VOT is necessary for perception of voiceless stops. In addition, the findings from the previous studies suggest that the position in a word where a voiceless stop occurs might have some influence on the perception of the stop. The interaction of VOT and this context effect, however, is not clear. One aim of the present study is to address this issue. If TM speakers simply rely on the VOT cue, short-lag stimuli would be consistently recognized as voiced no matter where a voiceless stop occurs in a word. Simple anchoring of VOT with voicing contrast would suggest an effect of L1 transfer of phonological contrast. On the other hand, if Mandarin listeners' voicing identification of stimuli changes based on the position of the stop in a word, this would suggest that a contextual effect modulates their VOT perception. Findings of the contextual effect on the degree of phonological contrast in a phonetic space have been reported previously (Ingram & Park 1998; Lisker 1957); Lisker (1957), for example, pointed out that the intervocalic context blurs the phonetic distinction between the English voiced and voiceless stops. Perhaps, a VCV context would make the voiced-voiceless contrast less distinct in Japanese as well since phonetic characteristics of voiceless stops would be minimized due to the influence from neighboring segments such as voicing assimilation, coarticulation, or adjacent vowel duration. If so, in order to identify the intervocalic stop as a voiceless stop, a greater amount of VOT might be necessary in the word-medial position than in the initial position. Thus, this greater VOT would indicate the contextual effect on VOT perception and suggest that TM listeners' perception of Japanese voiceless stops might be modified by this effect.

2. Perception Experiment

This experiment examined TM listeners' perception of a Japanese voiceless stop along a VOT continuum (0 to 120 ms) presented in the word-initial and word-medial positions. Three groups of TM listeners at different Japanese proficiency levels identified a voiceless bilabial stop [p] with a different VOT and their results were compared with those of native Japanese listeners. As the previous studies suggest, it was expected that TM listeners' performance would be influenced by the contexts and the identification of the voiceless stop would be more difficult in the word-medial position for the learners of Japanese.

2.1. Materials

A female native Japanese speaker (the author) recorded two nonsense words [pago] and [gopa] in a sound-attenuated booth, using a high quality microphone and a Protool digital recorder Version 7.1, sampled at 44.1 KHz, 16 bit. Each nonsense word was produced with two different pitch patterns, accented (High-Low) and unaccented (Low-High). Japanese uses pitch accents, and it has been reported that pitch accent affects VOT, such that VOT in an accented syllable is longer than in unaccented syllable (Homma 1981) although such VOT differences in the stimuli used for this experiment were small. Table 1 shows the phonetic information of the original recordings. VOT of initial [p] was 55.9 ms and 51.7 ms in the accented syllable and unaccented syllable, respectively, and VOT of medial [p] was 18.6 ms and 15.2 ms in the accented syllable and unaccented syllable, respectively.

Table 1. Duration and pitch at three points (onset-middle-offset) in [pago] and [gopa] (HL/LH)

Stimulus	[p] VOT	[a]		[g]	[o]	Whole
[pago]-HL	55.9ms	162.7ms		62.3ms	215.7ms	613.5ms
Pitch (Hz)		282–279–267			176–151–145	
[pago]-LH	51.7ms	179.7 ms		47.0 s	226.8ms	615.8 s
Pitch (Hz)		207–190–192			236–256–248	
Stimulus	[g]	[o]	Closure	[p] VOT	[a]	Whole
[gopa]-HL	53.5ms	156.9ms	114.5ms	18.6ms	216.0ms	557.6 s
Pitch (Hz)		266–296–302			226–182–164	
[gopa]-LH	20.2ms	166.4ms	95.4ms	15.2ms	179.3 ms	489.3 s
Pitch (Hz)		176–168–164			250–242–243	

For the HL stimuli, the onset of the F0 contour was manipulated to set at around 300 Hz in the first syllable and it moved downwards from 200 Hz to 150 Hz in the second syllable. For the LH stimuli, F0 moved downwards from around 200 Hz at the onset to around 180 Hz at the offset in the first syllable and it remained at 250 Hz in the second syllable. These F0 settings were chosen to try to be close to the average of the originals in order to sound natural. The VOT of each nonsense word was manipulated to create a VOT continuum varying in 5 ms steps from 5 ms to 100 ms along with stimuli with no burst (0 ms), with only burst (1 ms), VOT 110 ms, and 120 ms. Tokens with a VOT shorter than the original length were prepared by cutting VOT out at a zero crossing from the original recording, and tokens with a VOT longer than the original length were prepared by inserting frication noise at a zero crossing. Tokens with VOT of 0 ms were created by eliminating burst and frication noises, and tokens with only burst were created by eliminating only frication noises. The number of tokens for each continuum was 24, and the total number of tokens was 96 (2 nonsense words ×2 pitch patterns ×24 stimuli). In order to eliminate any noise during closure, a 100 ms silence was added before the beginning of [p] in each token as a closure. This duration was close to the average closure duration of the originals (114.5 ms and 95.4 ms). All the manipulation was conducted using Praat (Boersma & Weenink 2005). Figures 1a through 1d below illustrate the examples of the tokens.

