



Syllable sequence of /a+/ta/ can be heard as /atta/ in Japanese with visual or tactile cues

Takayuki Arai¹, Miho Yamada^{1,2}, Megumi Okusawa¹

¹Sophia University (Tokyo, Japan)

²Pannasonic (Japan)

arai@sophia.ac.jp

Abstract

In our previous work, we reported that the word /atta/ with a geminate consonant differs from the syllable sequence /a/+pause+/ta/ in Japanese; specifically, there are formant transitions at the end of the first syllable in /atta/ but not in /a/+pause+/ta/. We also showed that native Japanese speakers perceived /atta/ when a facial video of /atta/ was synchronously played with an audio signal of /a/+pause+/ta/. In that study, we utilized two video clips for the two utterances in which the speaker was asked to control only the timing of the articulatory closing. In that case, there was no guarantee that the videos would be the exactly same except for the timing. Therefore, in the current study, we use a physical model of the human vocal tract with a miniature robot hand unit to achieve articulatory movements for visual cues. We also provide tactile cues to the listener's finger because we want to test whether cues of another modality affect this perception in the same framework. Our findings showed that when either visual or tactile cues were presented with an audio stimulus, listeners more frequently responded that they heard /atta/ compared to audio-only presentations.

Index Terms: speech perception, multimodality, speech recognition, geminate consonant, visual/tactile cue

1. Introduction

Japanese has distinctions of certain geminate consonants from singletons. For example, the word pair of “kata” and “katta” has different meanings (“form” and “bought” in English). The main difference between such word pairs is the length of the consonant [1, 2]. Our prior research has shown that a word with the geminate consonant /atta/ differs from the syllable sequence /a/+pause+/ta/ (/a+/ta/, hereafter) in Japanese [3]. Specifically, we found that formant transitions were observed at the end of the first syllable in /atta/ but not in /a+/ta/. When we compared speech recognition rates by Japanese native listeners between /atta/ and /a+/ta/ responses with exactly the same speech sounds (including the total and silence durations) and only changed whether or not the formant transitions at the end of the first vowel exist, the /atta/ response dramatically increased with such formant transitions. We also found in a subsequent study that Japanese native listeners perceived /atta/ when a facial video of /atta/ was synchronously played with an audio signal of /a+/ta/ [4]. This suggests that native listeners of Japanese integrate both auditory and visual information (e.g., McGurk effect [5]), and that the articulatory closing compensates for the formant transitions lacking in the auditory information.

However, in that study [4], we used two video clips for the two utterances of /a+/ta/ and /atta/. In this case, the speaker was asked to control only the timing of the articulatory closing and to keep the rest as similar as possible. Unfortunately, as we used a human speaker, there was no guarantee that the video would be ideal. so in the present study, we utilize a set of physical models of the human vocal tract. For visual cues, the upper and lower jaws were attached to a miniature robot hand unit to produce opening and closing movements in a synchronous manner with the audio signal. We tested whether visual cues compensated for the formant transitions that were lacking in the auditory information.

We want to test whether cues of another modality might also affect this perception in the same framework. The interactions between multimodal/multisensory information and speech production/perception have been investigated in other research (e.g., the effect of somatosensory feedback on speech production [6]). Somatosensory/tactile information also affects human speech perception. For example, Ito et al. [7] examined the effect of somatosensory stimuli against the skin around a listener's mouth on vowel perception. Gick and Derrick [8] tested the effect of aero-tactile stimuli against a listener's neck or hand on consonant perception. Therefore, we also examine whether tactile cues can compensate for the formant transitions.

2. Three types of stimuli

In this section, we describe the three types of stimuli utilized in this study: audio-only (A), audio + visual (A+V), and audio + tactile (A+T).

