The perception of /shu/ and /shi/ with devoiced vowels included in geminate/non-geminate words by young normal-hearing listeners

Eri Iwagami1,*, Takayuki Ara1† and Keiichi Tajima2
1Sophia University,
7–1 Kioi-cho, Chiyoda-ku, Tokyo, 102–8544 Japan
2Hosei University,
2–17–1 Fujimi, Chiyoda-ku, Tokyo, 102–8160 Japan
(Received 27 March 2019, Accepted for publication 26 November 2019)

Abstract: In this paper, the results of the experiment for young normal-hearing listeners were presented, which is a part of the whole project, the perception experiment for young normal-hearing and elderly listeners. The perception experiment was conducted by using geminate and non-geminate words with /shu/ and /shi/ (referred to as /ʃu/ and /ʃi/), ‘shukan,’ ‘shukkan,’ ‘shikan’ and ‘shikkan,’ having devoiced and non-devoiced vowels to investigate if the lower limit frequency (F_L) and its falling affect to discriminate /ʃu/ from /ʃi/ with devoiced vowels by young normal-hearing listeners. The synthesised fricative sound was used for C1, and the frication has three types of F_L, 1,440 Hz, 2,760 Hz and 4,080 Hz. Moreover, there are two patterns of the frication with each F_L: 1) F_L of the frication was kept in the same frequency until the end of the frication in each F_L, and 2) the frequencies fall from each F_L to 480 Hz in the middle of the frication. The rate which they perceive as /ʃu/ (referred to as the response rate of /ʃu/) for stimuli with devoiced vowels in F_L falling condition was higher than in F_L non-falling condition. The response rates of /ʃu/ in the both devoiced/non-devoiced condition were also higher in case that the F_L was 1,440 Hz regardless of F_L falling. Hence, these results implied that F_L itself and F_L falling affect to distinguish /ʃu/ from /ʃi/ with devoiced vowels by young normal-hearing listeners. In the results of geminate perception, the tendency was observed that the response rates of geminate for the stimuli with devoiced vowels were almost same as the response rates for the stimuli with non-devoiced vowels, suggesting it was possible for young normal-hearing listeners to discriminate geminate from non-geminate even when vowels were devoiced.

Keywords: Misperception, Elderly listeners, Young normal-hearing listeners, Japanese, Devoiced vowels, Fricative sound, Geminate consonants

PACS number: 43.71.Es, 43.71.Ky, 43.71.Lz [doi:10.1250/ast.41.663]

1. INTRODUCTION

It has known that high vowels, /i/ and /u/ are devoiced as these vowels are between voiceless consonants or at the end of the words in Japanese, especially in Tokyo dialect [1–3]. In previous study [4], we investigated the perception of geminate and non-geminate words with devoiced vowels by young normal-hearing listeners and elderly listeners. As the result, the tendency was observed that young normal-hearing listeners recognise the words with both devoiced and non-devoiced vowels correctly, whereas elderly listeners misperceive consonants including high frequency components, especially /ɹ(u)/ and /ɹ(i)/ (the vowels in the parentheses show devoiced vowels), as the vowels are devoiced. Moreover, there was a tendency for the elderly listeners to perceive /ʃ(u)/ as /ʃ(i)/, thus spectra of /ʃ(u)/ and /ʃ(i)/ were analysed. As the result, the spectra differences of /ʃ(u)/ and /ʃ(i)/ were observed between 2kHz and 8kHz. Hence, it was possible that they misperceive /ʃ(u)/ as /ʃ(i)/, since it was hard to perceive the difference of spectra at high frequency regions for the elderly listeners having hearing threshold elevation at the high frequencies.

However, we only analysed these spectra in the previous study, thus a further investigation about the tendency is significant. Therefore, we considered it would be also required to investigate cues of the perception of
/ʃ(u)/ and /ʃ(ɪ)/ with devoiced vowels by both young normal-hearing and elderly listeners as a process to clarify the fact of the misperception by the elderly listeners.

Many researchers have investigated and reported the differences of acoustic characteristics of /ʃ/ comparing acoustic characteristics of /s/. It is known the friction of sibilant /s/ has the energy around 4–5 kHz in English or French, however, the spectral peak is located around 2–3 kHz in case of /ʃ/, since the front cavity becomes longer than in case of /s/ [5–8]. Moreover, acoustic cues such as spectral moment including kurtosis and skewness of the amplitude of the frication is important [9,10].

The palato-alveolar sibilant fricative /ʃ/ in English is produced with lip protrusion (rounding), and lip protrusion and lip rounding affect the acoustic characteristics of /ʃ/. Faber [11] observed lip protrusion and constricted portion of the oral cavity to investigate the production of sibilants, /s/ and /ʃ/ uttered in context and in isolation. As the results, it was found that a constricted portion was more anterior in /s/ than in /ʃ/, and /ʃ/ had more lip protrusion than /s/ in isolated utterances. Therefore, it was reported that the combination of lip protrusion and the constricted portion affected the centroid frequency of sibilant in isolated utterances, moreover lip protrusion and aperture more affected the centroid of sibilant’s frequency than the constricted portion in utterance with context. Moreover, Ladefoged et al. [12] reported that the peak frequency in palato-alveolar fricatives was lower since the front cavity is wider in palato-alveolar fricatives than dental or alveolar sounds because of the lip protrusion. Therefore, there are differences between the positions of the first peak which shows the resonance frequency of constricted portion of the front cavity. Proctor et al. [13] analysed the manner of articulation by using MRI as /s/ and /ʃ/ were produced individually and they produced with vowels. As the result, /ʃ/ was produced with lip rounding, and the acoustical differences were observed between /s/ and /ʃ/.

Toda [14] investigated effects of the lip rounding and following vowels in the production of alveolar and postalveolar sibilants. It was observed that lips were protruded when following vowel was /u/ than /i/ or /a/ in sibilant, lips were also protruded more as following vowel was /u/ in post-alveolar sibilant than in alveolar sibilant. Soli [15] analysed the spectra of the fricative peak frequencies of /s, z, j, ʒ/ uttered with three vowels, /a/, /i/ and /u/ before the vowel onset (60 ms). As the results, it was reported that the peak frequencies of the fricative sounds were located between 1.5 kHz–2 kHz. It was also clarified that the spectral peak was 100 Hz–300 Hz higher as they are before /i/ than as the sibilants are before /a/ and /u/.