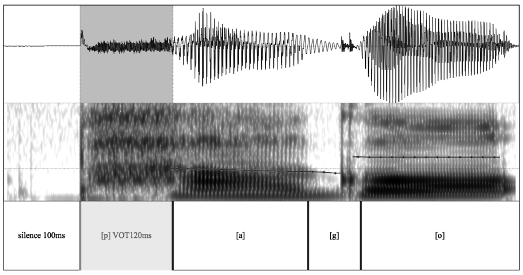

Figure 1a. Token [pago] (LH) with a VOT 120 ms

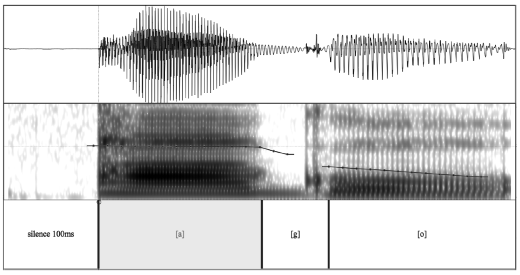

Figure 1b. Token [pago] (HL) with a VOT 0 ms

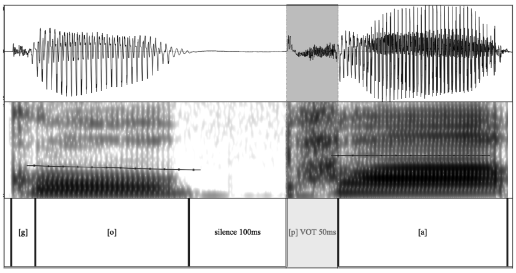

Figure 1c. Token [gopa] (LH) with a VOT 50 ms

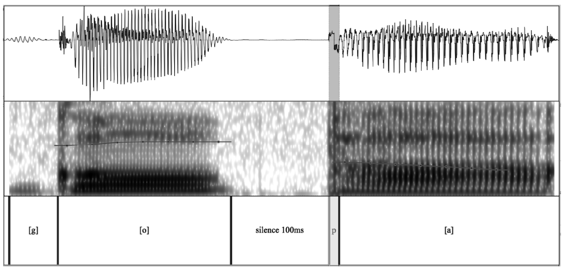

Figure 1d. Token [gopa] (HL) with a VOT 10 ms

2.2. Participants

Four groups of listeners participated in the experiment. None of them reported speech or hearing disorders. All participants were recruited in Taipei City, Taiwan. One group was native Japanese adults (n=33) who were learners of Mandarin, and their average length of living in Taiwan was 5.7 months at the time of the experiment. Their proficiency of Mandarin was beginning or intermediate level. Other groups were native speakers of Taiwan Mandarin, all of whom were university students. Taiwanese participants were divided in three groups based on their proficiency of Japanese: novice (n=37)−no prior training or knowledge of Japanese, beginning (n=31)−length of leaning Japanese was less than one year, and intermediate (n=31)−length of learning Japanese was more than 2 years or their Japanese proficiency was equivalent to two years study. Most of the Taiwanese participants were English majors except four subjects from the beginning group and eight subjects from the intermediate group. All participants received a small amount of money or credits for a course.

2.3. Procedure

The materials were pseudo-randomized, and two lists of materials with different orders of all the stimuli were created. Participants were randomly assigned one of the lists. Each participant was tested individually in a sound-attenuated booth. They were seated in front of a computer connected to a response box. E-Prime software (Psychology Software Tools, Inc.) was used to run the experiment and record responses. Participants wore high quality binaural headphones for listening to the stimuli. Oral instructions in Japanese were given to the native Japanese participants and oral instructions in Taiwan Mandarin were given to the Taiwanese participants by native speakers of TM.

