

ACOUSTICAL LETTER

Audiovisual synchrony perception of simplified speech sounds heard as speech and non-speech

Kaori Asakawa[1,*], Akihiro Tanaka[2,†], Shuichi Sakamoto[1,‡], Yukio Iwaya[1,§] and Yôiti Suzuki[1,¶]

[1]*Research Institute of Electrical Communication and Graduate School of Information Sciences, Tohoku University, 2–1–1 Katahira, Aoba-ku, Sendai, 980–8577 Japan*
[2]*Waseda Institute for Advanced Study, 1–6–1 Nishi Waseda, Shinjuku-ku, Tokyo, 169–8050 Japan*

(*Received 29 October 2010, Accepted for publication 5 January 2011*)

Keywords: Audiovisual speech integration, Synchrony perception, Sine-wave speech
PACS number: 43.71.+m [doi:10.1250/ast.32.125]

1. Introduction

Audiovisual synchrony is important for comfortable speech communication. We occasionally encounter a temporal mismatch between a speaker's face and speech sound, for instance, in a satellite broadcast or in video streaming via the Internet. Human observers perceive physically desynchronized audiovisual signals as synchronous within a certain temporal tolerance [1]. This audiovisual synchrony perception may be affected by both structural factors (i.e., bottom-up factors) and cognitive factors (i.e., top-down factors). For example, audiovisual spatial congruency [2] and stimulus complexity [3] are considered to be structural factors, while cognitive factors include, for instance, an instruction ("imagine" in Arnold *et al.* [4]) that audiovisual stimuli originate from the same source, and "assumption of unity." This assumption of unity means the following: when multimodal inputs have highly consistent properties, it is more likely that observers treat them as originating from a single source [5–7] (see [5,6] for review).

However, the contribution of structural and cognitive factors to multisensory integration is unclear [6] because these two factors are often intermingled and it is not easy to distinguish between them [6,8]. Vatakis and Spence [7] tried to dissociate them and control the structural factors (i.e., matched stimulus complexity) to investigate the assumption of unity. They showed that participants were less sensitive to audiovisual asynchrony when the auditory and visual stimuli originated from the same speech event than when they originated from different speech events. They speculated that the strength of the assumption of unity by observers depends on whether or not the stimulus origin is different between audition and vision, and that it is the assumption of unity that influences audiovisual synchrony perception. In their study, audiovisual structural factors could be nearly controlled. Nonetheless, their stimuli were different in terms of structural factors as well as of cognitive factors. Thus, the influence of purely cognitive factors on audiovisual synchrony perception is not clear, especially in the case of speech signals.

*e-mail: kaori@ais.riec.tohoku.ac.jp
†e-mail: a.tanaka@aoni.waseda.jp
‡e-mail: saka@ais.riec.tohoku.ac.jp
§e-mail: iwaya@riec.tohoku.ac.jp
¶e-mail: yoh@riec.tohoku.ac.jp

In this study, we attempted to investigate this cognitive effect on audiovisual synchrony perception. For this purpose, we used a simplified speech sound called "sine-wave speech (SWS)" [9]. In SWS, a natural speech signal is replaced with three sinusoids corresponding to the first three formant frequencies (Fig. 1). SWS is heard as either speech or non-speech altered by instruction; thus, manipulation of only the cognitive factor is possible. Listeners without information about SWS typically perceive this sound as non-speech such as a whistle or electronic sound. In contrast, once they are informed that SWS is a synthesized speech sound based on natural speech, they perceive SWS as speech [9,10]. This procedure using physically identical SWS sound in both groups enables us to discuss whether an instruction on sound (i.e., cognitive factor) modulates the audiovisual synchrony perception.

We measured audiovisual temporal resolution as an index of audiovisual synchrony perception [3,7]. We hypothesized that cognitive factor, i.e., multisensory inputs refer to the same event, would enhance the multisensory integration.