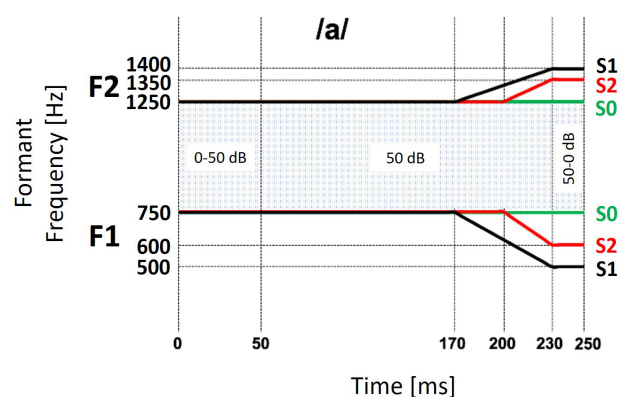


Figure 1: How F1 and F2 parameters change over time when synthesizing the first /a/ for audio signals.

2.1. Audio signals

Three audio signals were prepared: S0, S1, and S2. The simplest one, S0, is the vowel /a/ followed by the syllable /ta/ with a certain pause (denoted as /a+/ta/). The first and second vowels were both /a/ synthesized using XKL [9], while the consonant /t/ was taken from a natural utterance, as described in our previous study [4]. The frequencies of the first formant (F1) and second formant (F2) during the first /a/ were set as constant for the first vowel /a/ of S0, as shown in Fig. 1. The F1 frequency was 750 Hz and the F2 frequency was 1250 Hz.

Audio signals S1 and S2 are basically the same as S0, except for the formant transitions of F1 and F2 at the end of the first /a/. The time trajectories are also shown in Fig. 1. The only difference between S1 and S2 is the degree of the formant transitions, which were set to moderate in S2. Thus, the ending frequencies of F1 for S1 and S2 were 500 and 600 Hz, respectively, while those of F2 for S1 and S2 were 1400 and 1350 Hz. In all cases of S0, S1, S2, the gap between the first and the second syllables was 700 ms.

2.2. A+V stimuli

For the A+V stimuli, we used exactly the same audio signals discussed in Section 2.1. For the visual stimuli, we applied the physical models of the human vocal tract based on a previous study [10]. Figure 2 shows the setup used for the visual stimuli. The upper and lower jaws were attached to the robot hand unit (Vstone, Robovie-nano), and the open angle of the lower mechanism of the unit (i.e., the lower jaw) was achieved using a servo motor (Vstone, VS-S020A) controlled via Arduino.

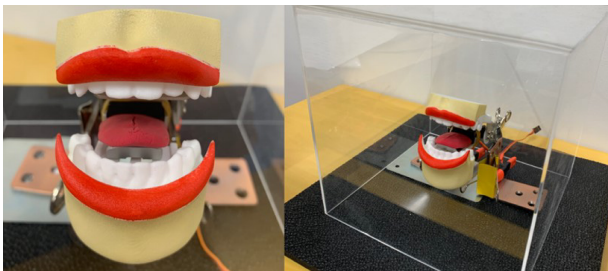


Figure 2: Physical models of human vocal tract used for visual stimuli. The upper and lower jaws were attached to the robot hand unit. The acrylic box (right) was used to reduce the noise level of motor sounds from the servo motor.

Finally, a Matlab script on a PC was programmed to synchronously play an audio signal and set the timing of the opening and closing of the vocal-tract model via Arduino.

We utilized two types of visual stimuli: VE and VL. The lower jaw was open at the beginning of each stimulus (A in Fig. 3). For the VE stimulus, the jaw started closing at the same time as the formant transition of the first vowel /a/ in S1 started (B in Fig. 3). For the VL stimulus, the jaw started closing 180 ms later than VE, i.e., during the closure of /t/ (C in Fig. 3). The lower jaw started opening when the burst of /t/ occurred (D in Fig. 3). Thus, the VE stimulus closes the jaw early (E), while the VL stimulus closes the jaw late (L).

The A+V stimuli were presented in an acrylic box with an anti-vibration pad, as shown on the right of Fig. 2. This prevented the participants from perceiving any motor sounds from the servo motor when it moved.