Fricative sound /ʃ/ in Japanese is produced without lip rounding [16], however, it is produced with the shrunk front of dorsal surface of tongue [17] although palato-alveolar in English is produced with the shrunk front of tongue in addition to lip rounding to form lip protrusion [18].

Moreover, the cues of the perception of fricative sounds have been researched. The factor which affects the perception of /s/ and /ʃ/ in English was investigated, and it was clarify that the fricative frequencies and the formant of vowels affected the discrimination of /s/ from /ʃ/ [19]. Nittrouer et al. [20] also showed that the weight was put to the formant transition in the perception of /s/ and /ʃ/ by infant children than by seniority children or adult. Li et al. [21] investigated the cue of the spectra of fricative. As the result, the important cue was included in the fricatives, /ʃ, ʒ, s, z, f, v/, moreover, there are the perception cues at 3.6 kHz and 8 kHz in /s, z/ and at 1.4 kHz and 4.2 kHz in /ʃ, ʒ/. Therefore, Li et al. implied that lower frequency of fricatives and fricative duration were important cues for the perception of fricatives.

Vowels are often devoiced in case that they are between voiceless consonants as above, therefore, following vowels of /ʃ/ are also devoiced as the following consonant was voiceless. It has reported that it was possible to distinguish /ʃ/ in both case that the following vowels are /i/ and /u/ since there are acoustic characteristics of velar or labial-velar as the following vowel is /a/ [22]. Beckman et al. [23] researched the characteristic of the perception other than the characteristic of spectra in /ʃi/ and /ʃu/ with devoiced vowels. As the result, it was found that the participants perceived /ʃi/ as the spectra peak was high even when the vowel was devoiced, whereas they perceived /ʃu/ as the spectra peak was low. Matsui [24] conducted a perception experiment by using the word with fricative at the end of the word. Continuum was created from the fricative sound /s/, which has constriction at alveolar to the fricative sound /ʃ/, which has constriction of the lips too. Moreover, two kinds of stimuli were created; stimulus with steady temporal frequency characteristics and the stimulus added transition at the end of fricative so that peak frequency falls. As the results, it was clarified that the vowel was expected as /u/ for the stimuli with the peak frequency transition, which was under 2,400 Hz, and it was possible to discriminate /ʃ(ɪ)/ from /ʃ(u)/ even when the vowels were devoiced.

It is possible to use the peak frequency transition for elderly listeners as one of the cues to distinguish /ʃ(u)/ from /ʃ(ɪ)/. As above, Li et al. has reported that there is the perception cue of fricative sounds at the lower frequencies; at 3.6 kHz and 8 kHz in /s, z/ and at 1.4 kHz and 4.2 kHz in /ʃ, ʒ/. Therefore, we presumed that the lower limited frequency (referred to as $F_L$) would affect the discrimination. Moreover, we hypothesised that it could be one of the cues of the discrimination in the case.
that $F_L$ falls toward the end of the frication (referred to as $F_L$ falling).

Hence, we conducted perception experiments for both young normal-hearing and elderly listeners as a whole project to investigate if $F_L$ itself and $F_L$ falling are used as cues to identify /ʃ(ʊ)/ from /ʃ(ɪ)/ with devoiced vowels by both listeners. The words used in the experiments included geminate and non-geminate words with /ʃ(ʊ)/ from /ʃ(ɪ)/. This paper describes the results of the experiment for young normal-hearing listeners, and the experiment for elderly listeners will be presented in another paper.

2. STIMULI

2.1. Recordings

We used four real Japanese words, ‘shikan (しかん)’ /ʃikan/ (relaxation), ‘shikkan (しっかん)’ /ʃikkan/ (disease), ‘shukan (しゅかん)’ /ʃukan/ (subjectivity) and ‘shukkan (しゅっかん)’ /ʃukkan/ (carrying out a coffin), which were three-mora words, ‘shikan’ and ‘shukan,’ (non-geminate) and four-mora words, ‘shikkan’ and ‘shukkan,’ (geminate) with $C_1V_1C_2V_2N$ and $C_1V_1C_2V_2N$ (‘ʼ’ shows geminate) structures, respectively (Hereafter, ‘word’ will be shown as these words are explained, on the other hand /IPA/ will be shown as the perception of these speech sound is explained). Accent patterns for these four words were all non-accented pattern. The familiarity of the speech sound for these words was 6.2, 4.3, 5.0 and 4.7 out of seven-point scale, respectively. The familiarity of the orthography (kanji) for the words was 4.3, 5.0, 5.3 and 4.8 out of seven-point scale, respectively. (It was not observed that the familiarities of the speech sound and character representation affected to the results in this experiment.) These four words inserted into the carrier sentence ‘watashi wa [word] to iimashita. (I said ____).’ were recorded by a female standard Japanese speaker by using microphone (SONY ECM-MS957) and PCM recorder (D & M Professional Marantz PMD660). The sampling rate and the quantisation bit rate were 48 kHz and 24 bit, respectively. $V_1$ was produced with non-devoiced vowel although it was between voiceless consonants as we investigate the perception of these words with devoiced vowels and non-devoiced vowels. $V_1$ and $C_1$ were target vowel and consonant in these words (we refer $C_1V_1$ to target syllable).

2.2. Processing of Recorded Speech Sounds

2.2.1. $V_1$ and $C_1$ of each word

Each recorded word was segmented from carrier sentences. The synthesised sound was used for each $C_1$ instead of uttered $C_1$ (the details of synthesis of the fricative sounds is described in Sect. 2.2.2). On the other hand, the uttered vowels (non-devoiced) /i, u/ of non-geminate words, ‘shikan’ and ‘shukan,’ were used for each $V_1$ of the minimal pairs, respectively. Each duration, frication and vowels durations, were set based on the original speech sounds of non-geminate words, ‘shukan’ ‘shikan,’ so that the similar durations as original durations could realise. The frication duration was set to 165 ms since frication duration of both words were approximately 165 ms. The $V_1$ durations were set to 50 ms. The original vowel durations of /i/ and /u/ were 60 ms and 50 ms.
respectively. It was considered that a speech sound could become unnatural by lengthening the vowel duration. Hence, the vowels were set to the shorter duration, which was the original duration of /u/.

