PERCEPTION BOUNDARY BETWEEN /s/ AND /ts/ IN JAPANESE AT VARIOUS SPEAKING RATES

Shigeaki Amano & Kimiko Yamakawa

Aichi Shukutoku University
psy@asu.aasa.ac.jp, jin@asu.aasa.ac.jp

ABSTRACT

To examine effects of speaking rate on a perception boundary between voiceless fricative /s/ and affricate /ts/ in Japanese, a stimulus continuum of /s/-/ts/ at various speaking rates was constructed by changing rise and steady+decay durations of these phonemes. The stimuli in the continuum were presented to 33 Japanese native speakers, and their response ratio to /ts/ was measured. Logistic regression analysis of the response ratio revealed that the perception boundaries systematically differed as a function of speaking rate. However, when the rise and steady+decay durations were normalized by a logarithm of the averaged mora duration, the boundaries at different speaking rates nearly coincided. These results suggest that the speech perception system discriminates /s/ and /ts/ at various speaking rates with reference to a single boundary calculated with the normalized variables.

Keywords: perception boundary, fricative, affricate, Japanese, speaking rate

1. INTRODUCTION

Japanese voiceless fricative /s/ and affricate /ts/ have similar acoustic features. They both consist of a frication, but the envelope of the frication's intensity tends to have long rise and steady components in /s/, whereas it tends to have short rise and steady components in /ts/.

With these characteristics, Yamakawa et al. [9] analyzed the variables that can discriminate /s/ and /ts/. They divided the intensity envelopes of /s/ and /ts/ into rise, steady, and decay components, and they approximated each component with a line of positive, zero, or negative slope (Fig. 1). They found that /s/ and /ts/ are well discriminated by two variables: the rise duration and the sum of the steady and decay durations (hereafter referred to as "steady+decay").

Based on the findings by Yamakawa et al. [9], Amano and Yamakawa [3] constructed a stimulus continuum between /s/ and /ts/ by changing the rise and steady+decay durations. They performed a perception experiment with the stimulus continuum and found that a perception boundary between /s/ and /ts/ is represented as a linear function of the two durational variables.

However, because Amano and Yamakawa [3] used the speech stimulus at only a normal speaking rate, speaking rate effects on the perception boundary have not been clarified. Because many previous studies have reported that speech perception depends on speaking rate, and the perception boundary of speech segments changes according to the speaking rate [e.g., 1, 5, 7, 8], the perception boundary found in Amano and Yamakawa's study [3] probably changes according to speaking rates.

On this background, the current study examined the effects of speaking rate on the perception boundary between /s/ and /ts/. The perception boundary is expected to systematically change as a function of speaking rate. The current study also examined the possibility that the discrepant perception boundaries at different speaking rates can coincide when the durational variables are normalized by a rate-dependent duration such as a logarithm of the averaged mora duration. Such coincidence would suggest that the speech perception system uses these normalized variables for discriminating /s/ and /ts/ at various speaking rates.

2. EXPERIMENT

2.1. Stimuli

Original materials were the Japanese words /su/ ("vinegar"), /suru/ ("do"), /suneru/ ("sulk"), and /sumagoto/ ("single-string harp"). These words form minimal pairs with /tsu/ ("harbor"), /tsuru/ ("hook"),

Figure 1: Schematic diagram of the intensity envelope for the stimulus continuum.

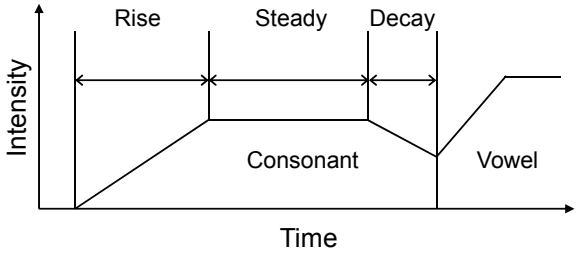


Table 1: Durations of rise, steady, and decay parts of the stimulus continuum (ms)

Speaking rate	Rise			Steady			Decay
	Min.	Max.	Step	Min.	Max.	Step	
Fast	12	30	3	10	70	10	10
Normal	15	39	4	10	100	15	14
Slow	25	55	5	10	142	22	18

/tsuneru/ ("pinch"), and /tsumagoto/ ("multi-string harp"), which have the same phoneme sequence except that the initial phoneme is replaced with /ts/.

The two words in each minimal pair have the same accent pattern and differ very little in auditory word familiarity [2]. Because of the small familiarity difference, no lexical bias [4] was expected for a category boundary between /s/ and /ts/ in the minimal pair.

The words were embedded in a carrier sentence, /__ mo tango desu/ ("__ is also a word"), and they were pronounced at a fast, normal, and slow speaking rate by one Japanese female speaker in her twenties. The pronounced words were digitally recorded with 16-bit quantization at a 48-kHz sampling rate, and used as original speech materials.

