ACOUSTIC FEATURES TO DISCRIMINATE BETWEEN AFFRICATES AND A FRICATIVE IN JAPANESE

Kimiko Yamakawa & Shigeaki Amano

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

ABSTRACT

To clarify the acoustic features that distinguish the alveolo-palatal affricates [tɕ] from the alveolar affricate [ts] or the alveolar fricative [s], Japanese words with these consonants, pronounced by single and multiple native speakers of Japanese, were analyzed with a one-third octave bandpass filter. In the frequency range of 2500–4000 Hz, [tɕ] had much higher intensity than [ts] or [s]. In addition, when the intensity in this frequency range was used as an independent variable, a discriminant analysis showed that [tɕ] was discriminated from [ts] or [s] with greater than 80% accuracy, but [ts] and [s] could not be discriminated. It was concluded that intensity in the frequency range of 2500–4000 Hz is a good acoustic feature to discriminate [tɕ] from [ts] or [s] but not to discriminate between [ts] and [s].

Keywords: acoustic feature, alveolar affricate, alveolo-palatal affricate, alveolar fricative, Japanese

1. INTRODUCTION

Non-native speakers of Japanese, such as Koreans and Thais [3, 6], frequently confuse the voiceless alveolo-palatal affricate [tɕ], voiceless alveolar affricate [ts], and voiceless alveolar fricative [s] in Japanese. The confusion between these consonants is probably caused by their similar acoustic features. For example, the affricates [tɕ] and [ts] and the fricative [s] consist of a noise with some duration. The affricates [tɕ] and [ts] have a burst at the beginning. It is worthwhile to clarify the acoustic features discriminating these consonants and then utilize them to refine a speech education method for non-native Japanese learners.

With this perspective, Yamakawa, Amano, and Itahashi [5] analyzed the acoustic features of [ts] and [s]. They divided the intensity envelopes of [ts] and [s] into rise, steady, and decay components, and they approximated each component with a line of positive, zero, or negative slope. They found that [ts] and [s] are well discriminated by two variables: the rise duration and the sum of the steady and decay durations (hereafter referred to as "steady+decay"). Moreover, they found that the production boundary between [ts] and [s] is represented by a linear function of these two variables.

Yamakawa and Amano [4] also reported that variables of the rise duration and steady+decay duration are effective in discriminating [tɕ] from [s]. However, they found that these two variables are not effective to discriminate [tɕ] from [ts].

What is a good acoustic feature by which to discriminate [tɕ] from [ts]? A previous study [2] showed that in English, the spectral peak is lower in [ʃ] (1600-2500 Hz) than in [s] (about 3500 Hz). This means that [ʃ] can be discriminated from [s] in the spectral domain. [ʃ] and [s] have the same manner but different places of articulation, similar to the case with [tɕ] and [ts]. Therefore, it seems probable that [tɕ] and [ts] can also be discriminated between in the spectral domain.

The present study investigated this possibility of spectral separation between [tɕ] and [ts] as well as between [tɕ] and [s]. That is, we aim to identify acoustic features by which to discriminate between [tɕ], [ts], and [s] in Japanese.

2. MATERIALS

2.1. Single-speaker word materials

2.1.1. Speaker

The speaker was a female Japanese native, 29 years of age. She was well experienced in pronunciation for recordings.

2.1.2. Word materials

Words were selected from the Japanese word familiarity database [1] in which about 70,000 words were pronounced at a normal speaking rate

by the abovementioned speaker. The spoken words in the database are stored as digital audio files with 16-bit quantization and 16-kHz sampling frequency. Low frequency noise in the digital audio files was removed by a high-pass finite-impulse-response filter with a 80-Hz cut-off frequency.

The word selection conditions were as follows:

1) the word length was 3 or 4 moras;
2) the initial phoneme was [tɕ], [ts], or [s];
3) the vowel /u/, which is not devoiced, followed the initial phoneme.

Because the database contains many 3- and 4-mora words that satisfy these conditions, the words were randomly selected. The number of words selected was 127 for [tɕ], 180 for [ts], and 181 for [s]. There were 488 word materials in total.

2.2. Multi-speaker word materials

2.2.1. Speakers

The speakers were 24 Japanese natives (12 males and 12 females). Their average age was 26.2 years (Min = 21, Max = 30, SD = 3.18).