Participants were asked to listen to each stimulus containing [p] and judge it in three ways: (1) the target stop was a good exemplar of [p]; (2) it sounded like [p] but with too much aspiration (excessively aspirated); or (3) it sounded like [b]. Participants were asked to press one button on the response box when they judged the stimulus as (1) or to press another button when they judged the stimulus as (2). They were asked not to press any button when their judgment was (3). This way of judgment created a two-alternatives forced choice if the subjects perceived [p], and a Go/No-Go task if they did not hear [p], both of which are commonly used in behavioral science (Shenoy & Yu 2012). This was a simpler task than a three-alternatives forced choice task, such that this method could minimize confusion and errors by the participants. The written criteria for sound identification were displayed on the computer monitor during the experiment. The maximum response window was five seconds from the onset of each stimulus, but when participants pressed a button, the next stimulus was played after one second. Prior to the test phase, participants listened to 20 of the experimental items[1] in the practice phase to become familiar with the stimuli and the task. Participants were asked to fill out a questionnaire about their own and parents' linguistic backgrounds at the completion of the experiment.

2.4. Results

The total number of data points was 12,672 (96 stimuli × 132 participants) and the number of [p] responses was 10,968 (86.6%) including judged as 1 (a good exemplar of [p]) and judged as 2 ([p] with too much aspiration) for the four VOT continua, [pago]-HL, [pago]-LH, [gopa]-HL, and [gopa]-LH. A generalized estimating equations analysis was performed with the two-way categorical data (identified [p]/[b]) in order to examine effects of the context (word-initial and word-medial), pitch of the target mora (H and L), subjects' Japanese proficiency (novice, beginning, intermediate, and native), and VOT values as main factors. A three-way interaction of proficiency, context, and VOT, and a two-way interaction of context and VOT were also added as factors. The analysis revealed that context (Wald $x^2(1)=28.16; p<.001$), VOT (Wald $x^2(1)=131.23; p<.001$), the three-way interaction (Wald $x^2(6)=22.74; p=.001$), and the two-way interaction (Wald $x^2(1)=28.15; p<.001$) were all significant except pitch (Wald $x^2(1)=1.44, p>.1$) and language proficiency (Wald $x^2(3)=4.39; p>.1$). Table 2 summarizes the parameter information.

Figures 2a and 2b below illustrate the frequencies of [p] responses sorted by groups in the initial and medial positions, respectively. The frequencies were calculated by collapsing the pitch factor, which was found insignificant in the analysis.

Table 2. Parameter estimates

Parameter	b	SE b	95% Cl		Wald x^2	df	Sig.
Proficiency = 1	−.127	.253	−.623	.370	.250	1	.617
Proficiency = 2	−.272	.293	−.846	.302	.861	1	.353
Proficiency = 3	−.508	.257	−1.011	−.006	3.927	1	.048
Context = 1	.648	.122	.409	.887	28.163	1	.000
Pitch = 1	.076	.0636	−.048	.201	1.446	1	.229
VOT	.076	.003	.009	.020	27.025	1	.000
[Proficiency = 1] * [Position = 1] * VOT	−.004	.004	−.011	.003	1.201	1	.273
[Proficiency = 1] * [Position = 2] * VOT	.006	.004	−.002	.013	1.957	1	.162
[Proficiency = 2] * [Position = 1] * VOT	.001	.005	−.008	.010	.042	1	.838
[Proficiency = 2] * [Position = 2] * VOT	.009	.005	.000	.018	3.832	1	.050
[Proficiency = 3] * [Position = 1] * VOT	.002	.004	−.006	.010	.276	1	.600
[Proficiency = 3] * [Position = 2] * VOT	.012	.004	.004	.021	7.888	1	.005
[Proficiency = 4] * [Position = 1] * VOT	−.013	.003	−.019	−.007	17.364	1	.000

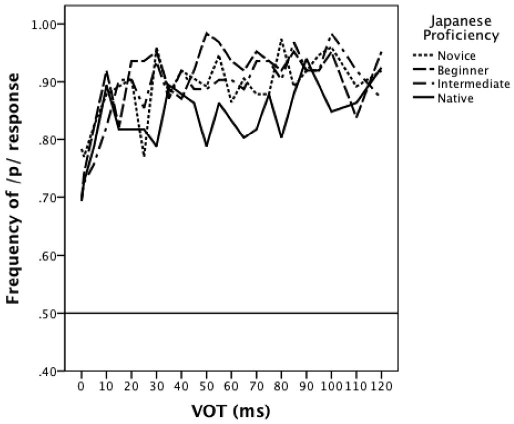

Figure 2a. [p] responses in [pago]

