Mapping Japanese alveolar fricative, alveolar affricate, and alveolo-palatal affricate in time-spectral space

Shigeaki Amano¹,* and Kimiko Yamakawa²,†

¹Faculty of Human Informatics, Aichi Shukutoku University, 2–9 Katahira, Nagakute, 480–1197 Japan
²Faculty of Contemporary Culture, Shokei University, 2–6–78 Kuhonji, Chuo-ku, Kumamoto, 862–8678 Japan

(Received 19 September 2020, Accepted for publication 28 December 2020)

Abstract: Previous studies of speech production have approached the acoustic features of fricatives and affricates in the time and spectral domains. To gain an integrated perspective on the consonants, this study examined whether alveolar fricative /s/, alveolar affricate /ts/, and alveolo-palatal affricate /tc/ can be mapped in a single time-spectral space. The time-domain variables were the rise duration and the sum of steady and decay duration of the consonants. The spectral-domain variable was the mean intensity of the consonants obtained with a one-third-octave bandpass filter with a center frequency of 3,150 Hz. Using these variables, canonical discriminant analysis was performed for /s/, /ts/, and /tc/ pronounced by a native Japanese speaker. The results showed that /s/, /ts/, and /tc/ were successfully mapped in the time-spectral space with high discriminant ratio. This means that the Japanese fricative and affricates having different acoustic features can be integrated into a single representation of the time-spectral domain with a good separation.

Keywords: Fricative, Affricate, Canonical discriminant analysis

1. INTRODUCTION

Many languages distinguish fricatives and affricates that have similar acoustic features. That is, both of fricatives and affricates consist of a friction that is a sustained non-periodic random noise. However, the affricates have a burst part at the beginning of the friction while the fricatives do not. Previous studies of speech production have approached the acoustic features of fricatives and affricates in the time and spectral domains.

An example of research in the time domain is the study of Howell and Rosen [1]. They measured the rise time of frication in affricate /f/ and fricative /f/ at initial, medial, and final positions in words and nonsense syllables spoken by four native English speakers. Their results showed that /f/ has a significantly shorter friction rise time than /f/. This means that the friction rise time is an acoustic feature that distinguishes affricates and fricatives. Liu et al. [2] analyzed the fricatives (/s/, /ʃ/, and /ʃ/) and the affricates (/tsʃ/, /tsʃ/, and /tcʃ/) spoken by 10 male native Mandarin speakers. Their results showed that fricatives have a longer frication than affricates, and that the occurrence of an initial burst is lower in fricatives than affricates.

An example of research in the spectral domain is the study of Strevens [3]. He reported that English fricatives have different spectral shapes according to their place of articulation. For example, the lowest frequency of spectrum is higher than 3,500 Hz for alveolar fricative /s/, whereas it is between 1,600 Hz and 2,500 Hz for alveolo-palatal fricative /ʃ/. Jongman et al. [4] investigated eight English fricatives, /f/, /v/, /θ/, /ð/, /s/, /z/, /ʃ/, and /ʒ/, spoken by 20 native English speakers. They showed that the spectral peak location, spectral moment, and normalized- and relative-amplitude contribute to the discrimination of fricative sub-categories defined by their place of articulation.

Fricatives and affricates are distinguished in Japanese. They include an alveolar fricative /s/, an alveolo-palatal fricative /ʃ/, alveolar affricate /ts/, and alveolo-palatal affricate /tc/. Probably because these consonants have similar acoustic features, non-native Japanese speakers, such as Korean, Chinese, Thai, and Vietnamese speakers frequently confuse them [e.g., 5–8].

However, frequency of confusion differs between the consonants. For example, Korean speakers frequently confuse /s/, /ts/, and /tc/ [e.g., 9] but not /ʃ/ [e.g., 10].
Thereby, Japanese words with /s/, /ts/, and /tɕ/, such as /surnakuru/ 'mathematics,' /tsurnakuru/ 'going to school,' and /tɕurnakuru/ 'junior high school,' are difficult for non-native Japanese speakers to pronounce with a clear distinction.

Because non-native Japanese speakers frequently confuse the /s/, /ts/, and /tɕ/, this study focused on these three consonants and tried to obtain scientific knowledge about acoustic features to distinguish them. Because native Japanese speakers can clearly distinguish the consonants, specifying the acoustic features that they use would be useful for non-native speakers. It is expected that the knowledge would contribute to develop a teaching method for non-native speakers to reduce mispronunciations and promote correct and clear pronunciations of these Japanese consonants in future.

