名古屋学院大学論集　言語・文化篇　第 24 巻　第 2 号　pp. 15-22

Voice Onset Time and the Perception of Japanese Voicing Contrasts

Ian Wilson, Yurika Hashimoto

Abstract

Much crosslinguistic research exists on the production and perception of voice onset time (VOT). However, most research on the perception of VOT uses synthetic stimuli instead of natural speech stimuli. Effects of synthetic speech on the perception of VOT are not known, but more research needs to be done to see if there are differences between perception using synthetic speech and perception using natural speech. This pilot study uses natural speech to investigate perception of Japanese VOT by Japanese listeners. Results clearly show that not just VOT, but other phonetic factors too must be responsible for differences in perception of the voiced-voiceless distinction in Japanese word-initial stop consonants.

1 Introduction

It is perhaps presumptuous to write an article on voice onset time (VOT) in a festschrift to Professor Katsumasa Shimizu. After all, he has written prolifically on VOT and it is difficult to believe that much more can be contributed. However, this paper contributes to Japanese VOT research by focusing on perception (as opposed to production) and by using real words (as opposed to synthetic sounds) as stimuli. The main research question that we investigated is whether or not it is feasible to use real words when testing Japanese native speakers' perceptual VOT boundary between phonologically voiced and voiceless stops. We did a pilot study to investigate this for undergraduate native speakers of Japanese, and also did an initial comparison of perception of VOT in Sino-Japanese words to loanwords.

As Shimizu (1977) stated, "the distinction between voiced and voiceless stop consonants is made not only by the existence or absence of laryngeal pulsing but also by voice onset time" (p. 25). In the production of word-initial stops, there are four salient temporal and spectral properties to distinguish voiced stops from voiceless ones: VOT, pitch (f0) at vowel onset (and f0 contour), release burst intensity, and F1 onset frequency and transition period (Shimizu, 1996, p. 9). In the perception of word-initial stops, on the other hand, most research has found that VOT and F1 transition are the most salient cues for a listener to distinguish voiced from voiceless stops (Shimizu, 1996). A recent study

(Kong et al., 2012) has shown that Japanese adults' perception of Japanese children's VOT productions cannot be explained by VOT alone, but can be greatly improved by factoring in f0 and spectral tilt (H1−H2) values.

1.1 Synthetic versus natural speech stimuli

It should be noted that most research on the perception of VOT has used synthetic stimuli (e.g., Abramson & Lisker, 1973; Caramazza et al., 1973; Itoh et al., 1986; Shimizu, 1977; Williams, 1977). However, Elman et al. (1977) objected to the use of synthetic nonsense syllables on the grounds that they "are not optimal stimuli for maintaining a language set" (pp. 971−972). After finishing their research project and obtaining different results from previous studies, Elman et al. (1977) attempted to replicate their results using synthetic syllables instead of natural speech syllables. They failed to get the same results, leaving us to believe that the type of stimuli (synthetic versus natural) could have an effect on the perceived voicing category of those stimuli.

Shimizu (1977) investigated the perceptual VOT boundary for the voiced-voiceless distinction of syllable initial stops. His participants were 12 undergraduate students at Nagoya Gakuin University (2 female, 10 male, ages 18−22). The mean VOT boundary for ba−pa was 18ms, for da−ta it was 26ms, and for ga−ka it was 26ms. In Shimizu's study, the stimuli were all synthetic speech sounds prepared at Haskins Laboratories in New Haven, CT, USA. Itoh et al. (1986) also investigated the VOT boundary for distinguishing voiced and voiceless stops using synthetic speech stimuli. Their participants included aphasic patients and also healthy listeners. Among the healthy listeners, Itoh et al. found that the mean VOT boundary for ga−ka was 24.4ms, a value quite close to that found by Shimizu (1977).

1.2 Trend towards longer Japanese VOT

Takada (2011) reported findings suggesting that longer VOT values in Tohoku dialects have been spreading south and west over the last 50 years. This trend toward longer VOT values could be due to a spread from Tohoku, but it is perhaps more likely that English, with its longer VOT values, has been influencing the pronunciation of Japanese (T. Vance, personal communication, October 15, 2012). Japanese students study English for 6 years in junior high school and senior high school, and many of them are exposed to more English in university and via television and the internet. They are also exposed to many more English loanwords now than ever before.

