Discrimination between Fricative and Affricate in Japanese Using Time and Spectral Domain Variables

Kimiko Yamakawa, Shigeaki Amano
Department of Human Informatics, Aichi Shukutoku University, Japan
jin@asu.aasa.ac.jp, psy@asu.aasa.ac.jp

Abstract

Previous studies on Japanese fricatives and affricates revealed that [s] and [ts] are classified by variables in the time domain (Yamakawa et al., 2012), whereas [ts] and [ʨ] as well as [s] and [ʨ] are classified by variables in the spectral domain (Yamakawa & Amano, 2011). To gain an integrated perspective on these findings, this study examined whether [s] [ts], and [ʨ] can be classified by a single discriminant model. Using variables in the time and spectral domains, canonical discriminant analysis was performed on word materials with [s], [ts], and [ʨ] pronounced by both single and multiple Japanese speakers. The time domain variables were a combination of the rise duration and sum of steady and decay durations of the three consonants. The spectral domain variable was their intensity obtained from a one-third-octave bandpass filter with a center frequency of 3150 Hz. The results showed that [s], [ts], and [ʨ] were successfully classified with high discriminant ratios of 92.6% and 90.2% for single and multi-speaker materials, respectively. This means that, in both cases, the single discriminant model can discriminate [s], [ts], and [ʨ] with the variables in the time and spectral domains.

Index Terms: fricative, affricate, time domain variable, spectral domain variable, discrimination

1. Introduction

Non-native speakers of Japanese, such as speakers of Korean and Thai frequently confuse the Japanese voiceless alveolar fricative [s], voiceless alveolar affricate [ts], and voiceless alveolo-palatal affricate [ʨ] (e.g., [1], [2], and [3]). For example, they pronounce [ʧuŋkʊkʊ] ‘junior high school’ as [ʧuŋkʊkʊ] ‘go to school’. These errors in pronunciation can cause misunderstandings in speech communication because listeners recognize a word different from what the speaker intended. The confusion between these consonants is probably caused by their similar acoustic features. 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.

Voiceless fricatives (e.g., [s], [ʃ], and [ɕ]) and affricates (e.g., [ts], [ʧ], and [ʨ]) consist of a frication. Previous studies revealed that the fricatives and affricate consonants have different acoustic features. For example, Howell and Rosen [4] showed that the friction rise time of [ʃ] is longer than that of [s] in English. Yamakawa et al. [5] reported that [s] and [ʦ] in Japanese are well discriminated by two variables of a rise time and steady-decay time of frication. In contrast to these studies of acoustic features in the time domain, some previous studies investigated acoustic features in the spectral domain.

For instance, Stevens [6] showed that English fricatives have different spectral peaks according to their place of articulation. Jongman et al. [7] and Maniwa et al. [8] showed that a spectral peak, a spectral moment, and a normalized- and relative-amplitude contribute to discrimination of English fricatives. Yamakawa and Amano [9] conducted discriminant analyses in a spectral domain and found that [ʦ] and [ʨ], as well as [s] and [ʨ], are well separated with an intensity in a frequency range of 2500 Hz to 4000 Hz, but [s] and [ʦ] are not.

However, the previous studies focused on acoustic features of fricatives and affricates in either the time or spectral domain. They did not treat the features together nor provide a view of how fricatives and affricates relate to each other in the space of acoustic features. On this background, this study analyzed a fricative [s] and affricates [ʦ] and [ʨ] at once to gain an integrated view of the characteristic of these consonants. That is, using the variables in the time and spectral domains, canonical discriminant analysis was performed on [s], [ʦ], and [ʨ] pronounced by single and multiple Japanese speakers. If discriminant ratio is high among consonants [s], [ʦ], and [ʨ], we can say that the independent variables in the canonical discriminant analysis represent distinctive acoustic features of the consonants.

2. Analysis 1: Single-speaker materials

2.1. Speaker

The data were from a female Japanese native speaker, aged 29. She was well experienced in pronunciation for recordings.

2.2. Word materials

Words were selected from the Japanese word familiarity database [10], in which about 70,000 words were pronounced at a normal speaking rate by the speaker mentioned above. 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 an 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 /s/, /ʦ/, or /ɕ/;
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 181, 180, and 127 for [s], [ʦ], and [ʨ], respectively. There were 488 words total.
2.3. Variables

2.3.1. Time domain variable

The durational variables used by Yamakawa et al. [5] were the time domain variables in this study. That is, the waveforms of fricatives and affricates were divided into three parts (rise, steady, and decay), and the intensity envelope of the rise, steady, and decay parts was approximated by three lines with positive, zero, and negative slopes, respectively (Figure 1). The lines for intensity envelope were determined by the automatic fitting method proposed by Yamakawa et al. [5]. This method finds the optimal intensity envelope in terms of the least-square error. From the determined lines, the duration of each part was calculated. The duration of rise \( x \) and duration of steady+decay \( y+z \) were used as the time domain variables.