Like the studies of fricatives and affricates in English [1,3,4] and Mandarin [2], previous studies of these consonants in Japanese have approached their acoustic features in the time and spectral domains. For example, Yamakawa et al. [11] analyzed the fricative /s/ and affricate /ts/ spoken by native Japanese speakers. They divided the intensity envelope of the consonants into rise, steady, and decay parts that were approximated by three lines with positive, zero, and negative slopes, respectively (Fig. 1). The lines were determined by an automatic fitting method [11] that finds the optimal intensity envelope in terms of the least square error. From the determined lines, the durations of rise (x1) and “steady + decay” (x2) were obtained. Their results showed that /s/ and /ts/ are discriminated well by the two variables x1 and x2. This means that x1 and x2 are acoustic features characterizing the distinction between /s/ and /ts/. In other words, /s/ and /ts/ are separated by these durational variables in the time domain.

An example of the study in the spectral domain is Yamakawa and Amano [12]. They analyzed affricates /ts/ and /tɕ/ in terms of intensity in a frequency band. In their study, the waveform of the affricates was passed through a high-pass filter at a center frequency of 3,150 Hz. They found that the mean intensity at the 3,150 Hz band can separate /ts/ and /tɕ/ with high accuracy. This means that the intensity at the 3,150 Hz band is an acoustic feature that distinguishes /ts/ and /tɕ/.

These previous studies have contributed to our knowledge of the acoustic features of fricatives and affricates in both the time and spectral domains. However, they focused on the acoustic features 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 integrated time-spectral domain.

Another problem is that the previous studies limited their investigation to individual consonant pairs, such as /s/ vs. /ts/ [11] or /ts/ vs. /tɕ/ [12]. To obtain a comprehensive view of the acoustic features of fricatives and affricates, it is necessary to jointly investigate more than two consonants. To deal with this problem, this study analyzed simultaneously /s/, /ts/, and /tɕ/ in the time-spectral domain to gain an integrated view of their characteristics.

It has been known that /s/ and /ts/ are distinguished in the time domain [11] while /ts/ and /tɕ/ are distinguished in the spectral domain [12]. Yamakawa et al. [11] used two variables in the time domain and Yamakawa and Amano [12] used one variable in the spectral domain. When these three variables are used, the distributions of the three consonants (/s/, /ts/, and /tɕ/) can be mapped on a three-dimensional space. However, to achieve a more concise and easily recognizable representation, this study attempted to map the consonants onto a two-dimensional space by applying canonical discriminant analysis to the consonant data. More precisely, this study examined whether /s/, /ts/, and /tɕ/ can be mapped and well separated in a time-spectral space using the time variables of rise duration [11], the sum of the steady and decay durations [11], and the spectral variable of the mean intensity of the 3,150 Hz band [11]. If the canonical discriminant analysis results in a high discriminant ratio, it indicates that the /s/, /ts/, and /tɕ/ are well separated in the time-spectral space. Moreover, this means that the consonants can be represented in a single time-spectral space.

2. ANALYSIS

2.1. Speech Materials

Speech materials were selected from a Japanese word familiarity database [13] that contains about 70,000 words spoken at a normal speaking rate by a 29-year-old female Japanese native speaker of a southern Kanto dialect. The spoken words in the database are stored as digital audio files with 16-bit quantization and 16-kHz sampling frequency. The criteria for selecting the speech materials were as follows:

![Fig. 1 Conceptual diagram of the intensity envelope of fricative and affricate consonants.](image)
Table 1 Mean and standard deviation of time-domain variables ($x_1$ and $x_2$) and spectral-domain variable ($x_3$) for three consonants ($/s/$, /ts/, and /tc/) in Sets 1 and 2.