2.2.2. Synthesis of C

White-noise was filtered by using MATLAB to synthesise sound. The bandpass filter was used for filtering, and FIR filter was designed by windowing. The number of FIR taps was 2,048. Figures 1, 2 and 3 show the frequency diagrams in case that the frequency was the same until 110 ms as spectrograms in Fig. 4 and 2) the frequency was set to the following two patterns: 1) the frequency was kept in the same F\textsubscript{L} until 165 ms as spectrograms in Fig. 4 and 2) the frequency falls in the middle. Figures 4 and 5 show spectrograms in F\textsubscript{L} ‘L/ M/ H’ for the two conditions, F\textsubscript{L} non-falling and F\textsubscript{L} falling, schematically. In the pattern 2), the frequency was the same F\textsubscript{L} until 110 ms, moreover F\textsubscript{L} between 110 ms and 165 ms falls from F\textsubscript{L} ‘L/ M/ H’ to 480 Hz as shown in Fig. 5, which shows the diagram of spectrogram for the synthesised sounds with F\textsubscript{L} ‘L’/ M’/ H’ as the F\textsubscript{L} falls towards the end of /ʃ/.

The pattern 1) is F\textsubscript{L} non-falling condition and referred to as S0 since the slope of F\textsubscript{L} is flat in this condition. The pattern 2) is F\textsubscript{L} falling condition and referred to as S1 since there is slopes of F\textsubscript{L} in case that the F\textsubscript{L} falls.

2.2.3. Creating devoiced/non-devoiced condition

Table 1 shows the combination of the target syllables using each F\textsubscript{L} in devoiced/non-devoiced conditions and characteristics which was used to synthesise fricative sound with MATLAB. The transition region of the upper limit frequency in the synthesised sound was set to 7,440–7,920 Hz. Beckman et al. [23] has reported that the spectral peak affects the discrimination of /ʃ/ with devoiced vowels, and the participants identify as /ʃ(u)/ as the spectra peak is low. Li et al. [21] also reported the lower frequency region affects the perception of fricative sounds. Therefore, the region of the lower limit frequency has three types in this experiment. Moreover, it has known that the spectral peaks of /s/ and /ʃ/ are around 5,000–8,000 Hz and 2,500–5,000 Hz, respectively [25]. The three types, 4,080–4,800 Hz, 2,760–4,800 Hz and 1,440–4,800 Hz were chosen, which are judged as natural fricative sounds by authors, in this experiment (see Fig. 4 which shows the diagram for the synthesised sounds with each F\textsubscript{L}).

We refer to the lower frequency of the transition region at the lower limit frequency as F\textsubscript{L}, moreover, each F\textsubscript{L} 4,080 Hz, 2,760 Hz and 1,440 Hz are called ‘H,’ ‘M’ and ‘L,’ respectively.

Matsui [24] reported that it is possible to predict the following vowel by using the peak frequency transition at the end of /ʃ/ even when the vowels are devoiced, moreover /u/ was perceived as the following vowel in case that the vowels are devoiced.

Table 1 The stimuli used in Experiment 1. H, M and L show that F\textsubscript{L} are 4,080 Hz, 2,760 Hz and 1,440 Hz, respectively. S0 and S1 show that F\textsubscript{L} non-falling and F\textsubscript{L} falling condition, respectively. D0 and D1 show non-devoiced and devoiced, respectively. Ø means that there are no vowels.

| \(F_L\) non-falling condition | \(F_L\) falling condition |
|-----------------------------|---------------------------|
| Non-devoiced                | Devoiced                  | Non-devoiced               | Devoiced                  |
| stimuli                     | \(V_1\)                   | stimuli                     | \(V_1\)                   |
| Non-devoiced stimuli        | \(V_1\)                   | Non-devoiced stimuli        | \(V_1\)                   |
| 4,080 Hz                    | H\textsubscript{S0,D0}    | /u/                         | H\textsubscript{S0,D1}    | Ø                          |
| 2,760 Hz                    | M\textsubscript{S0,D0}    | /u/                         | M\textsubscript{S0,D1}    | Ø                          |
| 1,440 Hz                    | L\textsubscript{S0,D0}    | /u/                         | L\textsubscript{S0,D1}    | Ø                          |
| 4,080 Hz                    | H\textsubscript{S1,D0}    | /u/                         | H\textsubscript{S1,D1}    | Ø                          |
| 2,760 Hz                    | M\textsubscript{S1,D0}    | /u/                         | M\textsubscript{S1,D1}    | Ø                          |
| 1,440 Hz                    | L\textsubscript{S1,D0}    | /u/                         | L\textsubscript{S1,D1}    | Ø                          |
FL falling (S1)/non-falling (S0) conditions. Moreover, the vowels used as V₁ in each stimulus is shown in the cell next to each stimulus.

Stimuli with two patterns, devoiced and non-devoiced, were created by using the speech sound materials, whose C₁ of the recorded target word changed to the synthesised fricative. There is uttered V₁ following C₁ in the materials with C₁ which is changed to synthesised fricative sounds, since every target word was recorded with non-devoiced vowels. The materials with uttered V₁ following C₁ is the stimuli in non-devoiced condition. On the other hand, the material which we deleted the uttered V₁ from the material in devoiced condition were devoiced condition.