Stimulus continua between /s/ and /ts/ were produced by modifying the rise and steady durations of /s/ in the original speech materials [3]. A schematic diagram of the intensity envelope for the stimulus continua is shown in Figure 1.

Table 1 shows the range and step for changing the duration of the rise, steady, and decay parts to make a stimulus continuum at each speaking rate. The intensity envelope of the rise was increased linearly as a function of time. The duration of the steady part was changed by cutting the tail of the original steady part if the target duration was shorter than the original value or by extending the original steady part with an overlap and add method if the target duration was longer than the original value. Decay duration was constant, but it differed among the speaking rates.

There were 49 stimuli (seven rise durations x seven steady durations) in each stimulus continuum for each of the four word pairs at three speaking rates, resulting in a total of 588 stimuli.

2.2. Participants

Thirty-three (nine male and 24 female) monolingual native Japanese speakers with normal hearing ability were paid for their participation in the experiment. Their average age was 26.5 years (Min. = 20; Max. = 35; SD = 0.86).

2.3. Procedure

The stimuli were diotically presented to the participants through headphones (MDR-Z900HD, SONY) at a comfortable sound level in a quiet room. The stimulus order was randomized for each participant.

When each stimulus was presented, two response buttons were displayed on a computer screen. One button showed a word with the initial phoneme /s/ and the other showed a word with the initial phoneme /ts/. Both words were written in Japanese hiragana orthography.

The participants' task was to make a two-alternative forced choice, deciding whether the word they heard began with /s/ or /ts/. After responding by clicking one of the two buttons, the participants were asked to click the "confirm" button.

At the beginning of the experiment, the participants performed 24 practice trials. After the practice trials, each participant proceeded with the experiment consisting of 588 trials broken into five blocks. The experiment was self-paced, but the computer prompted the participants to take two-minute breaks between blocks. It took the participants about 120 minutes to complete the experiment.

2.4. Results

2.4.1. Analysis without normalization

The response ratio for /ts/ was calculated by dividing the number of /ts/ responses by the total number of responses for each stimulus. For each speaking rate, a logistic function was fitted to the /ts/ response ratio as a dependent variable, with rise duration and steady+decay duration as independent variables. The fitted logistic functions for the fast, normal, and slow speaking rates were respectively Eq. 1, 2, and 3,

(1) $z = 1/(1 + \exp(-0.122x - 0.0840y + 6.99))$
(2) $z = 1/(1 + \exp(-0.104x - 0.0629y + 6.20))$
(3) $z = 1/(1 + \exp(-0.0569x - 0.0336y + 4.13))$

where x is the rise duration, y is the steady+decay duration, and z is the /ts/ response ratio.

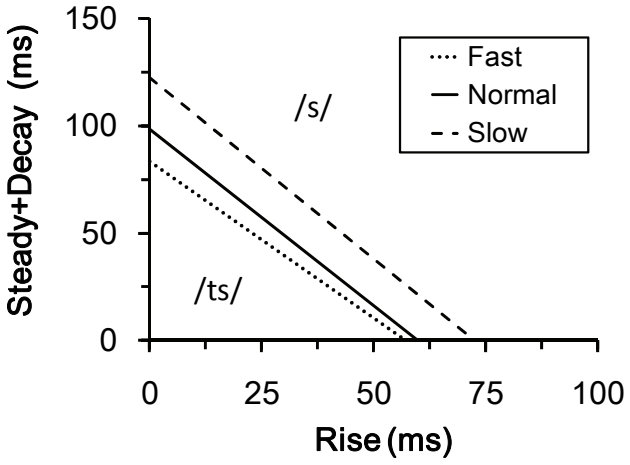
Fitting of the logistic functions was successful. For the fast speaking rate, the goodness of fit was significant ($\chi^2(2) = 2950.4$, $p < .0001$), and a Wald test indicated that each coefficient of the logistic function was significant (for the intercept, $\chi^2(1) = 1290.1$, $p < .0001$; for rise, $\chi^2(1) = 442.1$, $p < .0001$; for steady+decay, $\chi^2(1) = 1638.2$, $p < .0001$). For the normal speaking rate, the goodness of fit was significant ($\chi^2(2) = 3069.4$, $p < .0001$), and a Wald test indicated that each coefficient of the logistic function was significant (for the intercept, $\chi^2(1) = 1095.5$, $p < .0001$; for rise, $\chi^2(1) = 479.6$, $p < .0001$; for steady+decay, $\chi^2(1) = 1541.8$, $p < .0001$). For the slow speaking rate, the goodness of fit was significant ($\chi^2(2) = 2107.7$, $p < .0001$), and a Wald test indicated that each coefficient of the logistic function was significant (for the intercept, $\chi^2(1) = 612.4$, $p < .0001$; for rise, $\chi^2(1) = 277.0$, $p < .0001$; for steady+decay, $\chi^2(1) = 1270.0$, $p < .0001$).