2.2.2. Word materials

Thirty-six words (12 words × 3 phonemes) that match the conditions described in Section 2.1.2 were used as the word materials.

The word materials were pronounced by a speaker and recorded in a quiet room. In each trial, one of the words was presented on a computer screen in Japanese hiragana orthography. The speaker was asked to push the start button and then naturally pronounce the presented word at a normal speaking rate. The pronunciation was digitally recorded using a microphone and an A/D converter with 16-bit quantization and 48-kHz sampling frequency. The recording was then stored as a digital audio file on a computer. When the speaker finished each pronunciation, he/she was asked to push the stop button. The computer automatically checked the recorded pronunciation. It gave an alert when the intensity of the pronounced word was too low or too high, or when the beginning or end of the pronounced word was not properly recorded. In these cases, the word had to be recorded again. In addition to the checking done by the computer, an operator monitored the pronunciation and, if problems such as mispronunciation or hesitant pronunciation were found, the words were re-recorded at the end of the recording session. After recording, low-frequency noise in the digital audio files was removed by a high-pass finite-impulse-response filter with 70-Hz cut-off frequency. There were 864 word materials in total (24 speakers × 12 words × 3 phonemes).

3. ANALYSIS

We analyzed [ts], [tɕ], and [s] in the single- and multi-speaker word materials using a one-third-octave bandpass filter. Intensity (root mean square power) was computed for each band and expressed in dB with 10^{-10} as the reference level. The single-speaker word materials were filtered with a center frequency of 800–6300 Hz because the sampling frequency of these words was 16,000 Hz. Likewise, the multi-speaker word materials were filtered with a center frequency of 800–20,000 Hz because the sampling frequency of these words was 48,000 Hz. The mean intensity of each frequency band was calculated by averaging the intensity over the duration of the consonants.

Discriminant analyses of the pairs [tɕ]-[ts], [tɕ]-[s], and [ts]-[s] were conducted for each frequency band, with the mean intensity as the independent variable and the consonant as the dependent variable. The discriminant ratio between each consonant pair was obtained by a discriminant analysis.

4. RESULTS

4.1. Single-speaker word materials

Figure 1 shows the mean output intensity at each center frequency of the one-third-octave bandpass filter for [tɕ], [ts], and [s] in the single-speaker word materials. A two-factor analysis of variance of the consonant and center frequency indicated that the interaction between the consonant and center frequency was significant [$F(18,4761) = 652.7$, $p < .001$]. A simple main effect for each center frequency was significant at the 1% level. Multiple comparison tests showed that [tɕ] was significantly different from [ts] or [s] ($p < .001$) at all center frequencies, except for the [tɕ]-[ts] pair at 1250 Hz. It should be noted in Figure 1 that [ɕ] had much higher intensity than [ts] or [s] at center frequencies of 2500, 3150, 4000, and 5000 Hz. However, [ts] and [s] had almost the same intensity at these center frequencies.

Figure 2 shows the discriminant ratios between the consonant pairs ([tɕ]-[ts], [tɕ]-[s], and [ts]-[s]) as functions of the center frequency of the one-

third octave bandpass filter. The discriminant ratio between [tɕ] and [ts] or [s] was very high (more than 80%) at center frequencies of 2500, 3150, 4000, and 5000 Hz. Contrastingly, the discriminant ratio between [ts] and [s] was very low (near the chance level of 50%) at these center frequencies.

Figure 1: Mean output intensities of one-third octave bandpass filter (for [tɕ], [ts], and [s] in single-speaker word materials) as functions of center frequency. The error bars represent the standard deviation.