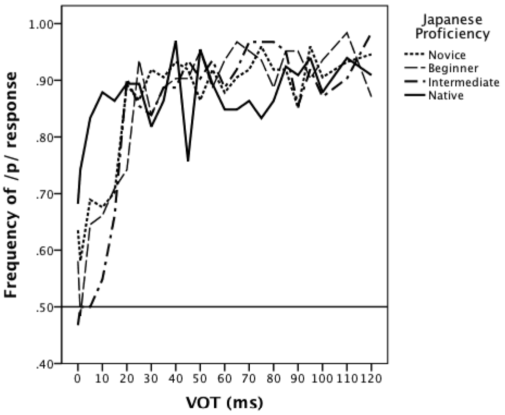

Figure 2b. [p] responses in [gopa]

Note that in Figure 2a, the TM listeners' identification rate of [p] was always above 70% through the VOT continuum, and it became slightly higher than the Japanese listeners' rate after VOT longer than 40 ms. Surprisingly, the TM listeners did not seem to have any difficulty in hearing [p] even with a short-lag VOT (0 to 25 ms) in the initial position. On the other hand, in the medial position (see Figure 2b) hearing [p] seemed more difficult for TM listeners compared with their performance in the initial context and also compared with native Japanese listeners' performance when a VOT was short. For the stimuli with a VOT less than 25 ms, TM listeners identified them as voiceless less than 70% of the time regardless of their Japanese proficiency level, and interestingly, the intermediate group showed the lowest frequency. On the other hand, native Japanese listeners maintained a high frequency through the VOT continuum. These results indicate that the contexts influenced the TM listeners' perception of voicing contrast.

In order to confirm the contextual effect on VOT perception, the frequencies of [p] responses to the stimuli with a VOT ranging from 0 to 25 ms were further analyzed by collapsing the pitch factor. Figure 3 shows the mean frequency of [p] responses sorted by the context and clustered by Japanese proficiency.

A series of analysis of variances (ANOVA) on the frequency of [p] responses was carried out with context, Japanese proficiency, and counterbalanced group as main factors. The context factor was a within-subjects factor and Japanese proficiency and counterbalanced group were between-subjects factors. The main effect of context ($F(1, 124)=27.95; p<.001$) and the interaction between context and proficiency ($F(3, 124)=5.88; p=.001$) were significant, but

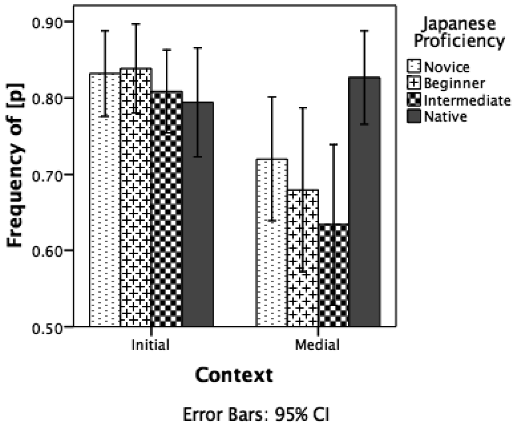

Figure 3. Mean frequencies of [p] responses to the stimuli with a short-lag (0～25 ms). VOT sorted by context (word-initial and word-medial) and clustered by subjects' Japanese proficiency (novice, beginner, intermediate, and native).

the main effect of proficiency did not reach significance ($F(3, 124)=1.26; p>.1$). In order to examine a simple effect of context for each proficiency group, the data were split across the proficiency factor. A simple effect of context was significant for all Taiwanese subject groups: for the novice group ($F(1, 35)=8.82; p=.005$); for the beginning group ($F(1, 29)=18.94; p<.001$); and for the intermediate group ($F(1, 29)=13.12; p=.001$); however, it was insignificant for the native Japanese group ($F(1, 31)=1.55; p>.1$). As Figure 3 and the statistical results revealed, the Taiwanese participants showed more difficulty in hearing [p] in the medial position than in the initial position across their Japanese proficiency levels when VOT was shorter than 25 ms; on the contrary, Japanese participants had no problem in hearing [p] with a short VOT in both positions. Furthermore, pair-wise comparisons were carried out to compare the performance of the Taiwanese groups with that of the Japanese group. The average frequencies of all Taiwanese participants and of native Japanese participants were compared, and one-way ANOVAs confirmed that Taiwanese participants' performance was as good as Japanese participants' in the initial position ($F(3, 131)=.48; p>.5$); nevertheless, Taiwanese participants' performance became poorer than Japanese participants' ($F(3, 131)=3.43; p<.05$) in the medial position.

