Temporal Recalibration in Audio-Visual Speech Integration Using a Simultaneity Judgment Task and the McGurk Identification Task

Kaori ASAKAWA[1] (kaori@ais.riec.tohoku.ac.jp)
Division of Psychology, Graduate School of Humanities, Tokyo Woman's Christian University
2-6-1 Zempukuji, Suginami-ku, Tokyo 167-8585, Japan

Akihiro TANAKA[2] (a.tanaka@uvt.nl)
Department of Psychology, Faculty of Letters/Graduate School of Humanities and Sociology, University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan

Hisato IMAI (hisato@lab.twcu.ac.jp)
Department of Psychology, College of Arts and Sciences, Tokyo Woman's Christian University
2-6-1 Zempukuji, Suginami-ku, Tokyo 167-8585, Japan

Abstract

Audio-visual synchrony is important for comfortable speech communication. Previous studies have revealed a temporal window during which human observers perceive physically desynchronized auditory and visual signals as synchronous in both speech and nonspeech signals. This temporal window of audio-visual integration is re-calibrated after adaptation to a constant timing difference between auditory and visual signals in nonspeech. In this study, we investigate whether or not the after-effects of the consequences of temporal recalibration occur even in speech stimuli. Our results suggest that the temporal recalibration occurs not only in nonspeech signals but also in monosyllabic speech.

Keywords: audio-visual integration; speech perception; temporal recalibration; simultaneity judgment; the McGurk effect

Introduction

Human beings perceive light and sound generated from an event via their respective organs, that is, eyes and ears. The light and sound are usually asynchronous. For example, a distant sound arrives later than a visual component because of their different velocities. People can detect a temporal mismatch due to a technical limitation in a live satellite broadcast. In addition, there is a difference between the sensory processing latencies of audition and vision (Spence & Squire, 2003). Thus, audio-visual temporal asynchrony occurs due to factors in the environment and in the brain processing.

Audio-visual temporal asynchrony is tolerated to some extent for auditory and visual stimuli to be perceived as a single event. This temporal tolerance, the so-called "temporal window," has been examined through various kinds of stimuli and tasks (e.g., Dixon & Spitz, 1980; Conrey & Pisoni, 2006; Grant & Greenberg, 2001; Vatakis & Spence, 2006a; Vatakis & Spence, 2006b). These studies have revealed that the audio-visual temporal window depends on the characteristics of the stimulus (e.g., complexity, duration, and ecological validity) and/or tasks (e.g., asynchrony detection, simultaneity/temporal order judgment, and speech identification).

Recent studies have shown that the audio-visual temporal window is recalibrated after adaptation to a constant timing difference between auditory and visual signals. Both Fujisaki, Shimojo, Kashino, and Nishida (2004) and Vroomen, Keetels, de Gelder, and Bertelson (2004) demonstrated shifts in the point of subjective simultaneity (PSS: the amount of time between the auditory and visual stimuli required for perceptual simultaneity) following exposure to a series of desynchronized simple audio-visual pairs toward the exposure lag.

The above studies used simple nonspeech pairs of light and sound. However, the temporal discrepancy that we encounter in the real world is more complex (e.g., speech, music, or action). Especially, speech signals play an important role in daily communications. Previous studies have investigated the temporal window of audio-visual speech integration in both synchrony perception (Dixon & Spitz, 1980; Conrey & Pisoni, 2006; Vatakis & Spence, 2006a; Vatakis & Spence, 2006b) and speech identification (Grant & Greenberg, 2001; Munhall, Gribble, Sacco, & Ward, 1996; van Wassenhove, Grant, & Poeppel, 2007). In addition to containing linguistic information, a speech signal has different characteristics from a simple nonspeech signal in terms of structural factors (e.g., physical complexity, saliency of onset/offset) and cognitive factors (e.g., ecological validity, familiarity). The question of whether speech is special has been discussed (see Jones & Jarick, 2006; Vatakis, Ghazanfar, & Spence, 2008; Tuomainen, Andersen, Tiippana, & Sams, 2005; Bernstein, Auer, & Moore, 2004). However, it has not been settled. Given these

[1] Present affiliation: Research Institute of Electrical Communication and Graduate School of Information Sciences, Tohoku University, 2-1-1 Katahira, Aoba-ku, Sendai 980-8577, Japan

[2] Present affiliation: Cognitive and Affective Neurosciences Laboratory, Tilburg University, PO Box 90153, 5000 LE Tilburg, The Netherlands

aspects of speech, in order to reveal the temporal recalibration of audio-visual speech, it is proper to use a speech signal, not a simple nonspeech signal.

