VOT and F0 cues in the perception of synthesized plosives by Japanese listeners *

◎Jiayin Gao[1,2,3], Jihyeon Yun[1,4], Takayuki Arai[1]
([1]Sophia Univ., [2]JSPS, [3]LPP, CNRS-Paris 3, [4]Chungnam National Univ.)

1 Introduction

Previous studies on production data have shown that in modern Tokyo Japanese, VOT cue is not robustly produced [1], while a large f0 (fundamental frequency) difference is realized, being higher after voiceless than voiced plosives [2]. The present study reports the perception of the voicing contrast by Tokyo Japanese (TJP) listeners. We aim to examine their perceptual boundary of VOT (voice onset time), as well as their sensitivity to the f0 cue. In a study with American listeners [3], it is shown that f0 plays a role only when VOT is ambiguous. However, when f0 conflicts with VOT, even when VOT is unambiguous, reaction times are lengthened, suggesting that f0 cue is still processed by listeners. Similar parameters are used in this study to compare the perceptual role of VOT and f0 in English and Japanese.

2 Experiment

2.1 Method

2.1.1 Participants

Twenty native speakers (8 males, 12 females) of TJP aged 20 to 31 (mean 23) participated in an identification test and 15 of them also participated in an AXB discrimination test.

2.1.2 Stimuli

Stimuli were created using a Klatt synthesizer KLSYN93. Two VOT continua of 13 stimuli were synthesized, each ranging from -60 to +60 ms with a step of 10 ms. The two continua differed in f0 onset: one started at 98 Hz and the other at 130 Hz. F0 then changed linearly to a steady state of 114 Hz over the first 50 ms of the vowel. The vowel duration was set to 250 ms.

2.1.3 Procedure

Participants were tested individually in a soundproof room, through a professional quality headphone connected to a laptop computer. The discrimination test preceded the identification test for listeners who participated in both tests.

For the three-VOT-step AXB discrimination test, at trial onset, a fixation cross was displayed at the centre of the screen; 500 ms later, the fixation cross disappeared and three auditory stimuli were presented consecutively, with an interstimulus-interval at 1 sec. Listeners were instructed to respond whether the second stimulus sounded more similar to the first or the third stimulus, by pressing the left or the right <SHIFT> key on the keyboard. Each continuum contained 10 A-B pairs, and each trial had 4 combinations (AAB, ABB, BAA, BBA), yielding a total of 320 trials (10 pairs*4 combinations*2 f0*4 repetitions).

For the identification test, at trial onset, a fixation cross was displayed at the centre of the screen; 500 ms later, one auditory stimulus was presented; at stimulus offset, the fixation cross was replaced with two possible responses written in *katakana* at the left and right side of the screen. Listeners were instructed to choose the syllable they heard by pressing the left or the right <SHIFT> key on the keyboard. Each stimulus was repeated 6 times, yielding a total of 156 trials (13 steps*2 f0*6 repetitions). Response times were measured from the offset of the auditory stimulus.

The response time-out was set to 2 seconds. The test phase was preceded by a training phase consisting of original stimuli with unambiguous prevoicing or aspiration, during which listeners were given feedback for their correctness and response time. During the test phase, the stimuli were presented in a different randomized order for each listener.

---

* 音声合成による破裂音を日本語母語話者が知覚する際の VOT と F0 の手がかり
高佳音[1,2,3], ユンジヒョン[1,4] 荒井隆行[1]（[1]上智大学, [2]日本学術振興会, [3]LPP, CNRS-Paris 3, [4]Chungnam National Univ）.

2.2 Results

2.2.1 AXB discrimination

Discrimination data were analyzed using d-prime. Sharp discrimination peaks could be observed, suggesting highly categorical perception of VOT. There was no clear shift of the peak between the two f0 onset versions.