The vowel /u/ was used as V₁ of non-devoiced pattern in FL falling condition since the speech sound becomes unnatural if /i/ is used. Whereas, /i/ was used for the stimuli whose FL were 4,080 Hz since the speech sound becomes unnatural sound in case that /u/ is used for V₁ following the high frequency of the frication. /i/ and /u/ were used for V₁ when FL of the stimuli was 2,760 Hz, and /u/ was used for the stimulus whose FL was 1,440 Hz since the speech sound also becomes unnatural sound in case that /i/ is used for V₁ following the low frequency of the frication. The possibility could be /u/ and /i/ are following to the fricative sound with high and low frequency component, respectively, for the fact that the spectral peak is 100 Hz–300 Hz higher in case that sibilant sounds are before /i/ than before /a/ and /u/. Therefore, we did not use the combination of the stimuli which could be unnatural speech sound in this experiment. Whereas, both /i/ and /u/ were used as FL is 2,760 Hz since this frequency did not make the speech sound unnatural in either case that /i/ or /u/ was followed. However, the vowel /i/ was used as a filler, therefore /i/ was excluded from the analysis in the stimuli with FL of 2,760 Hz. That is, we analysed how much the participants perceive as /u/ when the vowel /u/ was used for the stimuli in devoiced condition, moreover, how they perceive the stimuli when the vowel was devoiced.

‘D’ of each stimulus in captions of Table 1 shows if than for S₁,D₀, as noted Fig. 12. The vowel is devoiced or non-devoiced, hence, D₁ and D₀ mean devoiced and non-devoiced condition, respectively.

Figures 6 and 7 show the spectrograms of the stimuli with FL falling (S1) in devoiced (D1)/non-devoiced (D0) conditions as is 2,780 Hz.

2.2.4. Creating continuum
Continuum was created by changing the closure duration of C₂, the stop sound /k/, in two stimuli groups with devoiced and non-devoiced vowels with each FL to investigate the perception of geminate too. The closure duration of /k/ in devoiced/non-devoiced conditions was set to 5 steps (each 50 ms), from 100 ms to 300 ms, as the continua were created as shown in Figs. 6 and 7. The consonant, vowel and closure durations are shown in these two figures.

3. EXPERIMENT

3.1. Participants
10 young normal-hearing listeners (M 3, F 7) participated in this experiment. Their ages ranged from 18–25 years old with a mean of 21.1. All participants were native speakers of Japanese with normal hearing. Nine participants were from Great Tokyo area and one participant was from Hokkaido. That is, all participants were from regions with dialect that the vowel devoicing occurs.

3.2. Pre-experiment Tests
Pure tone audiometry test and gap detection test were conducted for the all 10 participants as pre-experiments.
3.2.1. Pure tone audiometry test

Seven pure tones with frequencies between 125 Hz and 8,000 Hz were presented intermittently to the all participants. The participants push a button while they can hear a sound, and stop pushing the button as they realise the sound has stopped. Figure 8 shows the audiogram of 10 participants. The average of absolute threshold of hearing with a good ear was 4.5 dB (HL). The average shows that the participants have normal hearings since their average was under 20 dB.

3.2.2. Gap detection test

Two types of white-noise were presented with in this gap detection test. One of the stimuli was a continuous white-noise and the other stimulus was a white-noise with a gap in the middle [26]. The first stimulus presented to the participants had a 100 ms gap, and the length of the gap was shortened form 100 ms by the two-up one-down method. The participants an either sound with a gap by clicking a button on the PC. The sounds were presented over headphones (Sennheiser HDA200) at an A-weighted sound pressure level of 70 dB as measured on a sound level meter (ONO SOKKI LA-5111) connected to an artificial ear (Brüel & Kjær Type 4153). The mean of the just noticeable difference (namely JND) of the gap detection test was 3.8 ms.

3.3. Experimental Procedure

We used 65 stimuli which is combined the following variability in this experiment; FL falling slope (falling (S1) and non-falling (S0)), FL (1,440 Hz (L), 2,760 Hz (M), and 4,080 Hz (H)), voicing (devoiced (D1) and non-devoiced (D0)), and closure duration (100–300 ms (five steps)). The total number of trials was 325 since all stimuli were repeated five times. The stimuli inserted into the single carrier sentence recorded in Sect. 2.1 were presented randomly to all of the participants by using Praat [27]. The experiment was conducted in a sound proof room at a university in Tokyo.

The stimuli were presented over headphones (Sennheiser HDA 200) through a USB sound adopter (Roland, UA-25EX) at an A-weighted sound pressure level of 70 dB, as measured in a sound level mater and with an artificial ear (Brüel & Kjær Type 4153). The participants chose one word which they heard from four words, ‘shukan,’ ‘shukkun,’ ‘shikan,’ and ‘shikkan’ by clicking a button with each word in Japanese orthography (kana) on PC screen. They took a break every 65 stimuli which were randomised in each block.

3.4. Results

3.4.1. Response rate of /ju/

Figure 9 shows the response rate which the participants responded as /ju/ (referred to as response rate of /ju/) for stimuli with each $F_L$ in both devoiced/non-devoiced conditions. The response rate of /ju/ for M,S0 and L,S0 in devoiced condition is 19% and 80%, respectively. Therefore, these response rates are lower than in non-devoiced condition (99% and 100%, respectively), as noted this figure. Figure 10 shows the response rate of /ju/ and the response rate which the participants responded as /i/ (referred to as response rate of /i/) for the stimuli with devoiced vowels in $F_L$ falling/non-falling conditions. The both response rates of /ju/ and /i/ are shown in Fig. 10 since one of the purposes in this experiment was that we investigate which the participants perceive /ju/ or /i/ in both $F_L$ falling/non-falling conditions, especially with devoiced vowels, on this experiment. Moreover, the response rate of /ju/ for the stimuli in both $F_L$ falling/non-falling conditions is compared in Fig. 11. For this figure, it can be seen that the response rate of /ju/ for H,S0 is 2% and the lowest, and the response rate of /ju/ for H,S1 is 89%, that is, it is the highest.

The three-way factorial within-subject ANOVA was conducted on influence of three independent variables (voicing of vowels (voicing), falling of vowels (falling slope), and $F_L$ itself ($F_L$)) on the response rate of /ju/. Voicing included two levels, devoiced/non-devoiced. Falling included two levels, $F_L$ falling slope/$F_L$ non-falling slope. $F_L$ included three levels, H, M and L.