Several studies have investigated temporal recalibration using a speech signal (Navarra, Vatakis, Zampini, Soto-Faraco, Humphreys, & Spence, 2005; Vatakis, Navarra, Soto-Faraco, & Spence, 2007; Vatakis, Navarra, Soto-Faraco, & Spence, 2008). Vatakis *et al*. (2007) conducted a study in which the participants had to make a temporal order judgment (TOJ) regarding pairs of asynchronous vowel-consonant-vowel (VCV) speech-sounds and visual-speech gestures while monitoring a continuous background consisting of an audio-visual speech stream of words (i.e., an on-line adaptation method). The continuous (adapting) speech stream could either be presented in synchrony or with the auditory stream lagging. Participants conducted the TOJ task either in a single task condition (i.e., devoted themselves to perform the TOJ task) or in a dual-task condition (i.e., conducted the TOJ task in parallel with counting the number of male names included in the asynchronous background speech stream). A significant PSS shift was observed in the direction of the adapting stimulus only when participants made a TOJ in the dual-task condition. This result suggests that the temporal perception for audio-visual speech is pulled toward the monitoring asynchrony only during a dual-task situation.

Although the above studies have demonstrated temporal recalibration using a speech signal, more consideration appears to be necessary for revealing the temporal recalibration for speech. First, their results were limited to a dual-task situation. Given these results, it is unclear what affects the temporal realignment in the on-line adaptation method; it might be the attention paid to background speech and/or a decline of an attentional resource for the target stimulus (Navarra *et al.* (2005) also showed, however, the temporal window was not affected when the background lag was too large (1000 ms). This suggests that it seems implausible to explain solely in terms of the reduced attentional source). In the latter case, the results from the on-line adaptation method might tap a different mechanism from those from an off-line adaptation method. Thus, it is not clear whether the temporal recalibration *following* exposure to asynchrony (i.e., off-line adaptation method), not in a dual-task situation, occurs for an audio-visual speech signal. Second, previous studies for temporal recalibration in speech have investigated using only a direct measure in which participants can judge the simultaneity explicitly. Because most of the information in any direct task is open to conscious inspection, the responses can easily reflect response strategies rather than a perceptual process (see Bertelson & de Gelder, 2004; Fujisaki *et al*., 2004). In order to reduce this possibility, it is useful to adopt an indirect (implicit) measure as well as a direct measure.

Therefore, in this study, we investigated the temporal recalibration for audio-visual speech following exposure to asynchrony using both direct and indirect measures. We used a simultaneity judgment task as a direct measure and a speech identification task as an indirect measure. We used a phenomenon which is known as the McGurk effect (McGurk & McDonald, 1976). The McGurk effect is created by dubbing an auditory /pa/ onto a visual /ka/. The participants observed this clip experience hearing /ta/ as a result of the integration of an auditory /pa/ and a visual /ka/. The more synchronous the visual and auditory speech are, the stronger the McGurk effect is (Munhall, *et al*., 1996; van Wassenhove, *et al*., 2007). We focused only on the aspect of the audio-delay of audio-visual asynchrony, given the ecological validity of audio-delay timing rather than visual-delay.

Experiment 1: Simultaneity Judgment Task

Methods

Participants Ten participants (mean age of 23.4 years) took part in Experiment 1. All of them reported normal hearing and normal or corrected-to-normal visual acuity**.** All were native Japanese speakers.

Materials The stimulus was based on digital audio and video recordings of a female native speaker of Japanese. The visual materials were recorded using a DV camera (HDR-A1J, Sony). Auditory stimuli were collected using a condenser microphone (ECM-77B, Sony). The audio uttering of /pa/ was dubbed onto an audio track of a video uttering of /ka/ (frontal view, including head and shoulders) using Sound Forge 8.0 (Sony) to get precise synchronization. The video clip (640 × 480 pixels, Cinepak Codec video compression, 30 frames/s) and the auditory speech (16-bit and 48-kHz audio signal digitization) were synthesized and desynchronized using Adobe Premiere Pro 2.0. A still image was extracted from the first and the last frame of the video clip and added to the beginning and end of the video clip to fill the blank field.