| Set | Consonant | n  | $x_1$ (ms) | $x_2$ (ms) | $x_3$ (dB) |
|-----|-----------|----|------------|------------|------------|
|     |           |    | Mean       | SD         | Mean       | SD         |
| 1   | /s/       | 63 | 47.8       | 12.9       | 85.7       | 19.1       | 39.7       | 2.8        |
|     | /ts/      | 63 | 27.0       | 10.1       | 46.0       | 10.6       | 40.3       | 2.6        |
|     | /tc/      | 63 | 32.2       | 8.9        | 46.8       | 11.4       | 53.7       | 3.4        |
| 2   | /s/       | 63 | 47.1       | 13.7       | 88.4       | 16.8       | 38.8       | 2.9        |
|     | /ts/      | 63 | 28.2       | 11.4       | 45.3       | 13.5       | 40.0       | 2.7        |
|     | /tc/      | 63 | 32.0       | 9.5        | 46.3       | 11.3       | 54.4       | 3.3        |

1) /s/, /ts/, or /tc/ is in word-initial position;
2) a non-devoiced vowel /u/ follows /s/, /ts/, or /tc/;
3) the word length is three or four morae.

From the words that satisfied these criteria, two sets of words were randomly selected. Set 1 was for obtaining canonical discriminant functions and Set 2 was for verifying generality of the functions. Sets 1 and 2 had 63 words each for /s/, /ts/, and /tc/, so that each set had 189 speech materials in total.

2.2. Procedure

The canonical discriminant model expressed in the discriminant functions (Eq. (1)) was used for the analysis.

$$f_i = a_{0i} + a_{1i}x_1 + a_{2i}x_2 + a_{3i}x_3.$$  (1)

In this equation, $i$ is the index of the discriminant function, $f_i$ is the discriminant score, $a_{0i}$ are discriminant coefficients, $x_1$ is the rise duration, $x_2$ is the “steady + decay” duration [11], and $x_3$ is the mean intensity of a one-third-octave bandpass filter with 3,150 Hz center frequency [12].

The index $i$ was set to two in the analysis, which means a conversion of three-dimensional data ($x_1$, $x_2$, $x_3$) to two-dimensional data ($f_1$, $f_2$) via Eq. (1). In other words, data reduction to a two-dimensional space was achieved using Eq. (1).

The canonical discriminant model (Eq. (1)) is a linear combination of the time-domain variables ($x_1$ and $x_2$) and the spectral-domain variable ($x_3$). This means that acoustic features in the time and spectral domains are incorporated into a single expression.

The values of $x_1$, $x_2$, and $x_3$ were calculated for the speech materials in Sets 1 and 2. Table 1 shows the means and standard deviations of $x_1$, $x_2$, and $x_3$. We can see in Table 1 that $x_1$, $x_2$, and $x_3$ are different between the consonants /s/, /ts/, and /tc/, suggesting that these variables are effective acoustic features to distinguish the consonants.

Using the three variables, canonical discriminant analysis with the model expressed in Eq. (1) was performed for the three consonants in Set 1. To verify the effectiveness of the canonical discriminant model, the discriminant functions obtained from Set 1 were applied to Set 2 and discriminant ratios of the three consonants were calculated.

2.3. Results

Canonical discriminant functions for Set 1 were obtained as Eqs. (2) and (3),

$$f_1 = 14.44 - 0.002_{x_1} - 0.015_{x_2} + 0.346_{x_3}$$  (2)

$$f_2 = -5.32 + 0.085_{x_1} + 0.073_{x_2} - 0.045_{x_3}$$  (3)

where $f_1$ and $f_2$ are discriminant scores of canonical discriminant functions and $x_1$, $x_2$, and $x_3$ are the rise duration, steady + decay duration, and mean intensity at the 3,150 Hz band, respectively.

By applying the varimax rotation to the data, Eqs. (2) and (3) were transformed into Eqs. (4) and (5),

$$F_1 = -0.020_{z_1} - 0.219_{z_2} + 1.030_{z_3}$$  (4)

$$F_2 = 0.915_{z_1} + 1.045_{z_2} - 0.135_{z_3}$$  (5)

where $F_1$ and $F_2$ are discriminant scores of rotated canonical discriminant functions and $z_1$, $z_2$, and $z_3$ are the normalized variables of $x_1$, $x_2$, and $x_3$ respectively. Figure 2 shows the items of /s/, /ts/, and /tc/ plotted on the plane defined by the discriminant functions of Eqs. (4) and (5). The items on the plane were discriminated into one of three consonant groups according to the distance from the centroid of each consonant group. Namely, the items were categorized by the three boundary lines in Fig. 2. Note that each of the three lines exists at the same distance from the two centroids out of three centroids of the consonant groups.