2. Experiment 1

2.1. Method

2.1.1. Participants

In Experiment 1, thirty-three participants with normal hearing and normal or corrected-to-normal visual acuity took part. All were native Japanese speakers. The experiments were approved by the Ethics Committee of the Research Institute of Electrical Communication, Tohoku University.

2.1.2. Stimuli

The stimuli consisted of video clips of faces and voices of three Japanese female speakers (frontal view, including head and shoulders) uttering monosyllables: /pa/, /ta/ and /ka/. Consequently nine tokens were used in total. The video clips (640×480 pixels, Cinepak Codec video compression, 30 frames/s, 16 bits and 48 kHz audio signal digitization) were edited by Adobe Premiere Pro 1.5 (Adobe Systems Inc.) to make the stimulus onset asynchronies (SOA). The SWS sound was synchronized with the corresponding video clip by replacing the original speech sound.

SWS sounds of these monosyllables were created using an interactive method as follows: First, the formant frequencies were estimated by using STRAIGHT [11] to conduct spectro-

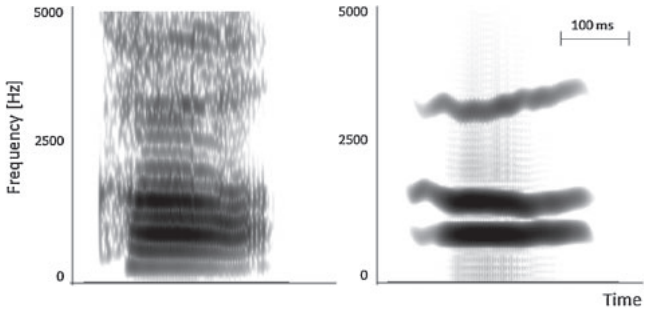

Fig. 1 Spectrograms of the natural /ka/ sound (left) and corresponding sine-wave speech (right) in this study.

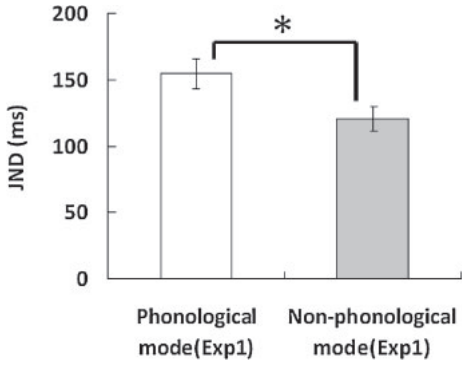

Fig. 2 Mean JND values for each condition in Experiment 1. The error bars represent the standard errors of the means. $^{*}p < 0.05$.

gram analysis of the original speech at an accuracy level of 1 ms. Corresponding amplitudes were extracted on the basis of these formant frequencies. These three formant and amplitude values were interpolated to resample at a sampling rate of 48 kHz and were low-pass filtered at 24 Hz to smooth their envelope. Then, three time-varying sine waves were synthesized in a manner to match the instantaneous frequencies derived from these formant frequencies and amplitudes. Consequently, SWS sounds retained the overall configuration of the original speech spectra despite lacking various characteristics of natural speech such as the fundamental frequency and broadband formant structure [9] (see Fig. 1).

2.1.3. Design

Two between-participants speech sound conditions were used: phonological mode condition and non-phonological mode condition. Participants in the phonological mode were instructed to regard SWS as synthetic speech, while those in the non-phonological mode did not receive such instruction. In addition, we provided participants in the former group with phonological information (i.e., original spoken syllables) of SWS sound to ensure the cognitive effect of instruction. Eighteen (two females and sixteen males, mean age 25.6 years) and fifteen (four females and eleven males, mean age 21.3 years) participants, respectively, took part in the phonological mode condition and in the non-phonological mode condition.

Another factor, stimulus onset asynchrony (SOA), was within-participants. We used 13 SOAs between the visual-speech and speech sound of the test stimulus (0, ±66, ±133, ±166, ±233, ±300, ±433 ms, where negative values signify speech sounds presented first).