Next, the frequencies of excessively aspirated (EA) [p] responses were analyzed in order to investigate whether the listeners would be sensitive to the amount of aspiration within the same phoneme. The number of data points obtained for EA [p] was 4,362 (39.8%). A binary logistic regression analysis was performed in order to examine the effect of the context (word-initial and word-medial), pitch of the target mora (H and L), subjects' Japanese proficiency (novice, beginner, intermediate, and native), and VOT values as main factors. The analysis[2] revealed that proficiency (Wald $x^2=363.94; p<.001$), context (Wald $x^2=11.43; p=.001$), and VOT (Wald $x^2=1075.31; p<.001$) were all significant, but pitch was insignificant (Wald $x^2=2.65; p>.5$). Figures 4a and 4b illustrate correlation of the frequency on the VOT continuum for each group in the initial and medial positions. Frequencies were calculated by collapsing the pitch factor.

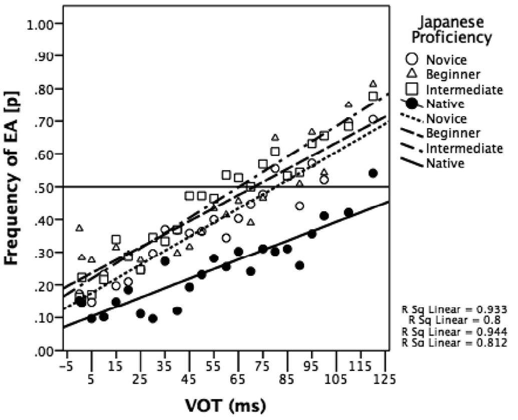

Figure 4a. Mean frequencies of EA [p] responses on the VOT continuum in the word-initial position for each group. The lines on the figure show correlation between the frequency and VOT

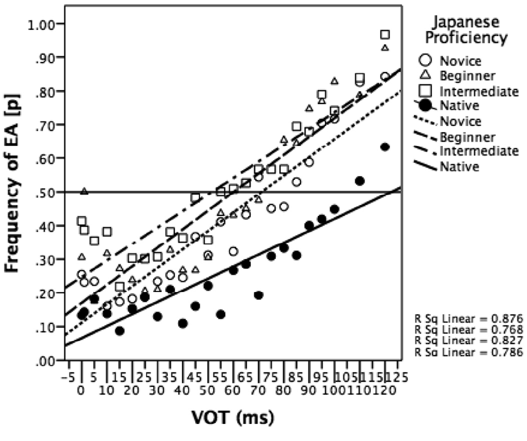

Figure 4b. Mean frequencies of EA [p] responses on the VOT continuum in the word-medial position for each group. The lines on the figure show correlation between the frequency and VOT.

In both figures, the correlation lines for Taiwanese groups are similar. In the word-initial context (Figure 4a), the frequency passed 50% at around VOT 65～85 ms, and in the word-medial context (Figure 4b), it passed 50% at around VOT 50～70 ms. These VOT values in perception correspond to VOT values of TM aspirated stops, 62 ms to 78 ms word-initially and 50 ms to 68 ms word-medially, reported in Ogasawara (2011). Stimuli with a VOT longer than the VOT of TM aspirated stops were judged as excessively aspirated exemplars by the TM listeners, which suggests that their perception of the Japanese stop VOT was based on their native category without adjusting their perception for the non-native sound when the task demanded they focus on the amount of aspiration. In addition, the Japanese listeners showed a different performance on judging the EA stimuli from the non-native groups. Neither word-initially nor -medially, the frequency of EA [p] crossed 50% throughout the continuum, which indicates that the native listeners were not very sensitive to the amount of VOT. Since Japanese stop categorization is based on voicing instead of aspiration, it is not surprising that they were not able to hear VOT difference among the stimuli within the [p] category.

3. General Discussion

This study examined the perception of the Japanese voiceless stop [p] by TM listeners and Japanese listeners. In the perception experiment, the Japanese [p] stimuli along the VOT continuum (0 to 120 ms) were tested with three TM groups at a different level of Japanese proficiency (novice, beginner, and intermediate) and a native Japanese group. Some points were discovered. First, there appeared to be a robust contextual effect on the perception of [p] for the TM listeners but not for the Japanese listeners. The TM listeners across Japanese proficiency levels had difficulty in hearing [p] in the word-medial position when VOT was short, but they were able to hear [p] even with a short VOT in the initial position. On the contrary, native listeners showed consistently good performance in hearing [p] regardless of the contexts. Next, the TM listeners were sensitive to the length of VOT within a category, but the native listeners were not. The TM listeners judged the stimuli with a VOT longer than that of their native aspirated stops as excessively aspirated [p]; whereas the native listeners failed to judge them as excessively aspirated even with a VOT over 100 ms.