The main effect for all factors, voicing, falling slope and $F_L$ were statistically significant ($F(1,9) = 23.626; p < 0.001$, $F(1,9) = 1044.414; p < 0.001$, $F(1,9) = 105.118, p < 0.001$). The interactions between voicing and falling slope, voicing and $F_L$, falling slope and $F_L$ were significant ($F(1,9) = 80.751; p < 0.001$, $F(2,18) = 245.801; p < 0.001$, $F(2,18) = 63.056; p < 0.001$), and two-way interaction was significant ($F(2,18) = 63.056; p < 0.001$).
The simple interaction effect test was conducted since two-way interaction was statistically significant. The voicing effect and the falling slope effect for \( F_L \) ‘H’ and ‘M’ were significant \((F(1,27) = 5.983; \ p = 0.021, F(1,27) = 193.540; \ p < 0.001)\). The voicing effect and the \( F_L \) effect for \( F_L \) falling slope, and the \( F_L \) falling slope effect and the \( F_L \) effect for both devoiced/non-devoiced were statistically significant, respectively \((F(2,36) = 65.285; \ p < 0.000, F(2,36) = 115.083; \ p < 0.001, F(2,36) = 196.432; \ p < 0.001)\). Furthermore, we conducted simple interaction effect test for two-way interaction. The simple main effect test was also conducted. The voicing effects for the combination of the \( F_L \) non-falling slope and the \( F_L \) ‘M’/‘L,’ and the \( F_L \) effect for the combination of the devoicing and the \( F_L \) ‘H/M’ were significant \((F(1,54) = 158.416; \ p < 0.001, F(1,54) = 9.901; \ p = 0.003, F(1,54) = 511.360; \ p < 0.001, F(1,54) = 365.185)\). In addition, the \( F_L \) effect for the non-devoiced and \( F_L \) ‘H’ were statistically significant; \((p < 0.001, F(1,54) = 676.550; \ p < 0.00)\). The \( F_L \) effect for the devoicing and the \( F_L \) non-falling slope, and the non-devoicing and the \( F_L \) non-falling slope were also significant \((F(1,72) = 122.444; \ p < 0.001, F(1,54) = 238.776; \ p < 0.0001)\).

In other words, the results show that the simple main effect of the voicing for M/L/S1 was statistically significant, however, the simple main effect of the voicing for H/S0 was non-significant. The results also show that the simple main effect of the \( F_L \) falling slope for each \( F_L \) was significant, however, the simple main effect of the \( F_L \) non-falling slope for each \( F_L \) was non-significant. Moreover, the simple main effect of the falling slope for \( F_L \) ‘H/M’ was significant, however, the simple main effect of the falling slope for \( F_L \) ‘L’ was non-significant.

We conducted multiple comparison by using Ryan method since the simple main effect was significant. The result showed that there were significant differences between the \( F_L \) ‘H’ and ‘L,’ \( F_L \) ‘M’ and ‘L,’ and \( F_L \) ‘H’ and ‘M.’

3.4.2. Response rate of geminate

Fig. 12 shows average of response rate of geminate for combinations of each stimulus with each closure duration \((F_L \) falling slope + devoiced (referred to as S1D1), \( F_L \) non-falling slope + devoiced (referred to as S0D1), \( F_L \) falling slope + non-devoiced (referred to as S1D0), \( F_L \) non-falling slope + non-devoiced (referred to as S0D0)). The y axis shows devoiced/non-devoiced vowels in both \( F_L \) falling/non-falling conditions are compared in Fig. 12, respectively. The three-way factorial within-subject ANOVA was conducted on influence of three independent variables, voicing of vowels (voicing), falling of \( F_L \) (falling slope), and closure duration, on the response rate of geminate. Voicing included two levels (devoiced/non-devoiced), falling slope included two levels \((F_L \) falling slope, \( F_L \) non-falling slope), and closure duration included five levels (100 ms, 150 ms, 200 ms, 25 ms, and 300 ms).
The main effects for the falling slope and the closure duration were statistically significant \((F(1, 9) = 8.542; \ p = 0.002, F(4, 36) = 145.548; \ p < 0.001)\). The interactions between the voicing and the closure duration, and the \(F_L\) and the closure duration were significant \((F(4, 36) = 4.107; \ p = 0.001, F(4, 36) = 3.986; \ p = 0.001)\), and the two-way interaction was also significant \((F(3, 36) = 3.749; \ p = 0.011)\).

The simple interaction effect test was conducted as two-way interaction was significant. The voicing effect and the \(F_L\) effect for the closure duration of 200 ms, the voicing effect and the closure duration effect for the \(F_L\) falling slope, and the falling slope effect and the closure duration effect for the devoiced \((F(1, 45) = 15.491; \ p < 0.001, F(4, 72) = 5.875; \ p < 0.001, F(4, 72) = 7.368; \ p < 0.001)\). We also conducted the simple main effect test.

The simple main effect of the voicing for the combination of the \(F_L\) falling slope and the closure duration of 200 ms/250 ms were qualified \((F(1, 90) = 16.988; \ p < 0.001)\). Moreover, the simple main effect of the voicing for the combination of the \(F_L\) non-falling slope and the closure duration of 250 ms were statistically significant \((F(1, 90) = 9.136; \ p < 0.003)\). The simple main effect of the \(F_L\) for the combination of the devoiced and closure duration of 200 ms, and the simple main effects of the closure duration for the combination of the devoiced and the \(F_L\) falling slope/non-falling slope and the combination for the non-devoiced and the \(F_L\) falling slope/non-falling slope were also significant \((F(1, 90) = 38.239; \ p < 0.001, F(4, 144) = 85.279; \ p < 0.001, F(4, 144) = 83.243; \ p < 0.001, F(4, 144) = 84.974; \ p < 0.001, F(4, 144) = 93.579; \ p < 0.001)\).