2.1.4. Procedure

The experiment was conducted in a soundproof room. Participants were seated approximately 51 cm from a 21-inch CRT monitor (CDT2128A; Takaoka Co. Ltd.), wearing headphones (HDA 200; Sennheiser Electronic GmbH and Co. KG). The speech-sound was presented at a sound pressure level of 70 dB. Pink noise was added to speech-sound at a sound pressure level of 60 dB (cf [2,7].).

Participants received a pre-experimental session to acquaint themselves with SWS sound before the experimental sessions. In the pre-experimental session, each SWS sound was presented once in random order without visual stimuli. Participants were instructed to categorize each SWS sound as /pa/, /ta/ or /ka/ in the phonological mode condition, while participants in the non-phonological mode condition were asked to describe their impressions of each SWS sound. This enabled us to ensure that participants in non-phonological mode did not hear the SWS sounds as /pa/, /ta/ or /ka/. In an experimental session, the test stimulus was presented with various SOAs using the method of constant stimuli. The task of participants was to judge whether the vision or sound of the test stimulus was presented first. The experimental session, lasting approximately 10 min, comprised 78 test trials (6 × 13 SOAs). Each participant performed in four experimental sessions.

2.1.5. Results and discussion

The proportions of 'vision first' responses were converted to their equivalent z-scores under the assumption of a cumulative normal distribution. Then, from each of the fitted functions, the slope and intercept values were calculated. Slope values were then used to calculate the just noticeable difference (JND = 0.675/slope; ±0.675 indicates 75% and 25% on the cumulative normal distribution; similar to the approach in [7]).

Analysis of the JND data in an unpaired t-test (two-tailed) revealed a significant difference between conditions ($t(31) = 2.16$, $p < 0.05$) (Fig. 2). Significantly poorer performance was observed in the phonological mode condition ($M = 155$ ms) than in the non-phonological mode condition ($M = 121$ ms). This result shows that audiovisual temporal integration was modulated by an instruction regarding SWS.

We conducted a supplementary experiment on another day to examine the correct rate of identification for the SWS sound used in Experiment 1. In this experiment, nine SWS sounds were presented ten times each using the method of constant stimuli. Participants who participated in the phonological mode in Experiment 1 were instructed to categorize each SWS sound as /pa/, /ta/ or /ka/. Results showed that the correct rate of identification for the SWS sound was very low; 36.5% (i.e., approximately chance level). Given this poor discriminability of sound, it is possible that this poor discriminability affected the results of Experiment 1 in which participants in the phonological mode were asked to regard the sound as phonological speech. In Experiment 2, we examined this possibility by using only two tokens.

3. Experiment 2

3.1. Method

3.1.1. Participants

In Experiment 2, seventeen (four females and thirteen males, mean age 23.4 years; eleven participants took part in Experiment 1) participants took part in the phonological mode condition, and seventeen (six females and eleven males, mean age 21.6 years) new participants took part in the non-phonological mode condition.

3.1.2. Design and procedure

The experimental design, apparatus and procedure were the same as in Experiment 1 except for the following: The stimuli consisted of video clips of faces and SWS sound of a single Japanese female speaker uttering monosyllables: /pa/ and /ta/. These two stimuli were also used in Experiment 1. In addition, we conducted a training session before the experimental sessions. The training session consisted of a learning phase and a test phase. SWS sounds were presented 24 times with immediate feedback in the learning phase in pseudorandom order. In a test phase in the training session with 24 trials, participants had to categorized SWS sound as /pa/ and /ta/ in the phonological mode condition, or as '1' and '2' in the non-phonological mode condition. Participants were trained until a criterion (75% correct) was met. As a result of this pre-experimental training, the correct rates of identification in phonological and non-phonological mode condition were 94.1% and 93.8%, respectively. This high score indicated that the above-mentioned procedure increased the rate of correct identification without any modification of the stimuli. This must have made it easier for participants to identify/discriminate the auditory stimuli than in Experiment 1. After this training session, participants were instructed to begin the experimental session consisting of the audiovisual TOJ task as in Experiment 1.