It was unexpected that the TM listeners were able to successfully identify the initial [p] with a short-lag VOT. One possible reason might be that they were already familiar with the voicing contrast of stops in the word-initial position. Many of the participants are fluent in English and some of them understand Taiwan Southern Min, both of which have a voicing contrast. In normal speech, stops occurring in the word-initial position in English and Taiwan Southern Min have a shorter VOT than that of the TM aspirated stops. As for the VOT in American English, Lisker and Abramson (1964) reported a range from 20 to 120 ms (58 ms in average) for [p] in words in isolation. The average VOT of [p^h] of Taiwan Southern Min in adults' speech was 39.37 ms and VOT of [p] was 10.24 ms in Tai's study (2009). Since the initial voiceless stops in those two languages are different from the TM aspirated stops in terms of VOT, perhaps the TM listeners might have already learned the new category of voiceless stops with a shorter VOT and established a strategy for correct perception as SLM (Flege 1995; 1999) predicts. Whereas, medial Japanese stops usually have a very short VOT (e.g., 13-23 ms in Ogasawara (2011)), which is analogous to the VOT of the TM unaspirated stops. Hence, it can be assumed that the Japanese stops with such a short VOT occurring in the word-medial position are absorbed into the unaspirated stop category. The TM listeners thus have not acquired a good perception strategy to successfully perceive the Japanese stops as voiceless.

Besides TM listeners' familiarity with the voicing contrast in the initial position, there may be some acoustic properties of stops other than VOT that might affect perception in different contexts. Lisker stated, "From the accepted phonetic descriptions we have every right to expect that in post-stressed intervocalic position the voiced-voiceless distinction is phonetically minimal." (1957: 43). Lisker (1986) listed 16 acoustic cues for distinguishing the English voiced and voiceless bilabial stops in the intervocalic position, such as closure duration, the duration of the preceding vowel and F1 transition, release burst intensity, and so on (see Lisker 1986, pp.5–6 for the complete list), some of which were found interacting with VOT. It is thus possible that if listeners do not make good use of those acoustic cues for the perception of medial stops, a longer VOT might be required to perceive voiceless stops. Lisker, Liberman, and Erickson (1977) and Summerfield and Haggard (1977) found that F1 transition interacts with VOT in CV syllables in the way that a longer VOT is necessary for perception of voiceless stops when voiced transition duration (VTD) is longer or F1 transition is lower. If this holds in Japanese, the F1 values in the onset of the following vowel [a] in the current study were in average 1000 Hz in the initial position ([pago]) and 793.7 Hz in the medial position ([gopa]) which needed a longer VOT. This might explain the reason why the TM listeners had more difficulty in recognizing the Japanese [p] with a short-lag VOT in the medial position. Since the focus of this study was not to investigate the interactions of different acoustic cues for stop perception, further study is needed to examine this issue.

In sum, it was found that the perception of non-native stops is influenced not only by phonetic similarity between non-native and native sounds but also by the context effect. In the case of Japanese stop perception by Mandarin listeners, the results of the current study were compatible with previous studies which found that Mandarin listeners are successful in perceiving voiceless stops in the word-initial position but not in the medial position. This study further suggests that not because the listeners have an ability to perceive a voiceless stop in the initial position, even when a VOT is short, but when a stop occurs intervocalically, a longer VOT is necessary for the perception of voiceless stops, which indicates a shift of perception by the phonetic contexts. Although the existing cross-linguistic speech learning models such as SLM (Flege 1995; 1999), PAM/PAM-L2 (Best, Hallé, Bohn, & Faber 2003; Hallé, Best, & Levitt 1999), and NLM (Iverson & Kuhl 1995) do not exclude allophonic variations, subcategories for allophones might need to be included in the models. In addition, how the learning of a non-native phoneme across allophonic categories is achieved should be taken into consideration. In other words, these models need to consider the perceptibility of non-native allophones with a different degree in an acoustic feature (e.g., the amount of VOT) and thus, which exhibit different acoustic similarity to a native sound based on the contexts.