This trend toward longer VOT values for both voiced and voiceless stops could be apparent in more proficient EFL speakers, or it could be apparent in younger speakers in general. It could also be apparent in speakers from Tohoku versus speakers from farther south in Japan. Takada (2004) found a correlation ($R^2 = 0.425$) between birth year and percentage of+VOT production for /d/. Specifically, VOT *production* for /d/ was more often positive for younger speakers than for older ones. On the other hand, VOT *perception* had no such systematic differences across different age groups. This result was

the same as what Itoh et al. (1986) found when they compared young adults (mean age＝28.7 years) to older adults (mean age＝69.6 years) and found no significant difference in the location of their VOT boundaries for the ga–ka distinction. So, from the results of both Tadaka and Itoh et al., it seems that perception is not as readily changing in young people as production is. However, this is something that needs more thorough testing.

Exposure to a second language also has the potential to affect phonetic production, and indeed perception, in one's first language. In both French-English bilinguals (Caramazza et al., 1973) and Spanish-English bilinguals (Elman et al., 1977; Williams, 1977) it is apparent that the perceptual VOT boundary values between /b/ and /p/ may be longer for those participants who have more exposure to English. For production of VOT, Johnson & Wilson (2002) found that young bilingual Japanese-English children's languages are each affected by the other, and that children can be differentiating their languages (in terms of VOT values) at a level that is not perceivable by adult listeners but is measurable using acoustic analysis equipment. Therefore, we would expect Japanese speakers who have an advanced level of English to have Japanese VOT values that are closer to native English VOT values for both production and perception.

This paper reports a pilot experiment on the perception of natural Japanese words by Japanese listeners, when such words are modified so that the VOT of word-initial /t/ and /d/ overlaps. We investigate whether or not other phonetic cues are sufficient to maintain perception of a stop-initial word when the VOT cue has been altered.

2 Method

2.1 Participants

A total of 41 participants took part in this perception experiment. All participants were third- and fourth-year undergraduate students at the University of Aizu, Fukushima, Japan. They were computer science majors and were members of an elective phonetics class. All participants had had 6 years of junior high school and senior high school EFL education, as well as two years of university-level EFL classes. Approximately 75% of the participants originated from the Tohoku area of Japan, and 90% were male. There were no reported cases of hearing problems among the participants.

2.2 Stimuli preparation

In choosing stimuli for a VOT perception experiment, there are a number of factors to consider. So many phonetic and linguistic properties can vary within a word, so it is imperative to control carefully the factors that are not under investigation. For example, if the stimuli are more than one syllable, then the VOT of the second stop consonant (if there is one) might affect the perception of the first stop consonant. The frequency of occurrence of a word in the language is another factor that must be

balanced (as much as possible) between choices in a given pair. It is assumed that a participant will choose the most frequently occurring word, especially in cases where the stimulus is neither a clear example of one of the two answer choices.

Japanese stop-initial words also have the property that before high vowels, the alveolar stop consonants (/t/ and /d/) are affricated. Affrication has a very large effect on VOT, so these combinations of alveolar stop-high vowel cannot be used. Another property of Japanese words that must be considered is pitch accent. If a pair of words which is only supposed to differ with respect to the word-initial stop also differs in pitch accent, then those words will be easy to distinguish no matter how similar the word-initial stop is made to sound.

Stimuli to be considered for use in this perception experiment were recorded by a third-year undergraduate female student at the University of Aizu. She is from Koriyama, Fukushima prefecture in the Tohoku area of Japan. She recorded all of the words in Table 1, repeating each word 3 times. From these recorded words, ideal tokens were selected, copied and manipulated (using Praat acoustic-analysis software), and then placed in a forced-choice identification task. Each word was used in its

Table 1. Real-word stimuli for Japanese VOT experiments

	Sino-Japanese (but *komatsu* is Yamato Japanese)					
	p	b	t	d	k	g
a	–	–	TAN 痰 (phlegm)	DAN 段 (step)	KAN 勘 (intuition)	GAN 癌 (cancer)
e	–	–	TENKI 天気 (weather)	DENKI 電気 (electricity)	KEN 県 (prefecture)	GEN 弦 (string)
o	–	–	TON 豚 (pork)	DON 丼 (rice bowl)	KOMATSU 小松 (small pine)	GOMATSU 語末 (word ending)
	Loanword Japanese					
	p	b	t	d	k	g
a	PAN パン (bread)	BAN バン (van)	TAN タン (tongue)	DAN ダン (down – baseball)	KAN カン (can)	GAN ガン (gun)
e	PEN ペン (pen)	BEN ベン (Ben – name)	TENISU テニス (tennis)	DENISU デニス (Dennis)	KEI ケイ (K – letter)	GEI ゲイ (gay)
o	PONDO ポンド (pound)	BONDO ボンド (bond – glue)	TON トン (tonne – [t])	DON ドン (Mafia boss)	KOMU コム (com – *Willcom*)	GOMU ゴム (rubber)

naturally produced state, as well as twice more (each one with modified VOT). For voiceless stops, the first modification to the naturally produced word was to reduce the VOT to 5ms by cutting an amount of aspiration. The second modification was to further reduce the VOT to 0ms by cutting all aspiration. For voiced stops, which always had positive VOT for this speaker, the first modification to the naturally produced word was to reduce the VOT to 0ms. The second modification was to add an amount of pre-voicing, which was copied from a pre-voicing recording made by the same speaker.