The perception boundary between /s/ and /ts/ was obtained as the linear function of the rise and steady+decay durations that gave a 50% /ts/ response ratio on the fitted logistic function. The linear functions for the perception boundary at the fast, normal, and slow speaking rates were respectively Eq. 4, 5, and 6.

(4) $y = -1.45x + 83.3$
(5) $y = -1.65x + 98.5$
(6) $y = -1.69x + 123$

These perception boundaries are shown in Figure 2. They are located at different positions. The perception boundary at the fast speaking rate is near to the point of origin, the perception boundary at the slow speaking rate is far from the point of origin, and the perception boundary at the normal speaking rate is located between these boundaries.

Figure 2: Perception boundary between /s/ and /ts/ at fast, normal, and slow speaking rates. Independent variables are rise duration and steady+decay duration.



These discrepancies might disappear when the durational variables are normalized by a rate-dependent duration such as a logarithm of the averaged mora duration. This possibility is examined in the next section.

2.4.2. Analysis with normalization

The response ratio for /ts/ was calculated with the same procedure as in Section 2.4.1. The averaged mora duration (ave_mora) in a word portion was calculated for each stimulus. With this averaged mora duration, normalized rise duration (nor_x) and normalized steady+decay duration (nor_y) were respectively obtained with Eq. 7 and 8,

(7) $nor\_x = x / \log(ave\_mora)$
(8) $nor\_y = y / \log(ave\_mora)$

where x is the rise duration and y is the steady+decay duration.

For each speaking rate, a logistic function was fitted to the /ts/ response ratio as a dependent variable, with the normalized rise duration and the normalized steady+decay duration as independent variables. The fitted logistic functions for the fast, normal, and slow speaking rates were respectively Eq. 9, 10, and 11,

(9) $z=1/(1+\exp(-0.260nor\_x-0.185nor\_y+7.71))$
(10) $z=1/(1+\exp(-0.228nor\_x-0.145nor\_y+6.53))$
(11) $z=1/(1+\exp(-0.140nor\_x-0.0864nor\_y+4.17))$

where z is the /ts/ response ratio.

Fitting of the logistic functions was successful. For the fast speaking rate, the goodness of fit was significant ($\chi^2(2) = 2979.0$, $p < .0001$), and a Wald test indicated that each coefficient of the logistic function was significant (for the intercept, $\chi^2(1) = 1299.9$, $p < .0001$; for normalized rise, $\chi^2(1) = 478.1$, $p < .0001$; for normalized steady+decay, $\chi^2(1) = 1671.1$, $p < .0001$). For the normal speaking rate, the goodness of fit was significant ($\chi^2(2) = 3028.8$, $p < .0001$), and a Wald test indicated that each coefficient of the logistic function was significant (for the intercept, $\chi^2(1) = 1094.2$, $p < .0001$; for normalized rise, $\chi^2(1) = 487.1$, $p < .0001$; for normalized steady+decay, $\chi^2(1) = 1572.7$, $p < .0001$). For the slow speaking rate, the goodness of fit was significant ($\chi^2(2) = 2066.8$, $p < .0001$), and a Wald test indicated that each coefficient of the logistic function was significant (for the intercept, $\chi^2(1) =$

598.4, $p < .0001$; for normalized rise, χ^2 (1) = 267.0, $p < .0001$; for normalized steady+decay, $\chi^2(1)$ = 1289.3, $p < .0001$).

The perception boundary between /s/ and /ts/ was obtained as the linear function of the normalized rise and steady+decay durations that gave a 50% /ts/ response ratio on the fitted logistic function. The linear functions for the perception boundary at the fast, normal, and slow speaking rates were respectively Eq. 12, 13, and 14.

(12) nor_y = -1.40nor_x + 41.6
(13) nor_y = -1.57nor_x + 45.0
(14) nor_y = -1.62nor_x + 48.2

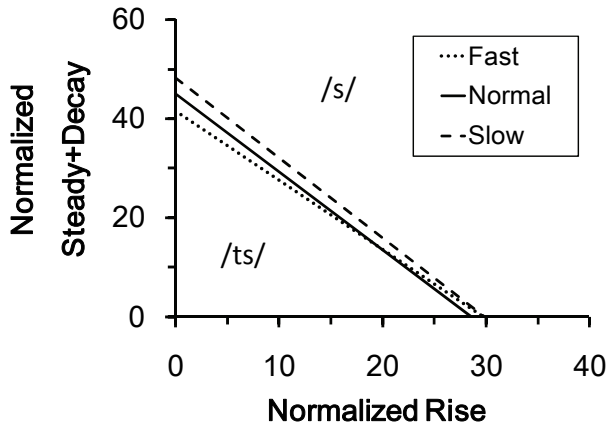
These perception boundaries are shown in Figure 3. They are located at almost the same position.