4. Conclusions

In conclusion, this study revealed some important points regarding the perception of the Japanese voiceless bilabial stop by native and TM listeners. Firstly, the non-native listeners' perception of the stop varied in different contexts, but this was not the case for the native listeners. Second, it seems that the non-native listeners' perception of the stop is not based on simple anchoring of VOT and voicing; instead, it is influenced by the interaction of VOT and the phonetic contexts.

Acknowledgements

I would like to thank Natasha Warner for comments on this material. My gratitude also goes to Ya-Hui Lin, Chia-Lin Shih, Ai Nagata, Hsiang-Ying Lo, Zi-Jia Huang, and Hui-Lian Chen for assisting the experiments. This research was supported by Grants for New Researchers, National Science Council, Taiwan (NSC 98-2410-H-003-035 and NSC 99-2410-H-003-072). All errors are my own.

Endnotes

1 Their VOT was 0, 20, 50, 80, or 120 ms for each of the four continua: [pago] with HL pitch and LH pitch and [gopa] with HL pitch and LH pitch.

2 $R^2 = .011$ (Hosmer & Lemeshow), .136 (Cox & Snell), .183 (Nagelkerke), Model $x^2(6) = 1598.35, p < .001$.

References

Aoyama, K., Flege, J. E., Guion, S. G., Akahane-Yamada, R., and Yamada, T. (2004). "Perceived phonetic dissimilarity and L2 speech learning: the case of Japanese /r/ and English /l/ and /r/," Journal of Phonetics, 32, 233–250.

Best, C. T., Hallé, P., Bohn, O.-S., and Faber, A. (2003). "Cross-language perception of nonnative vowels: phonological and phonetic effects of listeners' native languages," in *Proceedings of the 15th International Congress of Phonetic Sciences*, edited by M. Solé, D. Recasens, and J. Romero (Casual Productions, Barcelona), pp.2889–2892.

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

Boersma, P., and Weenink, D. (2005). Praat: doing phonetics by computer (Version 5.3.69) [Computer program]. Retrieved from http://www.praat.org/ (date last viewed 03/31/14)

Chao, K-Y., and Chen, L-M. (2008). "A cross-linguistic study of voice onset time in stop consonant productions," Computational Linguistics and Chinese Language Processing, 13, pp.215–232.

Chen, L-M., Chao, K-Y., and Peng, J-F. (2007). "VOT productions of word-initial stops in Mandarin and English: A cross-language study," in *Proceedings of the 19*th Conference on Computational Linguistics and Speech Processing, pp.303–317.

Chiang, P-H. (2010). "The production of Japanese intervocalic /t/ and /d/ by Taiwanese and Shang-Hai dialect speaking learners: with a special focus on durational control" (in Japanese), Taida nihongobun kenkyū, 19, 173–196.

Duanmu, S. (2000). *The Phonology of Standard Chinese* (Oxford University Press, Oxford), pp.9–96.

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, pp.47–65.

Flege, J. E. (1989). Chinese subjects' perception of the word-final English /t/-/d/ contrast: Performance before and after training. Journal of Acoustical Society of America, 86, pp.1684–1697.

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

Flege, J. E. (1999). "Age of learning and second-language speech," in *Second language acquisition and the critical period hypothesis*, edited by D. P. Birdsong (Lawrence Erlbaum, Hillsdale, NJ), pp.101–132.

Flege, J. E., Schirru, C., and MacKay, I. R. A. (2003). "Interaction between the native and second language phonetic subsystems," Speech Communication, 40, pp.467–491.

Fukuoka, M. (1995). "A cross-acquisition study of Japanese voiced and voiceless plosives targeting native

Mandarin and Shanghainese speakers" (in Japanese), Nihongo kyōiku, 9, pp.201-215.

Hallé, P., Best, C. T., and Levitt, A. (1999). "Phonetic vs. phonological influences on French listeners' perception of American English approximants," Journal of Phonetics, 27, pp.281–306.

Harada, T. (2003). "L2 influence on L1 speech in the production of VOT," in *Proceedings of the 15th International Congress of Phonetic Sciences*, edited by M. Solé, D. Recasens, and J. Romero (Casual Productions, Barcelona), pp.1085-1088.

Homma, Y. (1980). "Voice onset time in Japanese stops," Onsei Gakkai Kaihō, 163, pp.7-9.

Homma, Y. (1981). "Durational relationships between Japanese stops and vowels," Journal of Phonetics, 9, pp.273-281.

Ingram, J. C. L., and Park, S.-G. (1998). "Language, context, and speaker effects in the identification and discrimination of English /r/ and /l/ by Japanese and Korean listeners," Journal of the Acoustical Society of America, 103, pp.1161-1174.