In other words, these results show that the simple effect and the simple main effect of the voicing for the \(F_L\) falling slope is statistically significant, and the simple effect and the simple main effect of the falling slope for the devoiced is also qualified. It can be seen that the response rate of geminate for \(S_1\) with closure duration of 200 ms and 250 ms was lower for \(S_1\) for Fig. 12. Furthermore, the response rate of geminate for \(S_1\) the falling slope the with closure duration of 200 ms decreased than for \(S_0\) for, as noted Fig. 12 and the FL effect for the closure duration of 200 ms, the voicing effect and the closure duration effect for the FL falling slope, and the falling slope effect and the closure duration effect for the devoiced \((F(1, 45) = 15.491; \ p < 0.001, F(4, 72) = 5.875; \ p < 0.001, F(4, 72) = 7.368; \ p < 0.001)\). We also conducted the simple main effect test.

The simple main effect of the voicing for the combination of the \(F_L\) falling slope and the closure duration of 200 ms/250 ms were qualified \((F(1, 90) = 16.988; \ p < 0.001)\). Moreover, the simple main effect of the voicing for the combination of the \(F_L\) non-falling slope and the closure duration of 250 ms were statistically significant \((F(1, 90) = 9.136; \ p < 0.003)\). The simple main effect of the \(F_L\) for the combination of the devoiced and closure duration of 200 ms, and the simple main effects of the closure duration for the combination of the devoiced and the \(F_L\) falling slope/non-falling slope and the combination for the non-devoiced and the \(F_L\) falling slope/non-falling slope were also significant \((F(1, 90) = 38.239; \ p < 0.001, F(4, 144) = 85.279; \ p < 0.001, F(4, 144) = 83.243; \ p < 0.001, F(4, 144) = 84.974; \ p < 0.001, F(4, 144) = 93.579; \ p < 0.001)\).

In other words, these results show that the simple effect and the simple main effect of the voicing for the \(F_L\) falling slope is statistically significant, and the simple effect and the simple main effect of the falling slope for the devoiced is also qualified. It can be seen that the response rate of geminate for \(S_1\) with closure duration of 200 ms and 250 ms was lower for \(S_1\) for Fig. 12. Furthermore, the response rate of geminate for \(S_1\) the with closure duration of 200 ms decreased than for \(S_0\) for, as noted Fig. 12 and the FL effect for the closure duration of 200 ms, the voicing effect and the closure duration effect for the FL falling slope, and the falling slope effect and the closure duration effect for the devoiced \((F(1, 45) = 15.491; \ p < 0.001, F(4, 72) = 5.875; \ p < 0.001, F(4, 72) = 7.368; \ p < 0.001)\). We also conducted the simple main effect test.

4. DISCUSSION

4.1. The Response Rate of /fu/

It was found that young normal-hearing listeners could use \(F_L\) falling as a cue to distinguish /fi/ from /fu/ for Fig. 9. The results showed that there were no significant differences between \(L/M/H_D1S1\) and \(L/M/H_D0S1\), that is, the response rates of /fu/ for the stimuli were high in both conditions. On the other hand, a tendency was
founded that the response rate of /ju/ for M_S0 and L_S0 was low in devoiced condition although the response rate was high in non-devoiced condition. However, there was no significant difference between devoiced and non-devoiced conditions in H_S0. In H_S0, the response rate of /ji/ was almost 100% since /i/ was used as the following vowel of /j/ (see Sect. 2.2.3). These results show that the response rate of /ju/ for H_S1 was high even when the vowels were devoiced although the response rate for the H_S0 was low as the vowels were devoiced. Hence, these results imply that F_L falling affects to discriminate /ju/ from /ji/ in devoiced condition by young normal-hearing listeners.

It has been reported that the peak frequency transition affects to discriminate /ji/ from /ju/ by young normal-hearing listeners regardless of spectra peaks of frequency by Matsui [24]. In Fig. 10, we compared the response rate of /ju/ and /ji/ for both S1_D1 and S0_D1. For the result, the response rates of /ju/ for H_S1_D1 and M_S1_D1 were high, and there were no significant differences between the response rate L_S1_D1 and L_S0_D1. However, it was observed that the response rates of /ji/ and /ju/ of the both stimuli were high. Therefore, the result of this experiment suggested that young normal-hearing listeners perceive /ju/ in F_L falling condition even when vowels were devoiced. Hence, the results of this experiment supported the results in Matsui [24].

Moreover, the response rate of /ju/ for L_S0 was significantly higher than H/M_S0. The following reasons were considerable for this result. Beckman [23] reported that the spectral peak of /f(i)/ is higher than /f(u)/ when the vowels are devoiced, moreover, young normal-hearing listeners expect that the following vowel is /u/ for /f(u)/ with devoiced vowel as the spectral peak was low. F_L ‘L,’ 1,440 Hz, which is used in this experience was the lowest frequency in F_L. Furthermore, the frequency band of falling part of L_S1 was between 1,440 Hz and 480 Hz, that is, the slope of the F_L falling was gentle. Therefore, the possibility was assumed that it was hard for the participants to recognise that F_L was falling in case of F_L ‘L,’ thus the response rate of /ju/ for L_S0 was high since the frequency of ‘L’ was low.

These results of this experiment suggested that F_L falling was used as one of cues to distinguish /f(i)/ from /f(u)/, moreover F_L itself affects the discrimination of /f(i)/ from /f(u)/ since it is possible for young normal-hearing listeners to distinguish the difference of frequencies.

### 4.2. The Response Rate of Geminate

We compared the response rate of geminate for S1/S0_D1/D0 in Fig. 12, and the results showed the tendencies about the response rate. The tendency was observed that each response rate of geminate for the stimuli with devoiced vowels was almost same as the response rates for the stimuli with non-devoiced vowels for Fig. 12. Therefore, this result implies it is possible for young normal-hearing listeners to distinguish geminate from non-geminate with /j(i)/ and /j(u)/ even when vowels are devoiced.

However, the response rate of geminate for S1_D1 with closure duration of especially 200 ms decreased, and the overall response rate for the stimuli slightly shifted toward longer closure duration, as noted Fig. 12. It had been reported that young normal-hearing listeners expected that the following vowel was /u/ when they perceive fricative /j/ with the peak frequency transition by Matsui [24]. It can be clearly seen that S1_D1 were perceived as /j(u)/ for the result of this experiment as well. Therefore, it assumed that young normal-hearing listeners perceive F_L falling as the following vowel ‘/u/.’