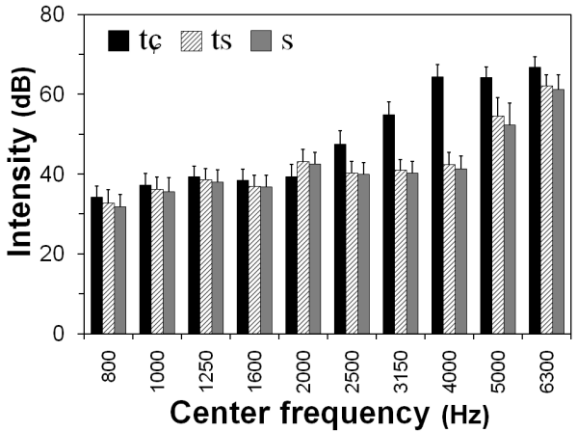
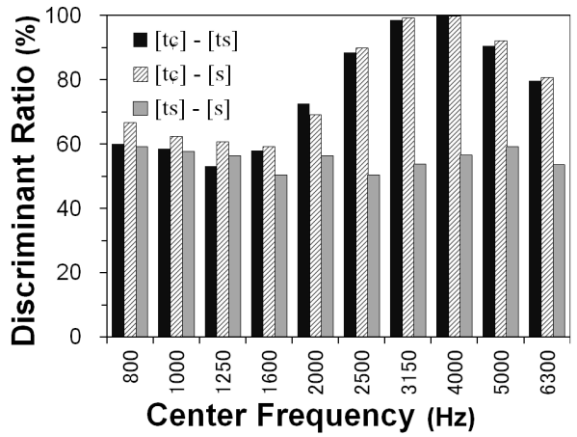


Figure 2: Discriminant ratios between consonant pairs ([tɕ]-[ts], [tɕ]-[s], and [ts]-[s]) of single-speaker word materials as functions of center frequency of one-third octave bandpass filter.



4.2. Multi-speaker word materials

Figure 3 shows the mean output intensity at each center frequency of the one-third octave bandpass filter for [tɕ], [ts], and [s] in the multi-speaker word materials. A two-factor analysis of variance of the consonant and center frequency indicated that interaction between the consonant and center frequency was significant [$F(28,644) = 343.7$, $p < .001$]. A simple main effect for each center frequency was significant at the 1% level. Multiple comparison tests showed that [tɕ] was significantly different from [ts] or [s] ($p < .05$) at all center frequencies, except for the [tɕ]-[s] pair at 1600, 16,000, or 20,000 Hz and the [tɕ]-[ts] pair at 6300 or 8000 Hz. It should be noted in Figure 3 that [tɕ] had much higher intensity than [ts] or [s] at center frequencies of 2500, 3150, and 4000 Hz. However, [ts] and [s] had almost the same intensity at these center frequencies.

Figure 3: Mean output intensities of one-third octave bandpass filter (for [tɕ], [ts], and [s]) in multi-speaker word materials as functions of center frequency. The error bars represent the standard deviation.

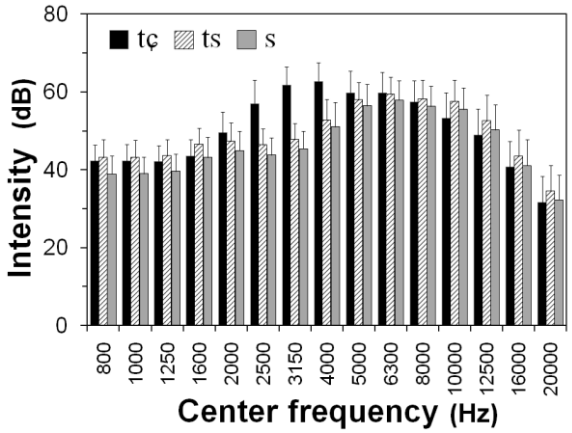


Figure 4: Discriminant ratios between consonant pairs ([tɕ]-[ts], [tɕ]-[s], and [ts]-[s]) of multi-speaker word materials as functions of center frequency of one-third octave bandpass filter.

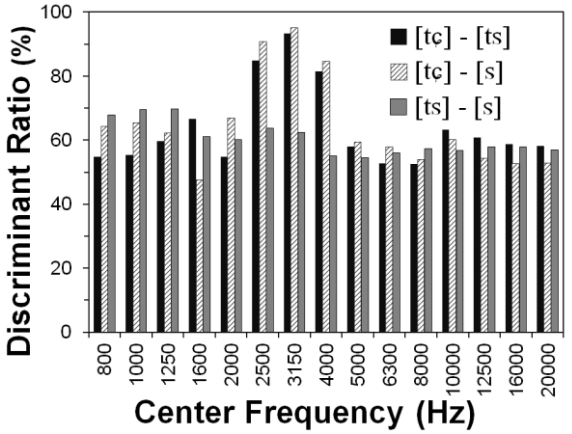


Figure 4 shows the discriminant ratios between the consonant pairs ([tɕ]-[ts], [tɕ]-[s], and [ts]-[s]) as functions of the center frequency of the one-third octave band-pass filter. The discriminant ratio between [tɕ] and [ts] or [s] was very high (more than 80%) at the center frequencies of 2500, 3150, and 4000 Hz. Contrastingly, the discriminant ratio

between [ts] and [s] was very low (55.0–63.7%) at these center frequencies.

