PERCEPTION AND PRODUCTION BOUNDARIES BETWEEN FRICATIVE [s] AND AFFRICATE [ts] IN JAPANESE

Shigeaki Amano & Kimiko Yamakawa

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

ABSTRACT

The perception and production boundaries between a fricative [s] and an affricate [ts] in Japanese were investigated to clarify the relationship between them. Regression and discriminant analyses revealed that the perception and production boundaries between [s] and [ts] were represented by very similar linear functions defined by the variables of rise duration and steady+decay duration. In addition, the perception boundary discriminated between [s] and [ts] in the production data with a small error rate (7.11%). These results indicate a correspondence between the perception and production boundaries separating [s] and [ts], and suggest a close connection between speech perception and production.

Keywords: speech perception, speech production, phoneme boundary, fricative, affricate

1. INTRODUCTION

Many researchers take it for granted that the perception and production of speech are closely connected. However, the correspondence between the perception and production boundaries has not been very clearly observed. For example, Miller, Green, and Reeves [3] reported that there is a discrepancy between the perception and production boundaries of /b/ and /p/ in terms of voice onset time. Moreover, the production boundary between /b/ and /p/ that they found [3] did not match the perception boundary reported in other studies (Miller and Volaitis [4]; Volaitis and Miller [7]; cf., Nagao & de Jong [5]).

However, a few studies have shown a correspondence between perception and production boundaries. For instance, Nagao and de Jong [5] reported that the perception and production boundaries between /b/ and /p/ in naturally spoken American English are almost the same in terms of voice onset time and syllable duration. In addition, Pind [6] found that the perception and production boundaries between VC: (vowel + long consonant) and V:C (long vowel + consonant) type words in Icelandic were very similar in terms of vowel duration and consonant duration. Finally, Amano and Hirata [1] found that the perception and production boundaries between single and geminate stops in Japanese are represented as a linear function defined by the variables of word duration and closure duration, and that these two boundaries correspond with each other.

To obtain additional evidence for a close connection between speech perception and production, this study selected a Japanese fricative [s] and an affricate [ts] and investigated the correspondence between the perception and production boundaries.

Yamakawa, Amano, and Itahashi [8, 9] modeled the intensity envelopes of [s] and [ts] as three polygonal lines. That is, they divided the intensity envelopes into rise, steady, and decay components and approximated each component with linear lines with positive, zero, and negative slopes, respectively. They found that the production boundary between [s] and [ts] is represented by a linear function defined by two variables: the rise duration and the sum of the steady and decay durations (hereafter referred to as “steady+decay”).

Based on their findings, these two variables (rise duration and steady+decay duration) were used in the current study to identify the perception and production boundaries between [s] and [ts], and to investigate the relationship between these boundaries.

2. PERCEPTION BOUNDARY

2.1. Stimuli

Original speech materials were Japanese words /su/, /suru/, /suneru/, and /sumagoto/ which were pronounced by one Japanese female speaker and registered in a Japanese word familiarity database [2]. These words have minimal pair words /tsu/, /tsuru/, /tsuneru/, and /tsumagoto/, which have the same phoneme sequence except that the initial phoneme is replaced with [ts]. In addition, the two words in each minimal pair have the same accent pattern and differ very little in auditory word familiarity [2].

The intensity envelope of [s] tends to have long rise and steady components, whereas the intensity envelope of [ts] tends to have short rise and steady components. Based on this tendency, stimulus continua between [s] and [ts] were produced by modifying the rise and steady durations of [s] in the original speech materials (Fig. 1). The rise duration of [s] was changed from 20 to 55 ms in 5-ms steps, and the intensity envelope of the rise was increased linearly as a function of rise time.

The duration of the steady component was changed from 0 to 100 ms in 10-ms steps. This was accomplished by cutting the tail of the original steady component if the target duration was shorter than the original value or by repeating the original steady component and then cutting its tail if the target duration was longer than the original value. The decay duration was fixed at 25 ms.

There were 88 stimuli (eight rise durations and 11 steady durations) in each stimulus continuum for each of the 4 word pairs, resulting in a total of 352 stimuli.

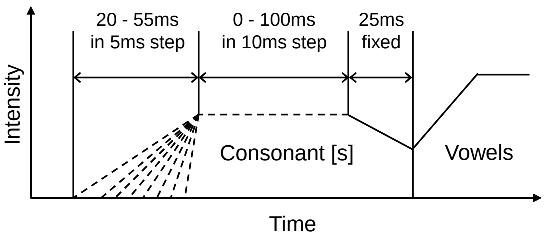
2.2. Participants

Forty (20 male and 20 female) monolingual native Japanese speakers with normal hearing ability were paid for their participation in the experiment. Their average age was 24.5 years (range = 20–35; SD = 3.23).