Discriminant ratios of /s/, /ts/, and /tc/ in Set 1 are shown in Table 2. The overall discriminant ratio of these three consonants was 98.4% ($n = 189$). This high discriminant ratio indicates that the canonical discriminant model successfully discriminates /s/, /ts/, and /tc/ in Set 1.

The discriminant functions of Eqs. (4) and (5) were applied to the items in Set 2 to get their discriminant scores. That is, $x_1$, $x_2$, and $x_3$ of the items were respectively
converted to $z_1$, $z_2$, and $z_3$ by dividing the difference of $x_i$ from its average by the square root of a diagonal element of pooled covariance matrix. From the $z_1$, $z_2$, and $z_3$, discriminant scores were obtained using Eqs. (4) and (5).

Using the discriminant scores, discriminant ratios of /s/, /ts/, and /tC/ in Set 2 were calculated with the same procedure as Set 1. That is, the items of Set 2 were categorized by the Set 1’s three boundary lines in Fig. 2. Discriminant ratios of Set 2 are shown in Table 2. The overall discriminant ratio of these three consonants was 98.9% ($n = 189$). This high discriminant ratio indicates that the canonical discriminant model is effective and stable in discriminating /s/, /ts/, and /tC/.

### 3. DISCUSSION

This study showed that the canonical discriminant model can discriminate /s/, /ts/, and /tC/ with a high discriminant ratio, and it can map the consonants on a two-dimensional plane. This means that the fricative (/s/) and affricates (/ts/ and /tC/) having different acoustic features can be integrated into a single time-spectral representation via linear functions of time and spectral variables. This is an advance in characterizing /s/, /ts/, and /tC/, because previous studies only treated the differences either in the time domain [e.g., 11] or in the spectral domain [e.g., 12], whereas this study analyzed the three consonants simultaneously and provides their concise representation (Fig. 2).

The current canonical discriminant model has an advantage over previous discriminant studies. For example, it integrates acoustic features in two domains into one simple form to provide a single time-spectral representation of the consonants. Moreover, because the relative weights of the variables of the time and spectral domains are expressed in the absolute values of the discriminant coefficients, we can compare and interpret the relative importance of each variable in consonant discrimination.

For example, the axes in Fig. 2 may be interpreted from the coefficients of Eqs. (4) and (5) as follows. In the horizontal axis (Eq. (4)), the coefficient of $z_3$ has a large absolute value, but those of $z_1$ and $z_2$ are nearly zero. Because $z_3$ is converted from the mean intensity at the 3,150 Hz band ($x_3$), it can be said that the horizontal axis heavily weights the acoustic features in the spectral domain. On the other hand, the vertical axis (Eq. (5)) heavily weights the acoustic features in the time domain, because the absolute value of the coefficient in Eq. (5) is nearly zero for $z_3$ but large for $z_1$ and $z_2$ which are respectively converted from the durational variables $x_1$ and $x_2$.

These comparisons and interpretations are impossible in previous discriminant studies, because they independently treated the time and spectral domain data. It can be said that the current study provides a comprehensive view of fricatives and affricates in the time-spectral domain.

In Fig. 2, /s/ and /tC/ are separated along the horizontal axis. This indicates that /ts/ and /tC/ differ in the spectral domain, which coincides with the findings of Yamakawa and Amano [12]. The alveolar affricate /ts/ and alveolo-palatal affricate /tC/ differ in the place of articulation which corresponds to a position of the tongue that define a shape of vocal tract. Because the shape of vocal tract is related to a spectral contour [14], the /ts/ and /tC/ having a different place of articulation differ in the spectral domain. By contrast, /s/ and /ts/ are separated along the vertical axis, which indicates that /s/ and /ts/ differ in the time domain. This result coincides with the findings of Yamakawa et al. [11]. The alveolar fricative /s/ and alveolar affricate /ts/ differ in the manner of articulation which corresponds to a movement of the tongue and jaw. Because their movement results in a time variation of speech waveform [15], the /s/ and /ts/ having a different manner of articulation differ in the time domain. These facts mean that two previous findings in the spectral domain.
domain [12] and time domain [11] are integrated into a single time-spectral space in this study.