2.3. A+T stimuli

For the A+T stimuli, we also used the same audio signals discussed in Section 2.1. For the tactile stimuli, we utilized the same unit (Robovie-nano), and the angle of the opening/closing mechanism (i.e., the lower jaw) was achieved using the same servo motor (VS-S020A) controlled via Arduino, as was done for the A+V stimuli. A Matlab script on a PC was programmed to synchronously play an audio signal and set the timing of the opening and closing of the unit via Arduino. Figure 4 shows the unit with an acrylic box and anti-vibration pad. Participants inserted an index finger into a hole to place the fingertip between the upper and lower mechanisms.

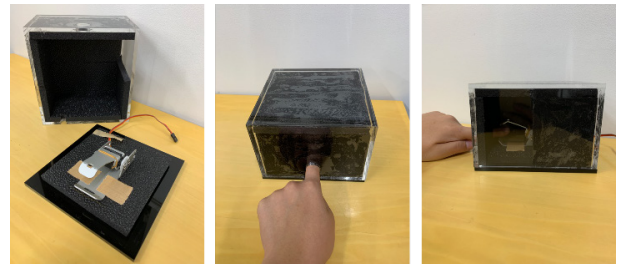


Figure 4: Same unit was used for tactile stimuli. Participants inserted an index finger into a hole of the acrylic box and placed the fingertip between the upper and lower mechanisms of the unit

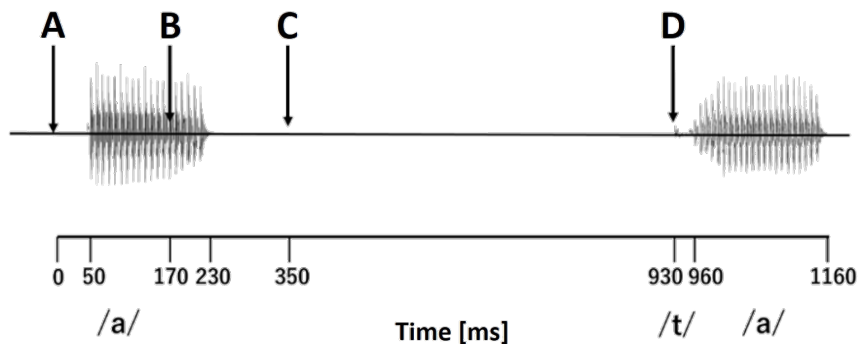


Figure 3: Waveform of audio signal along time axis. Point A: all of the visual/tactile stimuli are open at this point. Point B: early closing movements start to occur for VE and TE. Point C: late closing movements start to occur for VL and TL. Point D: all of the visual/tactile stimuli are open again at this point.

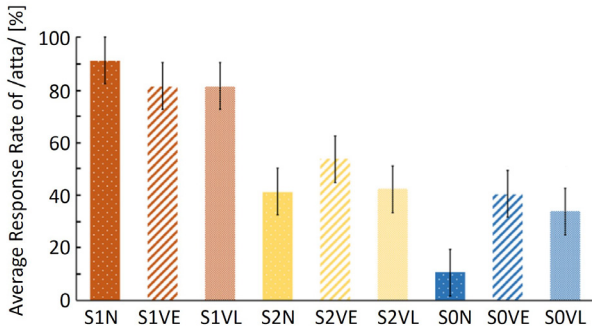


Figure 5: Response rate of /atta/ (%) under each stimulus condition in A+V session. All combinations of audio stimuli (S0, S1 and S2) and visual stimuli (VE and VL) are plotted.

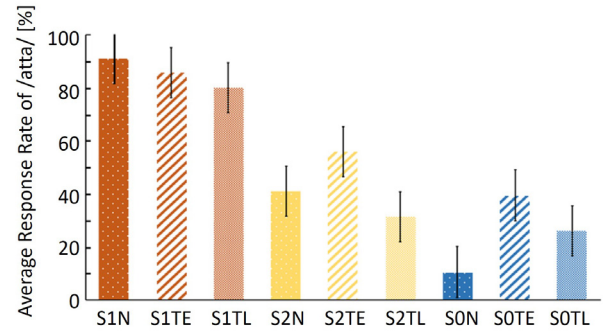


Figure 6: Response rates of /atta/ (%) under each stimulus condition in A+T session. All combinations of audio stimuli (S0, S1 and S2) and tactile stimuli (TE and TL) are plotted.