2.3. Procedure

The stimuli were diotically presented to the participants through headphones at a comfortable sound level in a quiet room. The 352 stimuli were presented five times in a randomized order for 1,760 trials for each participant.

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



When each auditory 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.

On each trial, the participants' task was to make a two-alternative forced choice, deciding whether the word they heard began with [s] or [ts]. After making a response by clicking one of the two buttons, the participants were asked to click the “next” button to hear the next stimulus.

At the beginning of the experiment, participants were given two sets of eight practice trials. After the practice sets, each participant proceeded with the experiment consisting of 1,760 trials that were broken into 10 blocks of 176 trials. The experiment was self-paced, but the computer prompted participants to take three-minute breaks between blocks. It took the participants 100–130 minutes to complete the experiment.

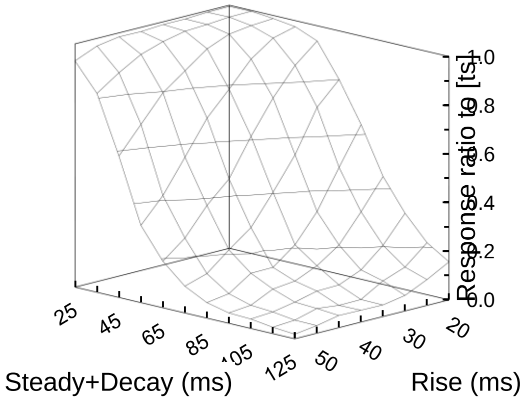
2.4. Results

The response ratio for [ts] was calculated by dividing the number of [ts] responses by the total number of responses. Fig. 2 shows the [ts] response ratio as a function of rise duration and steady+decay duration. A logistic function was fitted to this [ts] response ratio as a dependent variable, with rise duration and steady+decay duration as independent variables. The fitted logistic function was

$$z = 1/(1 + \exp(-0.0970x - 0.0756y + 9.0124)) \quad (1)$$

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

Figure 2: [ts] response ratio as a function of rise duration and steady+decay duration.



The logistic function fitting was successful: the goodness of fit of the logistic function was high and significant (chi-square(2) = 46276.74, p < .0001), and a Wald test indicated that each coefficient of the logistic function was significant (for the intercept, chi-square(1) = 15478.46, p < .0001; for rise, chi-square(1) = 7148.48, p < .0001; for steady+decay, chi-square(1) = 19103.29, p < .0001).

The perception boundary between [s] and [ts] was defined 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 function for the perception boundary was

$$y = -1.284x + 119.2 \qquad (2)$$

where x is the rise duration and y is the steady+decay duration. This linear function provides a good representation of the perception boundary because it is derived from a well-fitted logistic function for perception data.

3. PRODUCTION BOUNDARY

3.1. Spoken words

Japanese spoken words which were pronounced by one Japanese female speaker were selected from a Japanese word familiarity database [2]. The selection conditions for the spoken words were

1. the word length was one to four moras;
2. the initial phoneme was [s] or [ts] followed by the vowel /u/, which cannot be devoiced;
3. the vowel in the second mora was not a diphthong; and
4. the second mora was not a special mora, such as a nasal mora, an obstruent mora, or a lengthened vowel.

Table 1: Number of spoken words used to determine production boundary between [s] and [ts].

Word length (mora)	Initial phoneme		Total
	[s]	[ts]	
1	2	1	3
2	14	25	39
3	90	89	179
4	91	91	182
Total	197	206	403

Because the database does not contain many one- and two-mora words that satisfy these conditions, all of the one- and two-mora words were selected. On the other hand, because the database contains many three- and four-mora words that satisfy the conditions, about 90 three- and four-mora words were randomly selected for both the [s] and [ts] conditions. The number of selected spoken words is shown in Table 1.

3.2. Procedure

Intensity of each spoken word was calculated with a 6-ms window size and a 1-ms window shift. One author measured the rise, steady, and decay durations of [s] and [ts] in milliseconds while viewing the waveform and the intensity pattern. Another author checked and corrected the measured durations using the same procedure.

3.3. Results

Fig. 3 is a scattergram of [s] and [ts] with variables of rise duration and steady+decay duration. To obtain the production boundary between [s] and [ts], discriminant analysis was performed with these variables. The resulting discriminant function was

$$y = -1.227x + 108.8 \qquad (3)$$

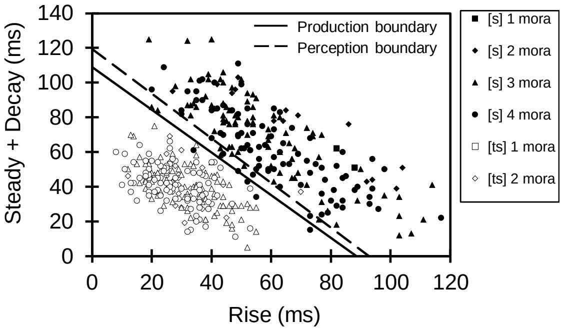
where x is the rise duration and y is the steady+decay duration. This discriminant function (Eq. 3) representing the production boundary is plotted by the solid line in Fig. 3.