The current results contribute to scientific knowledge about the phonetic characteristics of /s/, /ts/, and /tʃ/, but they can also contribute to education technology. As described in the introduction, /s/, /ts/, and /tʃ/ are frequently mispronounced by non-native Japanese speakers [e.g., 5–8]. The canonical discriminant functions obtained 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 canonical discriminant functions, a computer can automatically categorize the consonants pronounced by non-native Japanese speakers and provide an immediate feedback on whether their pronunciation correctly belongs to the intended phoneme category. Feedback can be done auditorily or visually. In visual feedback, a non-native Japanese speaker’s pronunciation can be mapped onto a plane defined by the two discriminant functions like Fig. 2 to help the speaker recognize in which consonant category their pronunciation falls. This feedback would help non-native Japanese speakers to learn normal Japanese pronunciation.

This study did not investigate the alveolo-palatal fricative /ʃ/. However, the results may be extended to distinguish /ʃ/ from /s/, /ts/, and /tʃ/. Previous studies suggested that consonants with different manners of articulation (i.e., fricative vs. affricate) can be discriminated by the time domain variables [e.g., 1, 2, 11, 16], whereas consonants with different places of articulation (i.e., alveolar vs. alveolo-palatal) can be discriminated by the spectral domain variables [e.g., 3, 4, 12, 16]. Therefore, the alveolo-palatal fricative /ʃ/ would be distinguished from the alveolo-palatal affricate /tʃ/ in the time domain, and from the alveolar fricative /s/ in the spectral domain. That is, if /ʃ/ is analyzed together with /s/, /ts/, and /tʃ/ using a canonical discriminant model, /ʃ/ will probably be located in the open space in the upper right corner of Fig. 2. If this prediction is confirmed in a future study, it would provide a more general, clearer perspective on the distinction between fricatives and affricates.

The current study treated voiceless consonants rather than voiced consonants. Voiced affricates (e.g., /dʒ/ or /dz/) and fricatives (e.g., /z/ or /ʒ/) might be characterized by different types of acoustic features than their voiceless counterparts. Previous studies have shown that voiceless and voiced fricatives could be categorized in the spectral domain [e.g., 4, 17, 18]. Therefore, the current model might be able to discriminate the voiced consonants with adjustments of parameters in the spectral domain.

Although this study provided clear and reliable results on mapping the consonants onto the time-spectral space, it should be noted that the speech materials were spoken by a single female Japanese speaker. If multi-speaker speech materials were used, individual differences in speech production might degrade the accuracy of consonant discrimination. However, it is expected that almost the same mapping would be obtained for each individual speaker, because individual speakers would probably have similar consonant boundaries, and individual differences would not be so great as to override consonant categories. Even if the individual differences were not negligible, a linear mixed model considering individual’s variation could be applied to the multi-speaker data to obtain a clear mapping. To construct such a linear mixed model, the results of this study can provide a guiding principle and a base model of the time-spectral representation of the data. This study has significance on these points, even though it dealt only with single-speaker data.

Speaking rate would be one issue that has to be considered in the refinement of the current canonical discriminant model. Because the duration of speech segments varies with speaking rate, the time-domain variables in the canonical discriminant model might be affected by the speaking rate. In this study, the speaking rate of the speech materials was normal rather than very fast or slow. This constraint might have unexpectedly boosted the model’s performance with respect to discrimination. If speech materials with a variety of speaking rates were used, the model’s performance might decrease. This means that the “raw” duration of each consonant is fragile as an acoustic feature in the time domain. However, a relative duration, such as the ratio of frication duration to word duration, is probably more stable against the speaking rates. If a relative duration is used instead of a raw duration, the speaking rate effects would be reduced or canceled, and the model’s performance would be improved. This possibility should be examined in a future study.

This study verified the mapping of fricatives and affricates to a time-spectral space from the perspective of acoustic features in speech production. Many researchers take it for granted that speech production and perception are closely related [e.g., 19, 20]. Therefore, it is worthwhile examining whether the mapping of affricates and fricatives can provide a reasonable account of consonant discrimination in speech perception. It is highly probable that the acoustic features to distinguish fricatives and affricates are the same in both production and perception of speech, and that the perception space has a very similar structure as the production space like Fig. 2. Future research is necessary to clarify these points.

ACKNOWLEDGMENT

This study was supported by JSPS KAKENHI.
REFERENCES

[1] P. Howell and P. Rosen, “Production and perception of rise time in the voiceless affricate/fricative distinction,” \textit{J. Acoust. Soc. Am.}, \textbf{73}, 976–984 (1983).