Design The experiment had two within-participants factors: Adaptation lag (audio delay: 0, 233 ms) and stimulus onset asynchronies (SOA) between the visual-speech and speech-sound of the test stimulus (audio delay: 0, 66, 133, 166, 233, 300, 433 ms). The audio-delay lag was used in terms of ecological validity based on the velocities of light and sound. This provided a half of the temporal window.

Procedure The experiment was conducted in a sound-proof room. The participant was seated at a distance of approximately 50 cm from a 17-inch CRT monitor (CPD-E220, Sony), wearing headphones (HDA 200, Sennheiser). The speech-sound was presented at the sound pressure level of approximately 65 dB. The pink noise overlapped with the speech-sound at 65 dB (i.e., S/N 0 dB). The video clip was presented on a black background using Real Player Ver. 10.5 (RealNetworks).

Each session started with an adaptation phase of 3 min with a constant time lag between the visual speech and

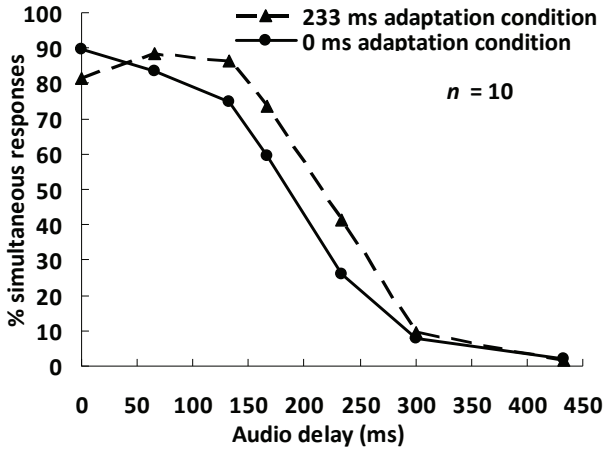

Fig.1 Percentages of simultaneous responses as a function of adapted lag and audio delay.

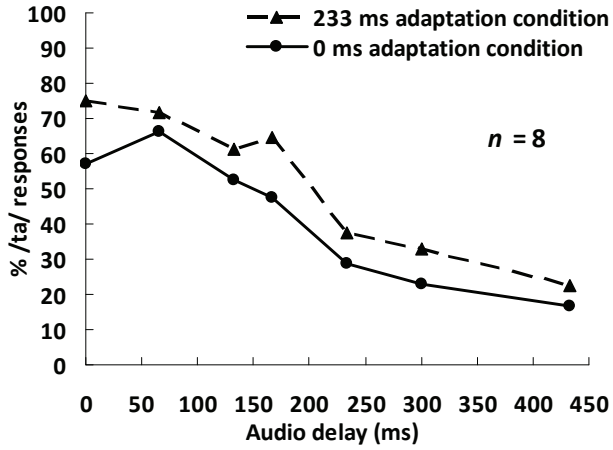

Fig.2 Percentages of /ta/ responses as a function of adapted lag and audio delay.

speech sound. The adaptation phase was followed by test trials, each preceded by a 10-s re-adaptation. After a 2-s red-tinted still image as a cue of the test trial, a test stimulus was presented with various SOAs. The participants' task was to judge whether auditory and visual stimuli were presented synchronously or asynchronously. Participants were instructed to respond accurately rather than quickly. The various SOAs of the test stimuli were randomly presented in each session using the method of constant stimuli. The experimental session, which lasted approximately 20 minutes, consisted of 42 test trials (6 repetitions of the 7 SOAs). Four experimental sessions were run for each adaptation condition. They participated in one adaptation condition per day.

Results and discussion

Figure 1 shows the percentages of simultaneous responses as a function of adapted lag and SOAs. The simultaneous responses decreased as the amount of SOAs increased. The percentages of simultaneous responses were higher in the audio delay (233 ms) adaptation condition than in the synchronous (0 ms) adaptation condition except when the SOAs were 0 and 433 ms. In a two-way analysis of variance (ANOVA), there were significant main effects of adapted lag [$F(1, 9) = 5.31$, $p < .05$] and test lag [$F(6, 54) = 131.32$, $p < .01$]. Interaction between those two factors was only marginal [$F(6, 54) = 1.87$, $p < .10$]. These results suggest that adaptation to audio delay timing affects the synchrony perception toward the adapted (audio-delay) lag.