2.3.2. Spectral domain variable

The intensity output of a one-third-octave bandpass filter with a center frequency of 3150 Hz was used as the spectral domain variable in this study, because the intensity of this frequency range was the most relevant to discriminate \([s]\) and \([ɕ]\), as well as \([s]\) and \([ɕ]\) in previous studies [5], [9]. The mean intensity of the 3150 Hz band for whole frication of each consonant was computed by averaging the output of the bandpass filter, and the intensity was expressed in dB with \(10^{10}\) as a reference level.

2.4. Discriminant analysis

Canonical discriminant analysis among the consonants \([s]-[ts]-[ʨ]\) was performed. Independent variables were the two time domain variables and the spectral domain variable. The dependent variable was a consonant category. Discriminant functions were obtained as Eqs. 1 and 2,

\[
-0.078x_1 - 0.064x_2 - 0.039x_3 + 8.634 = 0 \quad (1)
\]

\[
-0.004x_1 - 0.020x_2 + 0.241x_3 - 12.850 = 0 \quad (2)
\]

where \(x_1\), \(x_2\), and \(x_3\) are the rise duration, steady+decay duration, and intensity at the 3150 Hz band, respectively.

Figure 2 shows the consonant data \([s]-[ts]-[ʨ]\) plotted on a plane of the discriminant scores of Eqs. 1 and 2. Table 1 shows the mean and standard deviation of the discriminant scores. The discriminant ratio among the consonants \([s]-[ts]-[ʨ]\) was 92.6%. The individual discriminant ratios were 95.6%, 91.1%, and 90.6% for \([s]\), \([ts]\), and \([ʨ]\), respectively. These high discriminant ratios indicate that the three variables are effective for discrimination of the three sounds.

Figure 1: Schematic diagram of the intensity envelope pattern.

Figure 2: Scattergram of \([s]-[ts]-[ʨ]\) spoken by the same Japanese native speaker. The horizontal and vertical axis, respectively, represents the discriminant score of the discriminant function of Eqs. 1 and 2.
Table 1: Mean and standard deviation of discriminant score of the discriminant function of Eqs. 1 and 2.

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Eq. 1 Mean</th>
<th>Eq. 1 SD</th>
<th>Eq. 2 Mean</th>
<th>Eq. 2 SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>[s]</td>
<td>-2.54</td>
<td>1.18</td>
<td>0.04</td>
<td>1.18</td>
</tr>
<tr>
<td>[ts]</td>
<td>1.57</td>
<td>0.94</td>
<td>0.91</td>
<td>0.96</td>
</tr>
<tr>
<td>[ʨ]</td>
<td>1.40</td>
<td>0.78</td>
<td>-1.35</td>
<td>0.81</td>
</tr>
</tbody>
</table>

3. Analysis 2: Multi-speaker materials

3.1. Speaker

The data were from 24 Japanese native speakers (12 males and 12 females). Their average age was 26.2 years (range: 21–30, SD = 3.18).

3.2. Word materials

The word selection conditions were as follows:

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

Thirty-six words (12 words × 3 phonemes) that match the conditions described above 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. 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 (Sony, ECM-999) and A/D converter (Roland, UA25-EX) 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 pushed the stop button.

The computer automatically checked the recorded pronunciation. It issued an alert when the intensity of the pronounced word was too low or too high, or when the beginning or end of the 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 the recording, low-frequency noise in the digital audio files was removed by a high-pass finite-impulse-response filter with a 70-Hz cut-off frequency. There were 864 word materials in total (24 speakers × 12 words × 3 phonemes).

3.3. Variables

The same variables of Analysis 1 were used in Analysis 2.

3.4. Discriminant analysis

Canonical discriminant analysis among the consonants [s]-[ts]-[ʨ] was performed. The independent variables were the two time domain variables and the spectral domain variable. The dependent variable was consonant category. Discriminant functions were obtained as Eqs. 3 and 4,

\[
0.025x_1 + 0.026x_2 - 0.183x_3 + 6.359 = 0 \quad (3)
\]

\[
0.034x_1 + 0.032x_2 + 0.130x_3 - 10.700 = 0 \quad (4)
\]

Figure 3: Scattergram of [s], [ts], and [ʨ] spoken by Japanese speakers (n = 24). The horizontal and vertical axes, respectively, represent the discriminant score of the discriminant function of Eqs 3 and 4.
where $x_1$, $x_2$, and $x_3$ are the rise duration, steady+decay duration, and intensity at the 3150 Hz band, respectively.