Of the pairs of words in Table 1, only two pairs will be reported in this paper: the alveolar stop-initial words with the low vowel, namely tan (phlegm) and dan (step) in Sino-Japanese, and tan (tongue) and dan (down – used in baseball) among loanwords.

2.3 Procedure

The participants took part in a forced-choice identification task, where they listened to a word through headphones and then had to choose between two answers – one with a voiced stop and the other with a voiceless one – which word they had heard. Before beginning the task, participants were first shown images of the words they would hear. The images can be seen in Figure 1. The pilot study reported in this research paper deals only with the top row (i.e., the *tan-dan* distinction).

As mentioned in section 2.2, in addition to every original word, there were two modified words based on the original one. So, in total there were 12 words used: the 4 original words and 8 modifications of these 4 words. These 12 words were presented in a different random order to every participant, and this was accomplished by using Moodle, an open-source course management system. Within Moodle, there is a quiz module that instructors use to build quizzes. A 12-question quiz (with

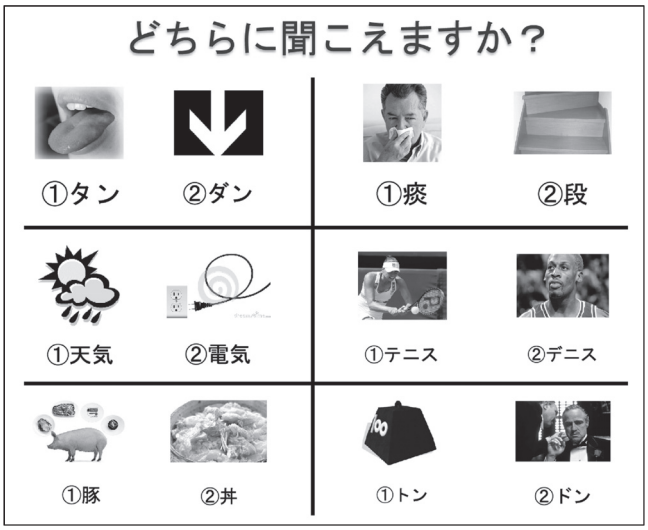

Figure 1. Stimuli images shown to participants before beginning the forced choice identification task.

question order being completely random for every subject) was built asking students (in Japanese), “Which word did you hear?” A sound file that could be clicked on to play appeared next to the text. Participants could click on the sound as many times as they wanted, and they had to choose between the given answers. Most participants finished the quiz in about 2 minutes.

3. Results

In Table 2, the results of the forced-choice identification task can be seen. The results for Sino-Japanese *tan* and *dan* are given first. These are words for which the kanji was given as the answer choice. Below the Sino-Japanese results, we can see the Loanword Japanese results. A number of things stand out in the table. First, we notice that the natural-speech rendition of Sino-Japanese tan had a very short VOT of 10ms in this speaker’s speech. 78% of listeners perceived this natural-speech *tan* as *dan*. For the tokens with decreased VOT, even more listeners perceived *tan* to be *dan*, and this is expected. With Sino-Japanese *dan*, on the other hand, the original natural-speech version had a short positive VOT of 10ms. This seemed to be confusing for listeners, and they split evenly (50%)

Table 2. Number of participants who identified a given stimulus as t-initial or d-initial

Sino-Japanese			
	VOT of stimulus	Participants who identified stimulus as *tan*	Participants who identified stimulus as *dan*
tan「痰」	＋10ms	9 (22%)	32 (78%)
tan「痰」	＋5ms	1 (2%)	40 (98%)
tan「痰」	0ms	6 (15%)	35 (85%)
dan「段」	＋10ms	20 (50%)	20 (50%)
dan「段」	0ms	10 (25%)	30 (75%)
dan「段」	－10ms	23 (57%)	17 (43%)
Loanword Japanese			
	VOT of stimulus	Participants who identified stimulus as *tan*	Participants who identified stimulus as *dan*
tan「タン」	＋20ms	35 (88%)	5 (12%)
tan「タン」	＋5ms	40 (100%)	0 (0%)
tan「タン」	0ms	34 (85%)	6 (15%)
dan「ダン」	＋10ms	1 (2%)	40 (98%)
dan「ダン」	0ms	10 (25%)	30 (75%)
dan「ダン」	－5ms	1 (2%)	39 (98%)

at identifying this as voiced or voiceless. For the *dan* token with 0ms VOT, the number of participants who identified it as *dan* rose to 75%. Finally, for the *dan* token with added pre-voicing, 57% of listeners thought this was voiceless *tan*.