Thus, our results suggest that the temporal window for speech was modulated toward the adapted lag using a direct measure (i.e., simultaneity judgment). In experiment 2, we used an indirect measure to reduce the possibility that this modulation of the temporal window resulted from post-perceptual influences.

Experiment 2: The McGurk Identification Task

Methods

Participants Eight participants (mean age of 22.3 years) took part in Experiment 2. All of them reported normal hearing and normal or corrected-to-normal visual acuity. All were native Japanese speakers.

Materials, Design and Procedure The participants' task was to choose what they heard while looking at a mouth. They had to give their answers among four choices (pa, ta, ka, or other). Except for this, the materials, design and procedure were the same as in Experiment 1.

Results and discussion

We adopted /ta/ responses as a typical McGurk illusion resulting from the integration of an auditory /pa/ and a visual /ka/ (see van Wassenhove *et al*., 2007). Figure 2 shows the percentages of /ta/ responses (the McGurk illusion) as a function of adapted lag and SOAs. The temporal range within which McGurk illusion is obtained is up toaround 200 ms audio-delay (267 ms in van Wassenhove et al. (2007) and240 ms in Munhall et al. (1996), although these studies adopteddifferent criteria for the McGurk effect). Consistent with thesestudies, the percentage of /ta/ responses was significantly lower whenthe audio-delay was 233 ms in our results. As the simultaneous responses in Experiment 1, the /ta/ responses decreased as the amount of SOAs increased. Also, the percentages were higher in the audio delay adaptation condition than in the synchronous adaptation condition. In a two-way ANOVA, there were significant main effects of adapted lag [$F(1, 7) = 9.28$, $p < .01$] and test lag [$F(6, 42) = 18.59$, $p < .01$]. Interaction was not significant [$F(6, 42) = 0.29$, $p > .10$].

These results suggest that adaptation to audio delay timing alters the McGurk illusion.

Our results showed that the temporal window changes not only with the direct measure but also indirect measure (i.e., speech identification). This finding supports the idea that temporal recalibration occurs not only for nonspeech signals but also for monosyllabic speech. In addition, this temporal recalibration could be regarded as a perceptual phenomenon, not a post-perceptual change, because temporal recalibration occurred with both direct and indirect measures.

General Discussion

In this study, we investigated the temporal recalibration for audio-visual speech following exposure to asynchrony (i.e., in an off-line adaptation method) using both a direct measure (i.e., the simultaneity judgment) and an indirect measure (i.e., the McGurk identification). The results showed that the temporal window for audio-visual speech is modulated in the direction of adapted lag (audio-delay) with both measures. These results suggest that the audio-visual temporal recalibration following exposure to audio-delay timing occurs not only for nonspeech signals but also for speech signals at the perceptual level.

Our results were consistent with those of Fujisaki *et al*. (2004) in that the audio-visual temporal window changed toward the adapted lag in the off-line adaptation method using both direct and indirect measures. Fujisaki *et al*. demonstrated temporal recalibration using both the simultaneity judgment and a stream/bounce illusion (Sekuler, Sekuler, & Lau, 1997) using simple nonspeech signals. The similar results of these studies could suggest that audio-visual temporal recalibration occurs perceptually both for simple nonspeech and speech. That is, the temporal window for speech might be recalibrated, nonetheless speech appears to have several characteristics, e.g., relatively high ecological validity, physical complexity, familiarity, and relatively low saliency of onset/offset. In a future study, it should be clarified whether the temporal window shifts toward the adapted lag and/or widens.

Using the off-line adaptation method, our results showed that the temporal window changes even in a single-task condition. A previous study using the on-line adaptation method showed that the temporal window changes only under the dual-task condition (Vatakis *et al*., 2007). This discrepancy suggests either that attention to the adaptation stimulus is needed for temporal recalibration or that the responses in these two adaptation methods reflect different mechanisms. Another possibility is that these two methods reflect different developmental phases of recalibration. Namely, the on-line method reflects an initial phase of recalibration while the off-line method reflects a later phase (cf. Navarra *et al*., 2005; Vatakis, Navarra, Soto-Faraco, & Spence, 2008). In order to investigate these possibilities, it is necessary to reveal temporal recalibration in audio-lead timing, and development in temporal recalibration.