Assuming the part of the falling of F_L is a vowel (referred to as F_L vowel here), the following possibility was considered. We compared the F_L vowel and the vowel in non-devoiced condition (referred to as non-devoiced vowel here). The duration of the non-devoiced vowel of the stimuli was 50 ms in this experiment. Whereas the duration of the falling of F_L was 55 ms since F_L falls between 110 ms and 165 ms in case of S1_D1, that is, the duration of F_L vowel was 55 ms. However, it might be suspected that the timing is ambiguous when the participants start to recognise the F_L vowel for following reason; non-devoiced vowel has the maximum energy of the vowel, whereas the F_L vowel has not the energy of vowel like non-devoiced vowel since the F_L vowel is the part of the frication. Therefore, it is possible that the duration of F_L vowel was perceived shorter than the duration of non-devoiced vowel if the participants recognise the F_L vowel is the middle or the end of F_L falling. It is known that the duration of preceding vowels is one cue of the perception of geminate. However, it is possible to be hard to perceive geminate in case that the judge of the F_L vowel’s duration was ambiguous and the duration was perceived shorter. Hence, it is assumed that the response rate of geminate for the stimuli with ambiguous closure duration decreased.

Furthermore, a slight difference is observed between the perception rate of geminate for S0_D1 and other stimuli, S0/S1_D0, with closure duration of 200 ms although there is no significant difference between them. Following reasons are considerable for these results; it is possible that the energy of vowel could be a cue to judge the boundary between the vowel and the closure duration because the energy of vowels is located in preceding the closure duration. On the other hand, the energy of the vowels is lost in devoiced condition, that is, the cue to
judge the boundary between the fricative and the closure durations are lost. Hence, it is implied that the response rate of geminate was ambiguous since it was hard to judge the boundary as there was only the energy of frication before the closure duration.

We discussed the possibilities of the response rates for two stimuli, S1_D1 and S0_D1, above. However, there is one more possibility for the both of the response rates. Extra closure durations for the cut vowel duration (50 ms) was not added as we generate the stimuli in devoiced condition, and the vowels were just cut (see Sect. 2.2.3). (There was no choice to lengthen C1 duration for the vowel duration since the C1 duration was not different between devoiced and non-devoiced conditions as the duration of the recorded original speech sound were measured.) That is, the total durations in devoiced condition is 50 ms shorter than in non-devoiced condition. Therefore, there is a possibility that the difference of the durations between devoiced and non-devoiced condition affects their perception of geminate.

However, it has not clarified how overall duration in devoiced condition should be same as the duration in non-devoiced condition. The perception boundary of geminate for S1_D1 could be same as the boundaries for the other stimuli if 50 ms closure duration is added to each closure duration used in this experiment (100–300 ms) of the stimuli in devoiced condition so that the overall duration in devoiced condition is same duration as non-devoiced duration. This possibility described above is also predicted for the line graph ‘S1_D1’ of Fig. 12 by shifting the line graph 50 ms toward the right.

On the other hands, it is possible that the boundary for S0_D1 would be shifted toward shorter closure duration in case that we use the way described above to length the total duration since the boundary was almost same as the boundary for the stimuli in non-devoiced condition on this result. The way is also possible that the 50 ms frication is added to the C1 duration used in this experiment. However, the original C1 durations uttered with devoiced and non-devoiced vowels were almost same as described above. Hence, we need to find how we should add extra 50 ms duration, and compare the response rates of geminate in each condition on this experiment and the response rates in case that the extra closure durations are added for the cut vowel duration for the further work.

5. CONCLUSION

In this paper, we showed that the results of the perception experiment for young normal-hearing listeners to investigate cues used to distinguish /f(u)/ from /f(i)/ as the vowels are devoiced. The tendency was observed that the response rate of /f(u)/ for S1_D1 higher than the response rate for S0_D1, and that the response rates in both devoiced/non-devoiced conditions were high as FL was 1,440 Hz regardless of F1 falling. Hence, it was suggested that F1 falling and F1 itself affect to identify /f(u)/ from /f(i)/ even when vowels are devoiced by young normal-hearing listeners, moreover these results supports Matsui’s suggestion that young normal hearing listeners are able to discriminate /f(u)/ from /f(i)/ as the peak frequency transition even when the vowels.

We also investigated the perception of geminate with /f(u)/ and /f(i)/ by young normal-hearing listeners. The tendency was observed that the response rates of geminate with devoiced vowels were almost same as the response rates with non-devoiced vowels, suggesting that young normal-hearing listeners were able to distinguish geminate from non-geminate with /f(u)/ and /f(i)/ even if the vowels are devoiced. However, it was observed that the response rate of geminate decreased as the closure duration was 200 ms, and the perception boundary of geminate for S1_D1 shifted toward longer closure duration. The tendency was also observed that the response rate for S0_D1 decreased as the closure duration was ambiguous. Therefore, this result implied that it was hard to distinguish fricative sounds and closure durations since the energy of vowels was lost and the energy of frication was left in devoiced condition although the energy of vowels could be used as the cue to judge the boundary between fricative duration and closure duration.

As the introduction, the elderly listeners misperceive /f(u)/ as /f(i)/ in both geminate and non-geminate words as the vowels are devoiced, whereas the young normal-hearing listeners was able to identify them regardless of devoiced vowels in the previous study, and the further investigation was required. For the results of this experiment for young normal-hearing listeners, it was possible that young normal hearing listeners were able to distinguish /f(u)/ from /f(i)/ correctly even when the vowels were devoiced since they use F1 falling and F1 itself as cues to identify them.

In the next paper, we will present the results of the perception experiment for elderly listeners, which is another experiment of this project. We will compare the results of the experiments for young normal-hearing and elderly in the paper. Moreover, we will analyse the relationship between the misperception of /f(u)/ and /f(i)/ by elderly listeners and the cues to perceive /f(u)/ and /f(i)/ which they use.