There were two types of tactile stimuli: TE and TL. The unit was open at the beginning of each stimulus (A in Fig. 3). For the TE stimulus, the unit started closing at the same time as the formant transition of the first vowel /a/ in S1 started (the same time as the VE stimulus, B in Fig. 3). For the TL stimulus, the unit started closing at the same time as the VL stimulus (C in Fig. 3). It started opening when the burst of /t/ occurred (D in Fig. 3). Thus, the TE stimulus closes the unit early (E), while the TL stimulus closes the unit late (L).

3. Experiment

3.1. Participants

Thirty-three normal listeners of Japanese (male: 10, female: 23, average age 22 years) participated in the experiment.

3.2. Procedure

The experiment was conducted in a sound-treated room. There were three sessions for each participant: audio only, A+V, and A+T. In each session, the stimuli discussed in Section 2 were presented eight times. The order of trials was randomized in each session. The order of the three sessions was also counterbalanced among the participants. The sounds and movements of the unit were controlled using a computer (Apple MacBook Pro), and the audio signals were presented through headphones (Sennheiser, HDA200) via a digital audio interface (Roland, Rubix24). A participant was asked to indicate what he/she heard for each trial: /atta/ or /a+/ta/.

3.3. Results

Two participants seemed not to understand the procedure, and we excluded their data from the rest of the analysis. Figure 5 shows the response rates (%) for all combinations of the audio stimuli (S0, S1, and S2) and visual stimuli (VE and VL). This figure also includes the results of the audio-only session, which are denoted as S0N, S1N, and S2N, respectively. The response rate (%) for each of the three audio signals was 91.1% for S1N, 41.1% for S2N, and 10.5% for S0N. Figure 6 shows the response rates (%) for all combinations of the audio stimuli (S0, S1 and S2) and tactile stimuli (TE and TL). This figure also includes the results of the audio-only session (S0N, S1N, and S2N).

Table 1: Results of the statistical analysis.

No.	Comparison	p-value for each pair		
		S0 vs. S1	S0 vs. S2	S1 vs. S2
(1)	S1N, S2N and S0N	< 0.001	0.028	< 0.001
		N vs. VE	N vs. VL	VE vs. VL
(2)	S0N, S0VE and S0VL	< 0.001	< 0.001	1.000
(3)	S1N, S1VE and S1VL	0.093	0.028	1.000
(4)	S2N, S2VE and S2VL	0.093	1.000	0.296
		N vs. TE	N vs. TL	TE vs. TL
(5)	S0N, S0TE and S0TL	< 0.001	0.003	1.000
(6)	S1N, S1TE and S1TL	0.432	0.147	1.000
(7)	S2N, S2TE and S2TL	0.126	0.197	<0.001

3.4. Statistical analysis

The results of statistical testing showed that there was a significant difference in the seven combinations in Table 1. Although the normality of these data was confirmed using the Shapiro-Wilk test prior to analysis, the normality was rejected in most cases. Since all combinations are paired three-group comparisons and a nonparametric test method that does not assume normality is required, all two groups were compared by the Friedman. If the Friedman test shows a significant difference between the groups, then pairwise comparisons with Bonferroni's correction were used. The statistical significance level was set to $\alpha = 0.05$ (two-tailed), and $p < 0.05$ was considered to be significantly different. SPSS Statistics 26 (IBM Corporation, Armonk, NY, USA) was used as the statistical analysis software.

Table 1 shows the results of the seven comparisons. (1) In the comparison of S1N, S2N, and S0N, a significant difference was observed in all combinations. Specifically, S1N was significantly higher than S0N, S2N was significantly higher than S0N, and S1N was significantly higher than S2N. (2) In the comparison of S0N, S0VE, and S0VL, we confirmed that S0VE and S0VL were significantly higher than S0N. No significant difference was found between S0VE and S0VL. (3) In the comparison of S1N, S1VE, and S1VL, we confirmed that S1N was significantly higher than S1VL. (4) In the comparison of S2N, S2VE, and S2VL, no significant difference was

observed in any of the combinations. (5) In the comparison of S0N, S0TE, and S0TL, we confirmed that S0TE and S0TL were significantly higher than S0N. (6) In the comparison of S1N, S1TE, and S1TL, no significant difference was observed in any of the combinations. (7) In the comparison of S2N, S2TE, and S2TL, we confirmed that S2TE was significantly higher than S2TL.