5. DISCUSSION

Analyses with a one-third octave bandpass filter and discriminant ratios indicate that [tɕ] can be discriminated from [ts] or [s] in terms of intensity in the frequency range of 2500-5000 Hz for single-speaker word materials and 2500-4000 Hz for multi-speaker word materials. The intersection of these frequency ranges clearly suggests that intensity in the frequency range of 2500-4000 Hz is a good acoustic feature by which to discriminate [tɕ] from [ts] or [s].

However, similar analyses reveal that intensity in the frequency range of 2500-4000 Hz cannot effectively discriminate between [ts] and [s]. This means that intensity in this frequency range is not a good acoustic feature by which to discriminate [ts] from [s]. If intensity in the frequency range of 2500–4000 Hz cannot discriminate [ts] from [s], what is a good acoustic feature by which to discriminate between them? As described in the introduction above, [ts] and [s] are well discriminated by rise duration and steady+decay duration [5]. In addition, variables of the rise duration and steady+decay duration also effectively discriminate [tɕ] from [s] [4]. However, these two variables were not able to discriminate [tɕ] from [ts] [4].

Combined with the previous results, the present results support the conclusion that [tɕ] and [ts] are discriminated from [s] by the rise and steady+decay durations, whereas [tɕ] is discriminated from [ts] and [s] by intensity in the frequency range of 2500–4000 Hz. In other words, consonants with the same place of articulation (such as [ts] and [s]) are discriminated in the time domain, but consonants with the same manner of articulation (such as [tɕ] and [ts]) are discriminated in the spectral domain.

The discriminant ratio between [ts] and [s] was found to be very high (96.2%; i.e., 3.8% error) in the time domain [5]. Similarly, the discriminant ratio between [tɕ] and [ts] or [s] was also found to be very high (93.2–99.1% at 3150 Hz) in the spectral domain. These results suggest that each domain independently contributes to the discrimination between [tɕ], [ts], and [s]. That is, any interactions between the time and spectral domains should be very small or negligible.

These conclusions may be extended to distinguish the alveolo-palatal fricative [ɕ] from the alveolo-palatal affricate [tɕ] or the alveolar fricative [s]. It is speculated that [ɕ] is separated from [tɕ] in the time domain, because each has the same place of articulation but a different manner of articulation. It is further speculated that [ɕ] is separated from [s] in the spectral domain, because each has the same manner of articulation but a different place of articulation. Future research should examine not only these speculations but also whether the present conclusion might be applied to the discrimination of affricates and fricatives for various phoneme environments and speaking rates. The findings of the present and future researches may contribute to the development of scientific and effective methods of speech education for non-native learners of Japanese.

6. ACKNOWLEDGEMENTS

This work was supported by a Grant-in-Aid for Scientific Research (C) (21530782) and by a Grant-in-Aid for Young Scientists (B) (22720173).

7. REFERENCES

[1] Amano, S., Kondo, T. 1999. *Lexical Properties of Japanese Vol. 1 (Nihongo no Goi-tokusei)*. Tokyo: Sanseido (in Japanese).

[2] Strevens, P. 1960. Spectra of fricative noise in human speech. *Language and Speech* 3(1), 32-49.

[3] Yamakawa, K. 2008. Pronunciation characteristics of Japanese affricate [ts] by Japanese learners: In case of Thais. *Proc. International Conference of Japanese Language Education* 2, 326-329.

[4] Yamakawa, K., Amano, S. 2010. Effectiveness of the noise part duration to distinguish between [ts], [tɕ] and [s]. *Proc. Phonetic Society of Japan 2010*, 137-141.

[5] Yamakawa, K., Amano, S., Itahashi, S. 2009. Production boundary between fricative and affricate in Japanese and Korean speakers. *Proc. Interspeech 2009*, 468-471.

[6] Yamakawa, K., Chisaki, Y., Usagawa, T. 2006. Subjective evaluation of Japanese voiceless affricate spoken by Korean. *Acoustical Science and Technology* 27, 236-238.