Figure 3 shows the consonant data ([s], [ts], and [ʨ]) plotted on a plane of the discriminant scores of Eqs. 3 and 4. Table 2 shows the mean and standard deviation of the discriminant score. The discriminant ratio among the consonants [s]-[ts]-[ʨ] was 90.2%. The individual discriminant ratio was 89.6%, 87.2%, and 93.8% for [s], [ts], and [ʨ]. These high discriminant ratios indicate that three variables are effective for the discrimination of [s]-[ts]-[ʨ].

Table 2: Mean and standard deviation of discriminant score of the discriminant function of Eqs. 3 and 4.

<table>
<thead>
<tr>
<th>Consonant</th>
<th>Eq.3 Mean</th>
<th>Eq.3 SD</th>
<th>Eq.4 Mean</th>
<th>Eq.4 SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>[s]</td>
<td>2.28</td>
<td>1.09</td>
<td>0.66</td>
<td>1.23</td>
</tr>
<tr>
<td>[ts]</td>
<td>0.15</td>
<td>0.89</td>
<td>-1.21</td>
<td>0.91</td>
</tr>
<tr>
<td>[ʨ]</td>
<td>-2.43</td>
<td>1.01</td>
<td>0.55</td>
<td>0.81</td>
</tr>
</tbody>
</table>

4. Discussion

The results for both single speaker materials (Analysis 1) and multi-speaker materials (Analysis 2) revealed that three variables, the duration of rise ($x$), duration of steady+decay ($y+z$), and intensity at the 3150 Hz band can well discriminate [s], [ts], and [ʨ]. In other words, the three consonants can be classified by a single discriminant model with a combination of variables in the time and spectral domains. This is an advance in the understanding of characteristic differences between [s], [ts], and [ʨ], because, in previous studies, the discriminant model could only treat differences either in the time domain ([s] vs. [ts]) [5] or the spectral domain ([s] vs. [ʨ]) [9]. It can be said that the current results provide an integrated model of the previous findings, which are represented by two unrelated models.

The current results provide a discriminant model for the consonants [s], [ts], and [ʨ]. These consonants are frequently mispronounced by non-native Japanese speakers from first language backgrounds such as Korean, Thai, and Vietnamese. The discriminant model in this study can be applied to a computer-aided instruction for non-native Japanese speakers to learn the correct pronunciation of the consonants. That is, using the discriminant model, a computer automatically categorizes the phonemes pronounced by non-native Japanese speakers, and it immediately gives them feedback on whether their pronunciation correctly belongs to their intended phoneme category. This feedback must promote non-native Japanese speakers to learn Japanese speech. The current results contribute to scientific knowledge about the phonetic characteristics of [s], [ts], and [ʨ], but they can also contribute to education technology, as described above.

This study verified the effectiveness of the discriminant model for affricates and fricatives from the perspective of the acoustic features in speech production. Many researchers take it for granted that the production and perception of speech are closely related (e.g., [11] or [12]). Therefore, it is worthwhile to verify whether the discriminant model of speech production in this study can give an accurate account of consonant discrimination in speech perception. It is possible that the acoustic features to distinguish between fricatives and affricates are the same in both the production and perception of speech.

However, the current study has some unsolved problems. For example, it did not investigate the Japanese alveolo-palatal fricative [ɕ]. A more general and clear perspective can be obtained for the distinction between fricatives and affricates if [ɕ] is included in the canonical discriminant analysis. Speaking rate is another problem to be considered for discrimination, because features in the time domain are usually affected by the speaking rate. The current study treated voiceless affricates and fricatives. Voiced affricates (e.g., [dz], [dk]) and fricatives (e.g., [z], [ʒ]) might have different acoustic features for discrimination. Future research should examine and clarify these problems.

5. Acknowledgment

This study was supported by JSPS KAKENHI Grant Number 22720173, 24652087, and 25284080. It was also supported by a cooperative-research grant of Aichi-Shukutoku University in 2013. A part of this study was supported by the collaborative research project “Construction of corpus Japanese linguistics” at the National Institute for Japanese Language and Linguistics.

6. References