3.1.3. Results and discussion

The JND values were calculated in the same manner as in Experiment 1 (Fig. 3). The JND value in the phonological mode appeared to be lower than that in Experiment 1, while the JND value in the non-phonological mode appeared to be not different from that in Experiment 1. No significant difference between the two conditions in Experiment 2 was revealed by analysis of the JND data in an unpaired two-tailed t-test ($t(25.8) = 0.76$, $p = 0.46$). This result showed that the sensitivities of the participants to the temporal order were not significantly different depending on whether they regarded an SWS sound and the natural face of a speaker as a speech sound with phonological information ($M = 118$ ms) or not ($M = 110$ ms) when they could substantially discriminate the sound. In addition, there was no significant difference between the performance of new and experienced participants in the phonological mode ($t(15) = 0.45$, $p = 0.66$).

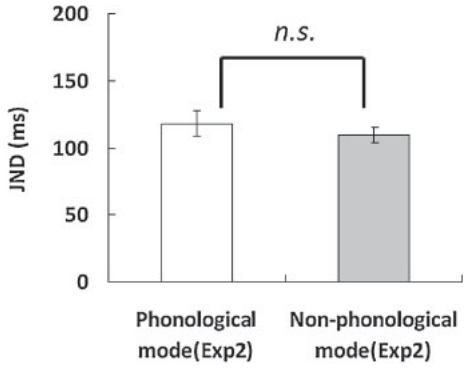

Fig. 3 Mean JND values for each condition in Experiment 2. The error bars represent the standard errors of the means.

4. General discussion

In this study, we investigated whether the instruction on phonological information of auditory stimuli could modulate the audiovisual synchrony perception in terms of temporal resolution. Our results of Experiment 2 showed that the phonological instruction did not affect the synchrony perception when the auditory stimuli were substantially discriminable. In contrast, the results of Experiment 1, where the same auditory stimuli were hardly discriminable, showed that the effect was significant. This suggests that cognitive factor such as instruction providing phonological information on audiovisual stimuli has little influence on the synchrony perception when the stimuli are physically identical between the conditions and sufficiently discriminable. Results similar to Experiment 2 have been recently reported in a study by Vroomen and Stekelenburg [12], who conducted an audiovisual TOJ experiment in which they used a single Dutch pseudoword stimulus (/tabi/) composed of SWS and a corresponding face video. They did not find any significant difference between the SWS speech mode condition (the original speech sound and SWS were alternately presented ten times before the experiment) and the SWS non-speech mode condition (participants were informed that the SWS was an artificial computer sound). Although their stimuli and experimental procedure were different from ours, their results support the generality of our findings that the sensitivity of the participants for the audiovisual synchrony is not modulated by whether they have phonological information or not.

However, in the present study, the results of Experiment 1 showed that the participants found it harder to judge the audiovisual temporal order when they were provided with phonological information about the poorly-discriminable SWS sound than when they were not provided with that information. This effect dissipated in Experiment 2 as the rate of correct identification was increased by training sound categorization and limiting variation of stimuli.

There is one possible explanation for the inconsistency between Experiments 1 and 2 as follows: Between the two experiments, there might have been a difference in terms of attentional resources for the TOJ task in the phonological mode. When participants were told that poorly discriminable SWS sounds were phonological, they could not fully concentrate on the TOJ task. That is, a certain portion of attentional resources could have been utilized for speech identification even if the participants were not concentrating speech identification but on the TOJ task. A previous study has reported poorer audiovisual temporal resolution (i.e., larger JND values) in a dual-task situation than in a single-task situation [13]. This suggests that the attentional cost can affect the performance of the TOJ task. Given these findings,

the poor temporal resolution in the phonological mode in Experiment 1 could have resulted from the reduced attentional resource attributed to the poorly discriminable sound.