[2] H. M. Liu, C. H. Tseng and F. M. Tsao, “Perceptual and acoustic analysis of speech intelligibility in Mandarin-speaking young adults with cerebral palsy,” \textit{Clin. Linguist. Phon.}, \textbf{14}, 447–464 (2000).

[3] P. Strevens, “Spectra of fricative noise in human speech,” \textit{Lang. Speech}, \textbf{3}, 32–49 (1960).

[4] A. Jongman, R. Wayland and S. Wong, “Acoustic characteristics of English fricatives,” \textit{J. Acoust. Soc. Am.}, \textbf{108}, 1252–1263 (2000).

[5] Y. Sukegawa, “Bogobetsu ni mita hatsuon no keikou [Utterance tendency of non-native Japanese speakers],” Nihongo onsei to Nihongo kyoiku [Japanese Speech and Education], Research Report of Grant-in-Aid for Scientific Research on Priority Areas by Ministry of Education, Science and Culture, pp. 187–244 (1993) (in Japanese).

[6] K. Yamakawa, Y. Chisaki and T. Usagawa, “Subjective evaluation of Japanese voiceless affricate spoken by Korean,” \textit{Acoust. Sci. & Tech.}, \textbf{27}, 236–238 (2006).

[7] K. Yamakawa, “Pronunciation characteristics of Japanese affricate /ts/ by Japanese learners: In case of Thais,” \textit{Proc. Int. Conf. Japanese Language Education}, vol. 2, 326–329 (2008).

[8] M. Kondo, “Language specific and language universal L1 interference on phonological acquisition of Japanese: Results of questionnaires to Japanese teachers,” \textit{Bull. Grad. Div. Lett. Arts Sci. Waseda Univ.}, \textbf{3}, \textit{57}, 21–34 (2012) (in Japanese).

[9] H. Matsuzaki, “Phonetic education of Japanese for Korean speakers,” \textit{J. Phon. Soc. Jpn.}, \textbf{3}(3), pp. 26–35 (1999) (in Japanese).

[10] Y. Nakato, “Phonetic features of Korean speakers of English and Japanese: An analysis based on the results of listening and pronunciation tests,” \textit{J. Phon. Soc. Jpn.}, \textbf{2}(1), pp. 72–82 (1998) (in Japanese).

[11] K. Yamakawa, S. Amano and S. Itahashi, “Variables to discriminate affricate [ts] and fricative [s] at word initial in spoken Japanese words,” \textit{Acoust. Sci. & Tech.}, \textbf{33}, 154–159 (2012).

[12] K. Yamakawa and S. Amano, “A variable to discriminate voiceless affricates at word initial in read and spontaneous Japanese speech,” \textit{J. Phon. Soc. Jpn.}, \textbf{18}(1), 40–44 (2014).

[13] S. Amano and T. Kondo, \textit{Nihongo no Goi-tokusei [Lexical properties of Japanese]}, vol. 1 (Sanseido, Tokyo, 1999) (in Japanese).

[14] K. N. Stevens, \textit{Acoustic Phonetics} (MIT Press, Cambridge, 2000), pp. 127–202.

[15] K. N. Stevens, \textit{Acoustic Phonetics} (MIT Press, Cambridge, 2000), pp. 323–485.

[16] K. Yamakawa and S. Amano, “Discrimination of Japanese fricatives and affricates by production boundaries in time and spectral domains: A case study of a female native speaker,” \textit{Acoust. Sci. & Tech.}, \textbf{36}, 296–301 (2015).

[17] G. W. Hughes and M. Halle, “Spectral properties of fricative consonants,” \textit{J. Acoust. Soc. Am.}, \textbf{28}, 303–310 (1956).

[18] S. J. Behrens and S. E. Blumstein, “Acoustic characteristics of English voiceless fricatives: A descriptive analysis,” \textit{J. Phon.}, \textbf{16}, 295–298 (1988).

[19] S. Amano and Y. Hirata, “Perception and production boundaries between single and geminate stops in Japanese,” \textit{J. Acoust. Soc. Am.}, \textbf{128}(4), 2049–2058 (2010).

[20] K. Nagao and K. de Jong, “Perceptual rate normalization in naturally produced rate-varied speech,” \textit{J. Acoust. Soc. Am.}, \textbf{121}, 2882–2898 (2007).