3. DISCUSSION

The current study revealed that the perception boundaries between /s/ and /ts/ systematically change as a function of speaking rate (Fig. 2). At a fast speaking rate, durational variables tend to have a shorter value than at a normal speaking rate, whereas they tend to have a longer value at a slow speaking rate than at a normal speaking rate. This tendency is consistent with the results of previous studies that indicate the speaking-rate dependency of the perception boundaries of speech segments such as English plosives [5, 8], Icelandic syllables [7], and Japanese geminate stops [1].

The current study also revealed that the speaking-rate dependency of the perception boundary disappears and the boundaries almost coincide (Fig. 3), when the rise and steady+decay durations are normalized by a logarithm of the averaged mora duration. This new finding strongly suggests that the speech perception system uses these normalized variables and it discriminates /s/ and /ts/ at various speaking rates with reference to a single boundary calculated with the normalized variables.

Figure 3: Perception boundary between /s/ and /ts/ at fast, normal, and slow speaking rates. Independent variables are rise duration and steady+decay duration, which are normalized by a logarithm of the averaged mora duration.



The logarithmic conversion is often found in perception such as loudness of sound, brightness of light, and heaviness of object. That is, the perception system obeys Fechner's law, which states that the perception of the stimulus is proportional to a logarithm of the stimulus's physical value. The results of the current study suggest that perception of speech segment duration also obeys Fechner's law and therefore, the logarithm of the duration can be a reliable speaking rate normalizer.

In the current results, the logarithm of the averaged mora duration is a good normalizer of the rise and steady+decay durations. This means that the rise and steady+decay durations invariantly relate to a logarithm of the averaged mora duration at any speaking rate. This notation of the current results provides support to relational invariance theory [6]. In relational invariance theory, acoustic invariance is relational in the sense that one part of speech relates invariantly to another in order to indicate a feature of speech, and speech perception is sensitive to this invariant relationship. The results of this study accord with these notions of the theory.

Amano and Yamakawa [3] have shown that the perception and production boundaries between /s/ and /ts/ are almost identical. However, they used stimuli at only a normal speaking rate. A future study should examine whether the perception and production boundaries coincide at fast or slow speaking rates. It should also examine whether production boundaries at various speaking rates can be located at nearly the same position when rise and steady+decay durations are normalized with a logarithm of the averaged mora duration. In addition, it should examine whether these production boundaries are located at nearly the same position as the perception boundaries obtained in this study. These investigations will provide a new scientific view for relational invariance theory and the relationships between speech perception and production.

4. ACKNOWLEDGEMENTS

This study was supported by JSPS KAKENHI Grant Numbers 25284080 and 26370464 and by a cooperative research grant (2013-2014) and a specified research grant (2015-2016) of Aichi Shukutoku University.

5. REFERENCES

[1] Amano, S., Hirata, Y. 2010. Perception and production boundaries between single and geminate stops in Japanese. J. Acoust. Soc. Am. 128, 2049-2058.

[2] Amano, S., Kondo, T. 1999. Nihongo-no Goi-Tokusei (Lexical properties of Japanese). Tokyo: Sanseido (In Japanese).

[3] Amano, S., Yamakawa, K. 2011. Perception and production boundaries between fricative [s] and affricate [ts] in Japanese. Proc. the 17th ICPhS, 228-231.

[4] Ganong III, W. F. 1980. Phonetic categorization in auditory word perception. J. Exp. Psychol. Human. 6, 110-125.

[5] Miller, J. L., Volaitis, L. E. 1989. Effect of speaking rate on the perceptual structure of a phonetic category. Percept. Psychophys. 46, 505-512.

[6] Pickett, E. R., Blumstein, S. E., Burton, M. W. 1999. Effects of speaking rate on the singleton/geminate consonant contrast in Italian. Phonetica 56, 135-157.

[7] Pind, J. 1995. Speaking rate, voice-onset time, and quantity: The search for higher-order invariants for two Icelandic speech cues. Percept. Psychophys. 57, 291-304.

[8] Volaitis, L. E., Miller, J. L. 1992. Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. J. Acoust. Soc. Am. 92, 723-735.

[9] Yamakawa, K., Amano, S., Itahashi, S. 2012. Variables to discriminate affricate [ts] and fricative [s] at word initial in spoken Japanese words, Acoust. Sci. & Tech. 33, 154-159.