As for the loanword *tan* and *dan* cases, results are much clearer, with an overwhelming majority identifying *tan* as *tan*, and *dan* as *dan*, but no matter what the VOT.

4. Discussion

From the results for the *tan-dan* loanword pair, it is very clear that something other than VOT must be giving the listeners a cue as to whether the word is *tan* or *dan*. Shimizu (1977) showed that the ta-da VOT perceptual boundary is at about 26ms. However, note that *all* of the words used in this research had VOT less than 26ms, and hence all should have been perceived as *dan*, not *tan*, if VOT were the only cue. In creating the modified stimuli, only the VOT was altered, so it is entirely likely that other cues such as f0 at vowel onset, as well as F1 at vowel onset, led the listener to believe that even the 0ms VOT version of *tan* was still *tan*.

One of the strangest findings in this study is the result for Sino-Japanese *dan*, particularly the third token. How is it possible that 23 out of 40 listeners perceived *tan*, when the word had −10ms VOT (i.e., pre-voicing)? Again, other phonetic factors must have contributed to this intriguing result.

5. Conclusions and future work

The pilot research described in this paper is a first step towards a more thorough understanding of the gradual lengthening of VOT by native Japanese speakers. The paper undoubtedly raises more questions than it answers, but it is hoped that future research will uncover more answers to the mysteries of VOT – especially the perception of VOT, and the relative importance of VOT for the perception of the voiced-voiceless stop distinction in Japanese.

In future work, there are a number of areas that need empirical research. For example, it would be interesting to further investigate whether increased exposure to English is affecting the VOT of Japanese people producing and perceiving their native language. We hypothesize that given a choice of two homonyms such as tan (tongue) and tan (phlegm), if a Japanese person hears a longer VOT, s/he will select the loanword (i.e., tongue). We also hypothesize that a correlation exists between a Japanese person's English proficiency and his/her VOT length. In other words, we hypothesize that the better at English a person is, the more English-like his/her Japanese VOT values are. It is hoped that future research will be able to elucidate these matters.

References

Abramson, A. S., & Lisker, L. (1973). Voice-timing perception in Spanish word-initial stops. *Journal of Phonetics*, 1, 1–8.

Caramazza, A., Yeni-Komshian, G. H., Zurif, E. B., & Carbone, E. (1973). The acquisition of a new phonological contrast: The case of stop consonants in French-English bilinguals. *Journal of the Acoustical Society of America*, 54, 421–428.

Elman, J. L., Diehl, R. L., & Buchwald, S. E. (1977). Perceptual switching in bilinguals. *Journal of the Acoustical Society of America*, 62, 971–974.

Itoh, M., Tatsumi, I. F., & Sasanuma, S. (1986). Voice onset time perception in Japanese aphasic patients. *Brain and Language*, 28, 71–85.

Johnson, C. E., & Wilson, I. L. (2002). Phonetic evidence for early language differentiation: Research issues and some preliminary data. *The International Journal of Bilingualism*, 6, 271–289.

Kong, E. J., Beckman, M. E., & Edwards, J. (2012). Voice onset time is necessary but not always sufficient to describe acquisition of voiced stops: The cases of Greek and Japanese. *Journal of Phonetics*, 40, 725–744.

Shimizu, K. (1977). Voicing features in the perception and production of stop consonants by Japanese speakers. *Studia Phonologica*, 11, 25–34.

Shimizu, K. (1996). *A cross-language study of voicing contrasts of stop consonants in Asian languages*. Tokyo: Seibido.

Takada, M. (2004). *日本語の語頭の有声歯茎破裂音/d/における＋VOT化と世代差* [＋VOT tendency in the initial voiced alveolar plosive /d/ in Japanese and the speakers' age]. *Journal of the Phonetic Society of Japan*, 8, 57–66.

Takada, M. (2011). *日本語の語頭閉鎖音の研究―VOTの共時的分布と通時的変化* [Research on the word-initial stops of Japanese: Synchronic distribution and diachronic change in VOT]. Tokyo: Kurosio.

Williams, L. (1977). The perception of stop consonant voicing by Spanish-English bilinguals. *Perception & Psychophysics*, 21, 289–297.