2.2.2 Identification

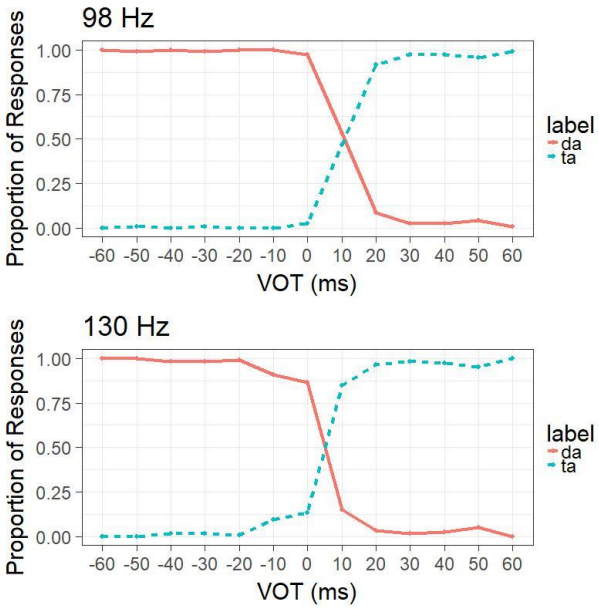

Fig. 1 Identification curves for /da/ and /ta/ responses as a function of VOT, for 98 Hz (upper panel) and 130 Hz (lower panel) stimuli.

Fig. 1 shows pooled identification curves for /da/ and /ta/ responses as a function of VOT, separately for the two f0 onset versions. We may observe S-shape curves, suggesting highly categorical perception of the VOT continuum. ROC curves were plotted and AUC was used to estimate the consistency of listeners' responses. It was 0.984 for 98 Hz stimuli and 0.98 for 130 Hz stimuli, suggesting nearly perfect categorical perception. Fig. 1 also shows a boundary shift toward the leftmost VOT endpoint by half a step with 130 Hz compared to 98 Hz stimuli. For pooled data, the optimal cutoff point between /da/ and /ta/ responses was estimated by the YOUDEN index at 10 ms for both f0 onset versions. Individual analyses estimated that 12 out of 20 listeners showed a boundary shift of one step between the two f0 onset versions while the other 8 listeners did not.

2.2.3 Response time (RT)

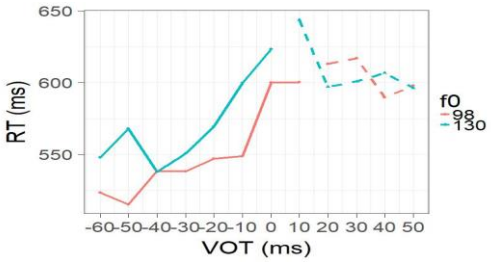

Fig. 2 RT as a function of VOT.

Fig. 2 shows pooled RT data for dominant responses (i.e., >50%) at each VOT step. When f0 conflicts with VOT (e.g., 130 Hz with negative VOT), a lengthening of RT can be globally observed. However, post-hoc pairwise comparisons showed a significant effect of f0 only for VOT at -50 and -10 ms (*t*s=-3.0, *p*s<.005). Contrary to [3], f0 had little effect on RT.

3 Conclusions

Our results showed that, despite a shift in production from VOT to f0 cue of plosives, VOT remains the primary cue of voicing perception of plosives by Japanese listeners and the role of F0 is minor and inconsistent among the listeners, at least in synthesized nonsense syllables. A follow-up study on modified natural speech has been conducted and preliminary results showed that listeners were more sensitive to f0 in natural than synthesized speech.

Acknowledgements

We are grateful to Yu SHI for sharing the Python code for the experiment platform. This project is supported by JSPS (*kakenhi* n. 17F17006).

References

[1] M. Takada, E. J. Kong, K. Yoneyama, and M. E. Beckman. "Loss of prevoicing in modern Japanese /g, d, b/," *Proc. ICPhS 2015*, paper n. 873, 1-5 (2015).

[2] J. Gao, and T. Arai. F0 perturbation in a "pitch-accent" language. *Proc. 6th Int'l Symposium on TAL*, paper n. 13, 1-5 (2018).

[3] D. H. Whalen, A. S. Abramson, L. Lisker, and M. Mody. F0 gives voicing information even with unambiguous voice onset times. *JASA*, **93**(4), 2152-2159 (1993).