Results of our study imply that sound discriminability may affect audiovisual synchrony perception of speech (average scores of the rate of correct identification in the phonological mode: 36.5% in Experiment 1 vs 94.1% in Experiment 2). Similar results have been reported by previous studies (e.g., [14]). Conrey and Pisoni [14] found a significant correlation between audiovisual synchrony perception for speech and audiovisual speech intelligibility. In our experiments, participants in the phonological mode showed a smaller JND, i.e., better performance, of audiovisual synchrony perception in Experiment 2, where the correct rate of identification was higher, than in Experiment 1, where the correct rate was just at the chance level. The correct rate of identification was comparable to intelligibility in their experiment. Thus, the present results are understood as being consistent with their findings.

In this study, we investigated the cognitive effect in audiovisual synchrony perception using physically identical stimuli. The findings reveal the role of a purely cognitive effect in audiovisual synchrony perception, and suggest that synchrony perception might be affected by the sound discriminability. Further studies are required to reveal the relevance of the relation between audiovisual synchrony perception and stimulus discriminability.

Acknowledgements

This work was supported by Grants-in-Aid for Specially Promoted Research No. 19001004 from the MEXT Japan to SY, for JSPS Fellows No. 21-8191 to AK, and the Cooperative Research Project Program of the Research Institute of Electrical Communication, Tohoku University (H22-A09) to TA. The authors thank Dr. H. Kawahara for permission to use the STRAIGHT system.

References

[1] V. van Wassenhove, K. W. Grant and D. Poeppel, "Temporal window of integration in bimodal speech," *Neuropsychologia*, **45**, 598–607 (2007).

[2] C. Spence, R. Baddeley, M. Zampini, R. James and D. I. Shore, "Multisensory temporal order judgments: When two locations are better than one," *Percept. Psychophys.*, **65**, 318–328 (2003).

[3] A. Vatakis and C. Spence, "Audiovisual synchrony perception for music, speech, and object actions," *Brain Res.*, **1111**, 134–142 (2006).

[4] D. H. Arnold, A. Johnston and S. Nishida, "Timing sight and sound," *Vision Res.*, **45**, 1275–1284 (2005).

[5] R. B. Welch and D. H. Warren, "Immediate perceptual response to intersensory discrepancy," *Psychol. Bull.*, **88**, 638–667 (1980).

[6] C. Spence, "Audiovisual multisensory integration," *Acoust. Sci. & Tech.*, **28**, 61–70 (2007).

[7] A. Vatakis and C. Spence, "Crossmodal binding: Evaluating the "unity assumption" using audiovisual speech stimuli," *Percept. Psychophys.*, **69**, 744–756 (2007).

[8] M. Radeau and P. Bertelson, "Adaptation to auditory-visual discordance and ventriloquism in semirealistic situations," *Percept. Psychophys.*, **22**, 137–146 (1977).

[9] R. E. Remez, P. E. Rubin, D. B. Pisoni and T. D. Carrell, "Speech perception without traditional speech cues," *Science*, **212**(4497), 947–949 (1981).

[10] J. Vroomen and M. Baart, "Phonetic recalibration only occurs in speech mode," *Cognition*, **110**, 254–259 (2009).

[11] H. Kawahara, I. Masuda-Katsuse and A. de Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," *Speech Commun.*, **27**, 187–207 (1999).

[12] J. Vroomen and J. J. Stekelenburg, "Perception of intersensory synchrony in audiovisual speech: Not that special," *Cognition*, **118**, 78–86 (2011).

[13] J. Navarra, A. Vatakis, M. Zampini, S. Soto-Faraco, W. Humphreys and C. Spence, "Exposure to asynchronous audiovisual speech extends the temporal window for audiovisual integration," *Cognit. Brain Res.*, **25**, 499–507 (2005).

[14] B. Conrey and D. B. Pisoni, "Auditory-visual speech perception and synchrony detection foe speech and nonspeech signals," *J. Acoust. Soc. Am.*, **119**, 4065–4073 (2006).