Conclusion

In this study, we investigated temporal recalibration for audio-visual speech following exposure to asynchrony using both a simultaneity judgment task and the McGurk identification task. The results showed that the temporal window modulated toward exposed (audio-delay) timing in both direct and indirect measures. These results suggest that temporal recalibration occurs for speech signals at the perceptual level.

Acknowledgments

A part of this work was supported by a Grant-in-Aid for Specially Promoted Research No. 19001004 from the Ministry of Education, Culture, Sports, Science and Technology, Japan.

References

Bernstein, L.E., Auer, E.T., & Moore, J.K. (2004). Audiovisual speech binding: convergence or association? In Calvert, G.A., Spence, C., & Stein, B.E. (eds) *The handbook of multisensory processing.* MIT, Cambridge, pp. 203–223.

Bertelson, P., & de Gelder, B. (2004). The psychology of multimodal perception. In Spence, C., & Driver, J. (eds) *Crossmodal space and crossmodal attention.* Oxford University Press, Oxford, pp. 141–177.

Conrey, B., & Pisoni, D.B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *Journal of the Acoustical Society of America, 119,* 4065–4073.

Dixon, N., & Spitz, L. (1980). The detection of Audiovisual desynchrony. *Perception, 9,* 719–721.

Fujisaki, W., Shimojo, S., Kashino, M., & Nishida, S. (2004). Recalibration of audiovisual simultaneity. *Nature Neuroscience, 7,* 773-778.

Grant, K.W., & Greenberg, S. (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. *Paper presented at the ISCA International Conference on Auditory-Visual Speech Processing*

Jones, J.A., & Jarick, M. (2006). Multisensory integration of speech signals: The relationship between space and time. *Experimental Brain Research, 174*, 588–594.

McGurk, H., & McDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–747.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk Effect. *Perception and Psychophysics, 58,* 351–362.

Navarra, J., Vatakis, A., Zampini, M., Soto-Faraco, S., Humphreys, W., & Spence, C. (2005). Exposure to asynchronous audiovisual speech increases the temporal window for audiovisual integration of non-speech stimuli. *Cognitive Brain Research, 25,* 499–507.

Roberts, M., & Summerfield, Q. (1981). Audiovisual presentation demonstrates that selective adaptation in speech perception is purely auditory. *Perception and Psychophysics, 30,* 309–314.

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound alters visual motion perception. *Nature, 385,* 308.

Spence, C., & Squire, S.B. (2003). Multisensory integration: maintaining the perception of synchrony. *Current Biology, 13,* R519–R521

Tanaka, A., Sakamoto, S., Tsumura, K., & Suzuki, Y. (2009). Visual speech improves the intelligibility of time-expanded auditory speech. *NeuroReport, 20,* 473–477.

Tuomainen, J., Andersen, T.S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition, 96*, B13–B22.

Vatakis, A., Ghazanfar, A.A., & Spence, C. (2008). Facilitation of multisensory integration by the 'unity effect' reveals that speech is special. *Journal of Vision, 8,* 1–11.

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2007). Temporal recalibration during asynchronous audiovisual speech perception. *Experimental Brain Research, 181,* 173–181.

Vatakis, A., Navarra, J., Soto-Faraco, S., & Spence, C. (2008). Audiovisual temporal adaptation of speech: Temporal order versus simultaneity judgments. *Experimental Brain Research, 185,* 521–529.

Vatakis, A., & Spence, C. (2006a). Audiovisual synchrony perception for music, speech, and object actions. *Brain Research, 1111,* 134–142.

Vatakis, A., & Spence, C. (2006b). Audiovisual synchrony perception for speech and music using a temporal order judgment task. *Neuroscience Letters, 393,* 40–44.

Vroomen, J., Keetels, M., de Gelder, B., & Bertelson, P. (2004). Recalibration of temporal order perception by exposure to audio-visual asynchrony. *Cognitive Brain Research, 22,* 32–35.

van Wassenhove, V., Grant, K.W., & Poeppel, D. (2007). Temporal window of integration in bimodal speech. *Neuropsychologia, 45,* 598-607.