4. Discussion

Although many studies have previously reported on influences of sensorimotor on speech perception (e.g., [11–13]), we conducted an experiment on the sensorimotor speech perception for the geminate consonant in Japanese. Our objective in this work is to clarify how visual and tactile stimuli affect the speech perception of the audio stimuli of Japanese utterances. We selected the pair of utterances /atta/ and /a+/ta/ due to the following background. The geminate consonants in Japanese are difficult to acquire for second language learners, in general. They often learn that the duration of the target consonant is important and try to elongate the closure duration of stops and the frication duration of fricatives. However, native listeners of Japanese are sometimes not able to recognize an utterance as /atta/ by a non-native speaker when /a/ and /ta/ are concatenated with a long silence in between. Instead, it sounds like /a+/ta/ for native Japanese, with the two syllables lined up to each other with a pause in between them. When native speakers of Japanese utter /atta/, on the other hand, we know that the formant transitions before the geminate consonant play an important role [3]. In other words, the formant transitions at the end of the first vowel /a/ act as “glue” connecting the two syllables together. Then, with a long silence portion between /a/ and /ta/, it sounds like /atta/ instead of /a+/ta/. This can never be achieved without such glue as the formant transitions.

The next question we had in our previous study [4] was whether or not it is necessary that the glue be the formant transitions. In particular, we focused on the visual cue of the closing gesture of our articulators during closure of the consonant /t/. We prepared a pair of /atta/ and /a+/ta/ utterances, which have basically the same acoustic properties. That is, the first syllable was the vowel /a/ with the same duration followed by the silent part, while the second syllables were exactly the same. Thus, the total durations were the same. The only difference was the existence of the formant transitions at the end of the first syllable. We found that even if an audio signal is /a+/ta/, native Japanese perceive /atta/ if the closing gestures of the lower jaw and the tongue are seen simultaneously with the silent part between /a/ and /ta/. This tells us that such a closing gesture of the articulators also acts as glue and helps with the perception of the germination.

In our previous study [4], videos of a human face were used. However, it was almost impossible to make visual stimuli featuring exactly the same articulatory movements except for the timing of the closing of the lower jaw and the tongue. Therefore, in the present study we utilized a mechanical model of human speech production. As for the timing of the closing gesture, Point B for /atta/ and Point C for /a+/ta/ in Fig. 3 were compared.

In the audio-only session, the response rates of /atta/ were high for S1 (with the higher degree of formant transitions), low for S0 (with no formant transitions), and medium for S2 (with the medium degree of formant transitions). This was expected from previous studies [3, 4] and confirms that the F1 and F2 transitions at the end of the first syllable in /atta/ play an

important role when distinguishing it from /a+/ta/. One difference from the previous study [4] is the duration of the silent part. We found in the present study that even if we used a longer duration (700 ms instead of the 380 ms in [4]), the response rates for S1 reached higher than 90% (91.1%). This confirms the validity of these stimuli and demonstrates that the formant transitions still act as strong glue.

In the A+V session, both S0VE and S0VL conditions (visual closing cues with no formant transitions in audio signals) yielded significantly higher response rates compared to the audio-only S0N condition (exactly the same condition as S0 in Fig. 5). Because the formant transitions are missing from the S0 stimuli, this shows that visual cues compensated for the formant transitions when lacking auditory information, as discussed in our previous work [4].

In the A+T session, a similar tendency was observed as in the A+V session. In other words, both S0TE and S0TL conditions (tactile closing cues with no formant transitions in audio signals) yielded significantly higher response rates compared to the audio-only S0N condition. In addition, for the S2 stimulus set, the S2TE condition yielded a significantly higher response rate than the S2TL condition. This tells us that the early timing of the closing tactile information affects the geminate perception of listeners when the audio stimulus only contains weak formant transitions.