The discriminant error by the discriminant function was 1.26%. This very small discriminant error indicates that the variables of rise duration and steady+decay duration were sufficient for discriminating between the production of [s] and [ts]. In other words, the production boundary between [s] and [ts] is represented as a linear function defined by the variables of rise duration and steady+decay duration.

4. BOUNDARY CORRESPONDENCE

To illustrate the relationship between speech perception and production, the perception boundary (Eq. 2) is plotted as a dashed line in Fig. 3. The perception boundary is very similar to the production boundary (solid line). In addition, when the perception boundary was treated as a discriminant function for the production data of [s] and [ts], the discriminant error was only 7.11%. This error is small enough to conclude that the perception boundary discriminates between the production of [s] and [ts] fairly well. This means that the perception boundary can be regarded as the production boundary. These results strongly suggest a correspondence between the perception and production boundaries for [s] and [ts].

Figure 3: Scattergram of production data for [s] and [ts] in 1- to 4-mora-long words. The horizontal axis (x) shows rise duration and the vertical axis (y) shows steady+decay duration. The solid line (y = 1.227x + 108.8) is the production boundary and the dashed line (y = 1.284x + 119.2) is the perception boundary between [s] and [ts].



5. DISCUSSION

The results of this study show that both the perception and production boundaries between a fricative [s] and an affricate [ts] in Japanese are represented by a linear function defined by the rise duration and the steady+decay duration. Moreover, the results strongly suggest that the perception and production boundaries correspond with each other. In other words, the speech perception and production systems use the same linear function with the same variables to categorize the phonemes [s] and [ts].

A correspondence between speech perception and production has been found in the boundary between voiced and voiceless plosives in English [5], between VC: and V:C type words in Icelandic [6], between single and geminate stops in Japanese [1], and between [s] and [ts] in this study. Although these findings differ in terms of phonemes and languages, they all reveal the correspondence between speech perception and production. This raises the possibility that correspondences between speech perception and production also occur for other phonemes and languages. In other words, the correspondence between speech perception and production may be general and universal. However, of course, this notion must be confirmed by further studies of the perception and production boundaries of other phonemes and languages.

The results of this study may pose certain problems in terms of generalization. Unlike previous studies [1, 5, 6], speaking rate was not controlled in this study. That is, this study used a single speaking rate to identify the perception and production boundaries. It is probable that the perception and production boundaries between [s] and [ts] depend on speaking rate. However, the formulation used in this study cannot deal with such a dependency. Therefore, it is necessary to conduct a future study using a new formulation for the boundaries that is independent of the speaking rate. One possibility is to use adjusted rise and steady+decay durations that are divided by the averaged mora duration.

6. ACKNOWLEDGEMENTS

This work was supported by Grant-in-Aid for Scientific Research (C) (21530782).

7. REFERENCES

[1] Amano, S., Hirata, Y. 2010. Perception and production boundaries between single and geminate stops in Japanese. *Journal of the Acoustical Society of America* 128, 2049-2058.

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

[3] Miller, J.L., Green, K.P., Reeves, A. 1986. Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. *Phonetica* 43, 106-115.

[4] Miller, J.L., Volaitis, L.E. 1989. Effect of speaking rate on the perceptual structure of a phonetic category. *Perception & Psychophysics* 46, 505-512.

[5] Nagao, K., de Jong, K. 2007. Perceptual rate normalization in naturally produced rate-varied speech. *Journal of the Acoustical Society of America* 121, 2882-2898.

[6] Pind, J. 1995. Speaking rate, voice-onset time, and quantity: The search for higher-order invariants for two Icelandic speech cues. *Perception & Psychophysics* 57, 291-304.

[7] Volaitis, L.E., Miller, J.L. 1992. Phonetic prototypes: Influence of place of articulation and speaking rate on the internal structure of voicing categories. *Journal of the Acoustical Society of America* 92, 723-735.

[8] Yamakawa, K., Amano, S., Itahashi, S. 2008. Production boundary between fricative [s] and affricate [ts] in Japanese. *Proceedings of Fall Meeting of The Acoustical Society of Japan*, 281-282 (In Japanese).

[9] Yamakawa, K., Amano, S., Itahashi, S. 2009. Variables of production boundary between fricative [s] and affricate [ts] in Japanese. *Proceedings of Spring Meeting of the Acoustical Society of Japan*, 321-322 (In Japanese).