ACKNOWLEDGEMENT

We would like to thank Professor Akira Kawanaka and Professor Mamiko Fujii, and Doctor Keiich Yasu, Doctor Kei Kobayashi, who gave us advices for our research.
REFERENCES

[1] E. Dupoux, K. Kakehi, Y. Hirose and C. Pallier, “Epenthetic vowels in Japanese: A perceptual illusion?,” J. Exp. Psychol.: Hum. Percept. Perform., 25, 1568–1578 (1999).

[2] T. J. Vance, An Introduction to Japanese Phonology (SUNY Press, New York, 1987).

[3] Y. Takubo, K. Maekawa, H. Kubozono, K. Homma, K. Shirai and S. Nakagawa, Onsei (Iwanami Shoten, Tokyo, 2004).

[4] E. IWAGAMI et al., /shu/–/shi/ with devoiced vowels by young listeners, J. Acoust. Soc. Am., 132, 2663–2675 (2012).

[5] M. Ohso, “A phonological study of some English loan words in Japanese,” Ohio State Univ. Work. Pap. Linguistics, 14, 1–26 (1973).

[6] M. Beckman and S. Atsuko, “Spectral and perceptual evidence for CV coarticulation in devoiced /si/ and /syu/ in Japanese,” Phonetica, 41, 61–71 (1984).

[7] F. Li, A. Trevino, A. Menon and J. B. Allen, “A psychoacoustic method for studying the necessary and sufficient perceptual cues of fricative consonants in noise,” J. Acoust. Soc. Am., 85, 1718–1725 (1988).

[8] S. Behrens and S. E. Blumstein, “On the role of the amplitude of the fricative noise in the perception of place of articulation in voiceless fricative consonants,” J. Acoust. Soc. Am., 84, 861–867 (1988).

[9] C. H. Shadle and S. J. Mair, “Quantifying spectral characteristics of fricatives,” Proc. Int. Conf. Spoken Language Processing (ICSLP), pp. 1521–1524 (1996).

[10] L. Faber, “Lip protrusion in sibilant production,” J. Acoust. Soc. Am., 85(S1), S113 (1989).

[11] P. Ladevoged and I. Maddieson, The Sounds of the World’s Languages (Blackwell, Oxford, 1996).

[12] S. D. Soli, “Second formants in fricatives: Acoustic consequences of fricative-vowel coarticulation,” J. Acoust. Soc. Am., 70, 976–984 (1981).

[13] S. Narayanan, A. A. Alwan and K. Haker, “An articulatory study of fricative consonants using magnetic resonance imaging,” J. Acoust. Soc. Am., 98, 1325–1347 (1995).

[14] M. Toda, S. Maeda, A. J. Carlen and L. Meftahi, “Lip gestures in English sibilants: Articulatory–acoustic relationship,” Proc. 7th Int. Conf. Spoken Language Processing (ICSLP), pp. 2165–2168 (2002).

[15] J. Heinz and K. Stevens, “On the properties of voiceless fricative consonants,” J. Acoust. Soc. Am., 33, 589–596 (1961).

[16] G. A. Miller and P. E. Nicely, “An analysis of perceptual confusions among some English consonants,” J. Acoust. Soc. Am., 27, 338–352 (1955).

[17] J. Proctor, C. Shadle and K. Iskarous, “An MRI study of fricative consonants using magnetic resonance imaging,” J. Showa Med. Assoc., 27, 338–352 (1955).

[18] F. Li, A. Menon and J. B. Allen, “A psychoacoustic method to find the perceptual cues of stop consonants in natural speech,” J. Acoust. Soc. Am., 127, 2599–2610 (2010).

[19] S. Nittrouer, “Age-related differences in perceptual effects of formant transitions within syllables and across syllable boundaries,” J. Phon., 20, 351–382 (1992).

[20] F. Li, A. Trevino, A. Menon and J. B. Allen, “A psychoacoustic method to find the perceptual cues of stop consonants in natural speech,” J. Acoust. Soc. Am., 127, 2599–2610 (2010).

APPENDIX

Fig. A-1 This figure shows the spectrum of the original speech sound of /ʃ(ə)/ in ‘shukan’ and a piecewise-linear approximation of this spectrum. The synthesised sound in case that $F_t$ is 2,760 Hz (M) in Fig. 2 was synthesised on the basis of the piecewise-linear approximation of the original speech sound. The synthesised sounds L and H were formed on the basis of the synthesised sound M to compare three types of lower limit frequencies.
Eri Iwagami received a M.A. in English Literature from Hosei University, Tokyo, Japan, in 2015, and a Ph.D. in Science and Technology in 2018 from Sophia University, Tokyo, Japan. She is currently a postdoctoral researcher in the Faculty of Science and Technology at Sophia University, Tokyo, Japan. Her research interests include acoustic and experimental phonetics and speech and hearing sciences.

Takayuki Arai received the B.E., M.E. and Ph.D. degrees in electrical engineering from Sophia Univ., Tokyo, Japan, in 1989, 1991 and 1994, respectively. In 1992–1993 and 1995–1996, he was with Oregon Graduate Institute of Science and Technology (Portland, OR, USA). In 1997–1998, he was with International Computer Science Institute (Berkeley, CA, USA). He is currently Professor of the Department of Information and Communication Sciences, Sophia Univ. In 2003–2004, he is a visiting scientist at Massachusetts Institute of Technology (Cambridge, MA, USA). His research interests include signal processing, acoustics, speech and hearing sciences, spoken language processing, and acoustic phonetics.

Keiichi Tajima received a B.A. in Linguistics and Computer Science in 1991 from Brown University, Providence, Rhode Island, USA, an M.A. in Linguistics in 1995, and a Ph.D. in Linguistics and Cognitive Science in 1998, both from Indiana University, Bloomington, Indiana, USA. After working as a visiting researcher at ATR Laboratories, Kyoto, Japan, he started teaching at Hosei University, Tokyo, Japan in 2003. He is currently professor in the Department of Psychology at Hosei University. In 2012–2013, he was a visiting scholar in the Speech Communication Group in the Research Laboratory of Electronics at MIT, Cambridge, Massachusetts, USA. His research interests include acoustic and experimental phonetics, phonology, speech production and perception, and second-language learning.