5. Conclusions

We investigated whether visual or tactile cues affect the speech perception of /a+/ta/ vs. /atta/. In the case of visual cues, we utilized a set of physical models of the human vocal tract along with a miniature robot hand unit. The same unit was used to provide tactile cues to a listener’s finger. When each visual or tactile stimulus was presented with an audio stimulus containing no formant transitions at the end of the first syllable, listeners more frequently responded to /atta/ compared to the audio-only presentations. This demonstrates that showing visual or tactile cues compensated for the missing formant transitions. The fact that visual or tactile information affected human speech perception in this study is probably due to the common properties between visual/tactile and audio cues, such as “closing gesture.” In other words, when native Japanese listeners perceive a geminate consonant, it seems they associate the gemination with the closing gesture during speech perception and it does not matter whether the closing gesture is achieved in any modality (audio, visual, or tactile).

In future work, we plan to investigate whether such common properties truly exist, and if so, whether there are other properties among different stimulus sets. This might also lead to a better understanding of the mechanisms underlying tactile signing. It is also interesting to discuss the topic of this study from a developmental point of view (e.g., [14]).

6. Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Number 18K02988 and Sophia University Special Grant for Academic Research (Research in Priority Areas). The experiments were approved by the Ethics Committee at Sophia University.

7. References

- [1] M. S. Han, "The feature of duration in Japanese," *Onsei no Kenkyu*, vol. 10, pp. 65–80, 1962.
 - [2] Y. Hirata, "Perception of geminated stops in Japanese word and sentence levels," *Onsei-gakkai-kaiho*, vol. 194, pp. 23–28, 1990.
 - [3] E. Yanagisawa and T. Arai, "Effects of formant transition and intensity damping of preceding vowel off-glide on perception of Japanese geminate consonant Sokuon," *J. Acoust. Soc. Jpn.*, vol. 71, no. 10, pp. 505–515, 2015.
 - [4] T. Arai, E. Iwagami and E. Yanagisawa, "Seeing closing gesture of articulators affects speech perception of geminate consonants," *J. Acoust. Soc. Am.*, vol. 141, no. 3, pp. EL319–EL325, 2017.
 - [5] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, 1976.
 - [6] S. Tremblay, D. M. Shiller and D. J. Ostry, "Somatosensory basis of speech production," *Nature*, vol. 423, pp. 866–869, 2003.
 - [7] T. Ito, M. Tiede and D. J. Ostry, "Somatosensory function in speech perception," *Proc. Natl. Acad. Sci.*, vol. 106, pp. 1245–48, 2009.
 - [8] B. Gick and D. Derrick, "Aero-tactile integration in speech perception," *Nature*, vol. 462, pp. 502–504, 2009.
 - [9] D. H. Klatt, "The new MIT speech VAX computer facility," *Speech Communication Group Working Papers IV*, Research Laboratory of Electronics, MIT, Cambridge, pp. 73–82, 1984.
 - [10] T. Arai, "Flexible tongue housed in a static model of the vocal tract with jaws, lips and teeth," *Proc. of INTERSPEECH*, pp. 171–172, 2018.
 - [11] G. Hickok, J. Houde and F. Rong, "Sensorimotor integration in speech processing: Computational basis and neural organization," *Neuron*, vol. 69, pp. 407–422, 2011.
 - [12] J. C. Myers, J. R. Mock and E. J. Golob, "Sensorimotor integration can enhance auditory perception," *Scientific Reports*, 10:1496, 2020.
 - [13] D. Silverman, "Bodily skill and internal representation in sensorimotor perception," *Phenomenology and the Cognitive Sciences*, vol. 17, pp. 157–173, 2018.
 - [14] D. Choi and J. F. Werker, "Speech perception and the sensorimotor system in Infancy," *J. Acoust. Soc. Am.*, vol. 150, A111, 2021.
-