Iverson, P., and Kuhl, P. K. (1995). "Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling," Journal of the Acoustical Society of America, 97, pp.553–562.

Kingston, J. (2003). "Learning foreign vowels," Language and Speech, 46, pp.295–349.

Kitamura, Y. (1999). "Acoustic analysis using soundscope: Japanese stops produced by Chinese speakers" (in Japanese), Tōkaidaigaku kiyō ryūgakusei kyōiku center, 19, pp.69-71.

Lin, Y-H. (2007). *The Sounds of Chinese* (Cambridge University Press, Cambridge), pp.19-120.

Lisker, L. (1957). "Closure duration and the intervocalic voiced-voiceless distinction in English," Language, 33, pp.42-49.

Lisker, L. (1986). ""Voicing" in English: A catalogue of acoustic features signaling /b/ versus /p/ in trochees," Language and Speech, 29, pp.3-11.

Lisker, L., and Abramson, A.S. (1964). "A cross-language study of voicing in initial stops: Acoustical measurements," Word, 20, pp.384-422.

Lisker, L., Liberman, A. M., and Erickson, D. M. (1977). "On pushing the voice-onset-time (VOT) boundary about," Language and Speech, 20, pp.209-216.

Liu, J.-Q. (2005). "A study on perception of Japanese voiced and voiceless plosives by Chinese (Mandarin and Shanghainese) dialect speakers: Focusing on beginning learners" (in Japanese), Wasedadaigaku nihongo kyōiku kenkyū, 6, pp.79-90.

Liu, Y., and Nishigōri, J. (2006). "Perception and learning of Japanese voiced and voiceless consonants by Chinese/Shanghainese learners: Learning effectiveness of explanation and repetition drill" (in Japanese), Nihongo kenkyū, 26, pp.75-84.

Muraki, M., and Nakaoka, N. (1990). "Coda nasal and geminate: Pronunciation of English and Chinese speakers (in Japanese)," in *Kōza nihongo to Nihongo kyōiku 3 Nihongo no onsei onin Vol.2*, edited by Sugitō, M. (Meiji Shoin, Tokyo), pp.139-177.

Nishigōri, J., Komatsu, K., Ozaki, W., and Feng, Q.-Y. (2004). "Study on perception of voiced and voiceless sounds by beginning Chinese learners of Japanese: Development of multi media educational material and its learning effectiveness" (in Japanese), Nihongo kenkyū, 24, pp.31-45.

Ogasawara, N. (2011). "Acoustic analysis of voice-onset time in Taiwan Mandarin and Japanese," Concentric: Studies in Linguistics, 37, pp.155-178.

Riney, T. J., Takagi, N., Ota, K., and Uchida, Y. (2007). "The intermediate degree of VOT in Japanese initial voiceless stops," Journal of Phonetics, 35, pp.439-443.

Sharma, A., and Dorman, M. F. (2000). "Neurophysiologic correlates of cross-language phonetic perception," Journal of the Acoustical Society of America, 107, pp.2697–2703.

Shenoy, P., & Yu, A. J. (2012). Strategic impatience in Go/NoGo versus forced-choice decision-making. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.), Advances in Neural Information Processing Systems, 25 (pp.2123-2131). NY: Curran Associates, Inc.

Shimizu, K. (1989). "A cross-language study of voicing contrasts of stops," Studia phonologica, 23, pp.1-12.

Shimizu, K. (1996). *A Cross-Language Study of Voicing Contrasts of Stop Consonants in Six Asian Languages* (Seibido, Tokyo), pp.1-205.

Summerfield, Q., and Haggard, M. (1977). "On the dissociation of spectral and temporal cues to the voicing distinction in initial stop consonants," Journal of the Acoustical Society of America, 62, pp.435–448.

Tai, Y-L. (2009). *A Phonetic Study of the Acquisition of Voicing in Taiwan Southern Min* (Master thesis, National Chung Cheng University, Taiwan).

Vance, T. J. (2008). *The Sounds of Japanese* (Cambridge University Press, NY), pp.74-112.

Winkler, I., Lehtoksoki, A., Alku, P., Vainio, M., Czugler, I., Csepe, V., Aaltonen, O., Raimo, I., Alho, K., Lang, H., Iivonen, A., and Nätäänen, R. (1999). "Pre-attentive detection of vowel contrasts utilizes both phonetic and auditory memory representations.," Cognitive Brain Research, 7, pp.357–369.